AARHUS UNIVERSITY

Integrative cancer genomics analysis of gene expression and DNA methylation

PhD dissertation

Michał Świtnicki

Health
Aarhus University
2015

Integrative cancer genomics analysis of gene expression and DNA methylation

PhD dissertation

Michał Świtnicki

Health
Aarhus University
Department of Molecular Medicine

Preface

This dissertation contains my primary scientific contributions made during my employment as a PhD student at the Department of Molecular Medicine (MOMA), Aarhus University Hospital, Denmark, from April 2012 to June 2015. This dissertation is part of the fulfilment of the requirements for the Doctorate of Philosophy at Aarhus University, in the Translational Molecular Medicine programme offered by the Graduate School of Health.

Supervisors:

Jakob Skou Pedersen, MSc, PhD (Principal Supervisor); Professor at Department of Molecular Medicine, Aarhus University Hospital and at Bioinformatics Research Centre, Aarhus University, Denmark

Karina Dalsgaard Sørensen, MSc, PhD; Associate Professor at Department of Molecular Medicine, Aarhus University Hospital, Denmark

Torben Falck Ørntoft, MD, PhD, Professor and Head of Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark

Carsten Wiuf, MSc, PhD; Professor at Department of Mathematical Sciences, University of Copenhagen, Denmark

Roald Forsberg, MSc, PhD; General Manager at QIAGEN Aarhus, Denmark, Vice President of Engineering, Bioinformatics at QIAGEN, Netherlands

The dissertation contains 10 chapters of which two (Chapters 8 and 10) contain articles currently in peer review at *Bioinformatics* journal while Chapter 9 contains a manuscript in preparation. The manuscripts are listed below:

Manuscript 1:

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

***Michał P. Świtnicki**, Malene Juul, Tobias Madsen, Karina D. Sørensen, Jakob S. Pedersen*

Manuscript 2:

Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data

***Michał P. Świtnicki**[*], Tobias Madsen[*], Jakob S. Pedersen* * shared first authorship*

Manuscript 3:

ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction

*Sudhakar Sahoo, **Michał P. Świtnicki**, Jakob S. Pedersen*

Acknowledgements

I express my deepest appreciation to my main supervisor Jakob Skou Pedersen for his exceptional guidance, inputs to my works, for being understanding and pragmatic, for the occasional scientific sparring and the personal support I received from him. I second my gratitude to my other supervisor, Karina Dalsgaard Sørensen for the opportunity to be involved in this project, for critically reviewing my works and for many scientific discussions throughout my studies. I also thank Torben Ørntoft, Carsten Wiuf and Roald Forsberg for being available to supervise aspects of the project and discuss relevant matters, when needed.

I would also like to express my gratitude to my dear colleagues at MOMA, to Philippe Lamy for scientific sparring, to Siri Strand and Tobias Madsen for fruitful collaborations, to Christa Haldrup for clinical consultations, to Morten Muhlig for initial coding advices, to Frank Sørensen and Søren Vang for their advices on using the operating system and for being helpful, to Malene Juul for statistical consultations and for help with translations, to Sudhakar Sahoo, for keeping me a good company throughout most of my time at MOMA, for interesting discussions and fruitful collaboration. I also thank all the people at MOMA for warm welcome in the beginning, for teaching me about the Danish customs and for occasional exercises in Danish with me.

I would also like to thank Aarhus University, Aarhus University Hospital, CLC bio A/S, The Danish Strategic Research Council and The Danish Council for Independent Research for making this project possible to carry out, also financially.

I want to especially express my gratitude to my best friend Adam for discussions about the matter of life and supplying me with great music recommendations that kept me going through the busy periods of finalizing this project.

I also thank my parents and the rest of family for supporting me throughout this extended period, for being there when needed, backing me with a good word and for a constant supply of comfort.

Lastly, I thank Jotun Hein, Zoltan Szallasi and Jakob Grove for agreeing to examine my PhD thesis.

Michał Świtnicki

September 2015

Department of Molecular Medicine

Aarhus University

Abbreviations

AUC: area under the receiver operating characteristic curve

BRCA: breast cancer adenocarcinoma

cDNA: complementary deoxyribonucleic acid

CMCT: 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene

DMS: dimethyl sulphate

EM: expectation-maximization algorithm

FDR: false discovery rate

FFPE: formalin-fixed paraffin-embedded (tissue)

HNSCC: head and neck spinocellular carcinoma

HR: hazard ratio

ICGC: International Cancer Genome Consortium

KL: Kullback-Leibler divergence

KS: Klinefelter syndrome

lincRNA: large intergenic non-protein coding ribonucleic acid

lncRNA: long non-protein coding ribonucleic acid

LSCC: laryngeal spinocellular carcinoma

MDS: multidimensional scaling

MOMA: Department of Molecular Medicine, Aarhus University Hospital, Denmark

mRNA: messenger ribonucleic acid

NGS: next generation sequencing

PC: prostate cancer

PGM: probabilistic graphical model

PCR: polymerase chain reaction

R: The R Project for Statistical Computing

RNA: ribonucleic acid

RNA-seq: sequencing of reverse-transcribed ribonucleic acid

rRNA: ribosomal ribonucleic acid

SCFGs: stochastic context-free grammars

SHAPE: selective 2’ hydroxyl acylation analysed by primer extension

TCGA: The Cancer Genome Atlas

450k: Infinium HumanMethylation450 BeadChip®

Contents

Chapter 1: Introduction

Cancer and cancer diagnostics

Cancer can be viewed as a disease of the genome characterized by abnormal growth and spread of cells. The development and progression of cancer disease are driven by a complex pattern of genomic and epigenomic changes, happening on an evolutionary basis. A number of hallmarks were defined as necessary for development of tumours, including sustaining proliferation, evading growth suppression, resisting cell death, enabling replicative immortality, inducing development of blood vasculature, activating invasive mechanisms of metastasis, evading destruction by the host immune system and reprogramming of energy metabolism (Hanahan and Weinberg, 2011).

A standard diagnosis of the cancer disease is organ-specific, but generally, the most robust disease confirmation can be obtained using histological evaluation of biopsies. However, diagnoses using biopsies are usually not performed early on, and fail to discriminate tumours based on their susceptibility to available treatments. Also, prognostication is very difficult based on the histological picture alone. Therefore, molecular profiling strategies have been developed to improve the management of cancer patients, focusing on the detection of biomarkers implicated in the cancer hallmarks. These biomarkers, apart from adding diagnostic value, may also be used for prognostication in a personalized medicine regime (Kalia, 2015). Molecular profiling may also contribute to early diagnostics, as some of molecular biomarkers can be detected in physiological fluids (Sethi, et al., 2013).

Traditional molecular biomarkers are based on a single data type, i.e. DNA mutations, gene expression, copy number status or specific protein expression levels. However, a gene might be disrupted by either of sequence mutation, or aberrant methylation, or differential expression, each having the same functional effect on the disease of interest. The traditional biomarkers implicitly assume that the same mechanism of disruption is diagnostic for every case, which is often not the case. This might partly explain our inability to properly diagnose many tumours - a notable example being the prostate cancer (Felgueiras, et al., 2014) with very high rates of overtreatment of clinically insignificant tumours. This assumption limits the predictive power of known biomarkers, and reduces the number of identifiable new ones. In particular, the required functional change might be produced by silencing with DNA methylation or by loss of a copy number but neither occurs frequently enough to be identified by statistical analyses and hence considered for further biomarker validation. In some cases, the required change might only be produced by a simultaneous change in both methylation and copy number, but none of the two changes reach statistical significance alone due to small individual effect size.

These kinds of considerations lead us to a new chapter in the biomarker field, first defined formally for neurodegenerative diseases in a recent publication (Carreiro, et al., 2015), the integrative biomarkers. For example, an integrative biomarker can be a gene for which we measure several of its molecular characteristics like DNA methylation and gene expression. Then, based on the combined assessment of these complementary pieces of information, we can perform diagnosis and/or prognosis. Hence, integrated analysis of the different data types should be of focus in future biomarker discovery endeavours.

RNA secondary structure prediction

For a long time, ribonucleic acid (RNA) molecules had been perceived as passive messenger molecules between genes encoded in DNA (with exception of RNA viruses) and the ribosomes producing proteins based on them. This view started to change when novel RNAs with distinct catalytic functions were discovered in the 1980s (Cech and Bass, 1986). Today, we know of a number of different types of RNAs involved in various processes: protein synthesis, post-transcriptional modification, DNA replication, gene regulation and parasitism (Atkins, et al., 2011). The key to the various functions of these diverse RNAs lies not only in their primary nucleotide sequence, but largely in their secondary and tertiary structures. The secondary structure results from the pairing of the ribonucleic bases that gives rise to stems (stretches of paired bases) and loops (unpaired bases between stem strands). Further, the tertiary structure is based on the scaffolds provided by the secondary structure, further stabilized by a presence of metal ions or the hydrogen bonds. Having recognized the importance and dynamics of RNAs in the cells, there has been great interest in predicting these structures.

De-novo structure prediction would seem to be intractable at first glance, as each residue can theoretically take up to seven torsion angles with the ribose ring (Das and Baker, 2007). Hence the number of possible conformations reaches astronomical levels, similarly to the problem of protein folding. In practice, there is a set of constraints that are imposed on folded RNA molecules that permits researchers to define folding algorithms based on, for example, physical models minimizing the free energy. However, the free energy minimization algorithms produce results correct in only about 70% of cases (Mathews, et al., 2010). This is due to the fact that RNA molecules are folded and function in a cellular environments with different pH levels, RNA chaperones (Rajkowitsch, et al., 2007), and other non-neutral molecules. Therefore, the minimal-energy predicted RNA molecule is often not the same as the functional one. To address these concerns and limit the folding space, researchers invented a number of experimental RNA structure probing procedures such as selective 2’ hydroxyl acylation analysed by primer extension (SHAPE) (Rice, et al., 2014) or similar. In recent years, with the advent of next-generation sequencing, these SHAPE assays were massively parallelized, which allowed researchers to

obtain high yields of secondary structure information (Poulsen, et al., 2015). Therefore, it has become a best practice to include these diverse structure probing sets in the predictive RNA structure models in a step towards predicting RNA molecules genome-wide for our better understanding of molecular evolution and diseases.

References

Atkins, J.F., Gesteland, R.F. and Cech, T. (2011) *RNA worlds : from life's origins to diversity in gene regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Carreiro, A.V., *et al.* (2015) Integrative biomarker discovery in neurodegenerative diseases, *Wiley interdisciplinary reviews. Systems biology and medicine*.

Cech, T.R. and Bass, B.L. (1986) Biological catalysis by RNA, *Annual review of biochemistry*, **55**, 599-629.

Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 14664-14669.

Felgueiras, J., Silva, J.V. and Fardilha, M. (2014) Prostate cancer: the need for biomarkers and new therapeutic targets, *Journal of Zhejiang University. Science. B*, **15**, 16-42.

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

Kalia, M. (2015) Biomarkers for personalized oncology: recent advances and future challenges, *Metabolism: clinical and experimental*, **64**, S16-21.

Mathews, D.H., Moss, W.N. and Turner, D.H. (2010) Folding and finding RNA secondary structure, *Cold Spring Harbor perspectives in biology*, **2**, a003665.

Poulsen, L.D., *et al.* (2015) SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data, *Rna*, **21**, 1042-1052.

Rajkowitsch, L., *et al.* (2007) RNA chaperones, RNA annealers and RNA helicases, *RNA biology*, **4**, 118-130.

Rice, G.M., Leonard, C.W. and Weeks, K.M. (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE, *Rna*, **20**, 846-854.

Sethi, S., *et al.* (2013) Clinical advances in molecular biomarkers for cancer diagnosis and therapy, *International journal of molecular sciences*, **14**, 14771-14784.

Chapter 2: Hypotheses and aims of dissertation

DNA methylation

In our DNA methylation studies, we hypothesize that:

- DNA methylation patterns, being the long-term mechanism for encoding temporal epigenetic changes, have sizeable role in development of many human diseases including cancer;
- DNA methylation at specific CpG sites can be used to define effective biomarkers for the diseases under study.

Based on this, our aims were to:

- to study DNA methylation in prostate and laryngeal cancers, in Klinefelter syndrome and in hormonal regulation of genes;
- to identify biomarkers of diagnostic and prognostic value for prostate cancer.

Integrative analysis of gene expression and DNA methylation

In our integrative studies of gene expression and DNA methylation, we hypothesize that:

- changes in methylation at CpG sites have context-specific interpretation by transcriptomic machinery and hence are interdependent with changes in gene expression;
- this interdependency can be effectively accounted for when analysing both data types, what may lead to identification of new and integrative biomarkers characterized by improved predictive power over single data type biomarkers.

Based on this, our aims were to:

- to define and evaluate an integrative model of gene expression and DNA methylation.
- to analyse a big breast cancer cohort in search for integrative biomarkers of cancer development and progression;
- to define a parameter-sparse implementation of the integrative model of gene expression and DNA methylation that would facilitate small cohort analyses;
- to analyse the same big breast cancer cohort with the parameter-sparse implementation and to define new set of biomarker candidates. We would also compare the merits of using both standard and parameter-sparse implementations of the proposed integrative model.

Integrating probing datasets with secondary RNA structure prediction methods

In our study of RNA secondary structure probing datasets, we hypothesize that:

- probing data can be effectively integrated with sequence-only RNA secondary structure predictive methods to increase their predictive accuracy;
- the use of probing data can be largely automated to allow rapid integration of diverse structure probing sets made with different chemical agents or protocols.

Based on this, our aims were:

- to use probabilistic graphical models to develop and evaluate a method for integration of diverse structural probing datasets with sequence-only RNA secondary structure prediction method using the stochastic context-free grammars;
- to demonstrate the model's ability to readily integrate diverse structure probing data sets by informing the base predictive model with SHAPE. DMS and CMCT probing sets.

Chapter 3: Presentation of methodologies

DNA methylation profiling with 450k microarray

The Infinium HumanMethylation450 BeadChip® from Illumina (Bibikova, et al., 2011) represents a powerful approach for profiling methylation status genome-wide. The platform measures genome-wide DNA methylation at >482,000 CpG loci overlapping 99% of RefSeq genes (21,231) and 96% of CpG islands (26,658). The methylation level at each CpG site is measured by two types of probes: one measures the intensity of methylated signal, while the other probe measures the un-methylated signal. The methylation level is then often reported as Beta-value, the ratio between the intensity of methylated signal to the sum of methylated and un-methylated signals, and hence takes values in the range between 0 and 1, were 0 signifies no methylation, 0.5 signifies methylation of 50% of sites in the sample, and 1 full methylation.

Processing of 450k data

The probes in the 450k platform were designed using two chemical assays: Infinium I and Infinium II (135,501 and 350,076 sites, respectively). These two slightly different assays are characterized by different signal resolution (Fig. 1 A), which should be corrected for to ensure comparability between probes of different designs. We correct for this lower resolution of Infinium II probes using a peak-based correction procedure proposed by (Dedeurwaerder, et al., 2011) that effectively scales type II probe peaks to match those of type I.

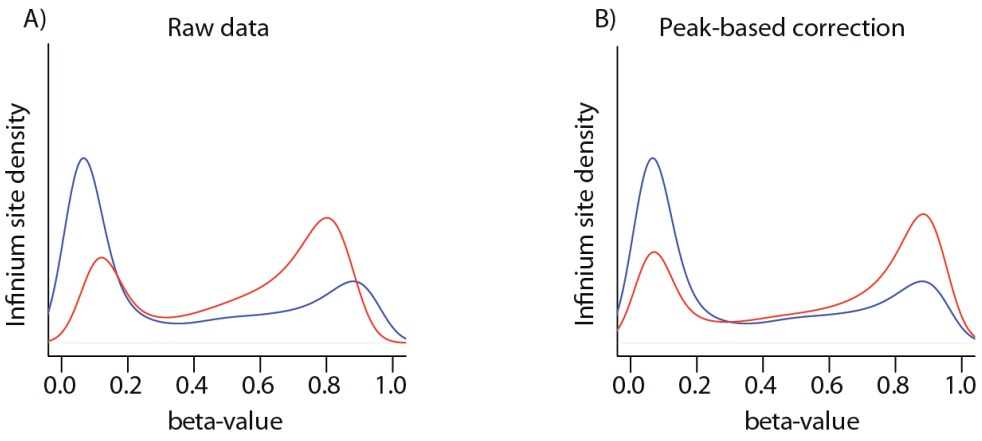


Fig. 1 Density plots of the beta-values for the two Infinium assay types considered (blue: Infinium I; red: Infinium II) before (A) and after (B) peak-correction procedure.

Many high-throughput statistical analysis methods assume the data is *homoscedastic*, i.e., the variance is approximately constant in the entire spectrum of the values the data can take. The violation of this assumption, which is described as *heteroscedasticity* in statistics, imposes serious challenges when applying these methods to high-throughput data. Beta-value, being the default methylation metrics for 450k platform, has significant heteroscedasticity in the low (<0.2) and

high (>0.8) methylation range as demonstrated by (Du, et al., 2010) (Fig. 2 A). This is resolved after transforming Beta-value to M-value, a widely used metrics for microarray data analysis, as suggested by (Du, et al., 2010). The transformation is done by calculating logit of Beta-value (Equation 1), and effectively makes the set approximately homoscedastic (Fig. 2 B), providing a better basis for downstream analyses that assume it. Therefore, M-values are preferred for differential analysis (and used for this purpose throughout in this thesis) while Beta-values are preferred for biological interpretation as they represent a fraction of methylated sites in the sample.

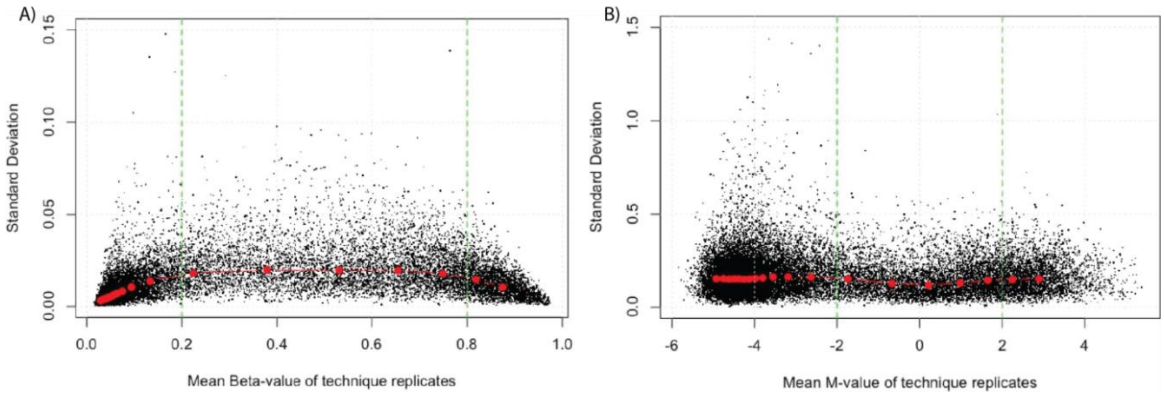


Fig. 2 The mean and standard deviation relations of technical replicates for beta-value (A) and M-value (B). Modified from (Du, et al., 2010).

$$M = log_2\left(\frac{Beta}{1 - Beta}\right);\ Beta = \frac{2^M}{2^M + 1}$$

Equation 1 Relationship between M- and Beta- values.

Differential methylation analysis

Having pre-processed the data, one may proceed to differential methylation analysis. Pne of the most powerful methods for this task was proposed by authors of *limma* (Ritchie, et al., 2015). Originally designed for analysis of differential gene expression in microarrays (Smyth, 2004), it uses a hierarchical model with an empirical prior on the variance, effectively reducing the number of parameters to be estimated for each region/site in comparison to competing approaches. It translates to using a moderated t-statistic with augmented degrees of freedom. *Limma* has been demonstrated to be very powerful in analysis of numerous datasets with many practical advantages over other methods (Jeanmougin, et al., 2010), and became the statistical framework of choice for most of our methylation analyses.

One of the alternatives to *limma*, often used in the cancer research field, is the Welch's t-test (Welch, 1947). This test is an extension of the classical t-test to populations with different variances. Welch's t-test is especially suitable for cancer cohort analyses as cancers typically

exhibit much higher variance than the control normal tissue samples (examples of that in *Chapter 4: Ongoing methylation studies*). Although we prefer using *limma* to Welch's t-test, it was not always possible to use the preferred method for all of our analyses, for example for analysis of synthetically generated data sets (Manuscript 1).

At last, we control for the False Discovery Rate (FDR) of performed statistical analyses using the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995). This multiple testing correction technique is suitable for exploratory studies like ours, as it is less conservative than competing methods that produce high rates of type II errors (false negatives) like Bonferroni, Holm-Bonferroni or Sidak corrections. Generally, lower number of false negatives is more desired than lower number of false positives in genome-wide screens as results are always validated through various robust laboratory techniques.

Hypothesis-free exploration of the data

It is often desirable to perform data mining to phrase new and sometimes unexpected hypotheses from the studied data. Indeed, such hypothesis-free exploration of 450k data was performed in many instances throughout projects described in this thesis. In particular, most variable CpG sites located in particular types of genomic regions can be used for such exploratory analysis, ranking them for example by the cohort-wise standard deviation. Thereafter, visual inspection of samples can be performed with the aid of the exploratory techniques, for example correlating known clinical variables with the patterns formed by samples. One such aiding technique is clustering (Hartigan, 1975), another is multidimensional scaling (Gower, 1966), both being used routinely in our methylation studies.

The most used type of clustering in genome-wide studies is the 2-way variant with clustering of features first (CpG sites, genes, transcripts) and by samples second. In short, pairwise Euclidean distances (dissimilarities) are calculated from the features in the multi-dimensional space, and then a linkage method is used to find the hierarchical structure. The results of such double clustering can be conveniently presented using a heatmap (a false colour image) with both clustering dendrograms (see examples in Chapter 4: Ongoing methylation studies). The heatmap helps with visual identification of subsets of CpG sites modulated similarly by the clinical variable.

To estimate the relative dissimilarity between samples, one could look at and sum the heights of branches separating the leaves (representing samples, for instance) in constructed dendrograms, or use a different visual aiding technique, the Multidimensional Scaling (MDS). MDS, aka principal coordinate analysis (Gower, 1966), takes a set of dissimilarities, and returns a set of points in n-dimensional space (2-dimensional, in most cases) such that the distances between the points are approximately equal to the dissimilarities. It is a dimensionality reduction technique that yields

results similar to clustering – except the dissimilarities are shown using XY plots, instead of with the hierarchy. It is a similar method to the more often used Principal Component Analysis which casts features into lower dimensions (Mardia, et al., 1979).

These hypothesis-free analyses can give researchers an idea about the amount of variability in their data, potentially reveal disease molecular subtypes, identify which sites are driving separation of samples when correlated with known clinical variables or merely can serve as a technical validation of sample labelling, ensuring mismatches are not present.

Regional methylation analysis

In some cases, it is preferable to perform the above analyses on regions, as opposed to individual CpG sites. In such case, a mean is typically derived from probes encompassed by defined regions, for example using Illumina's gene-centric annotations: TSS1500, TSS200, 5'UTR, 1st EXON, GENE BODY and 3' UTR. On average, each gene's region has a number of CpGs measured by the 450k technology (Fig. 3).

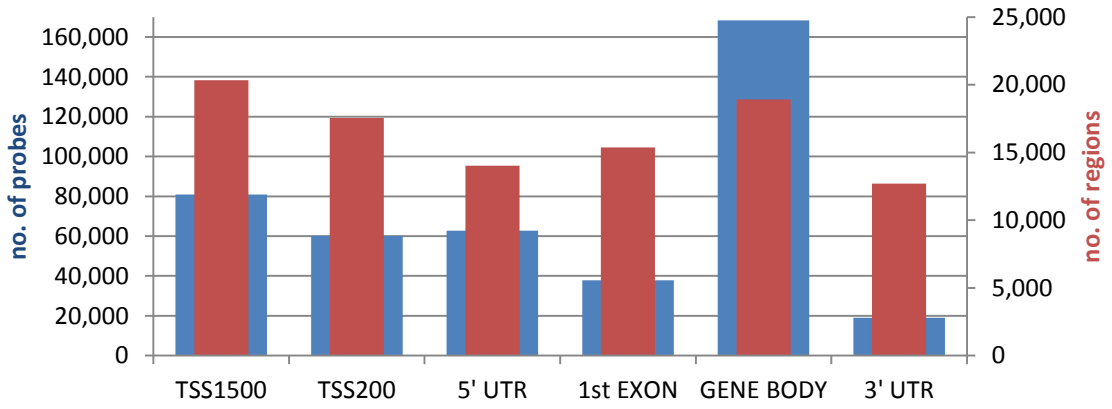


Fig. 3 Illustration of numbers behind the probes (blue bars, scale on the left) vs regions (red bars, scale on the right) for various gene-centric categories as defined for 450k platform.

We have also developed our own regional methylation analysis method with focus on gene body (3'UTR and GENE BODY categories) and promoter regions (TSS1500, TSS200, 5'UTR and 1st EXON categories), as described in Manuscript 1. In short, the method models the regional methylation with Gaussian kernels (assuming constant level of technical variance as demonstrated by (Du, et al., 2010)) and evaluates differences between tested groups non-parametrically using a random expectation of the likelihood ratio test statistic (Neyman, 1933).

Regional DNA methylation analysis could potentially help elucidate signal from noisy data such as in case of cancer, as the CpG sites within the same region are expected to have correlated methylation levels. On the other hand, if the annotation is inaccurate, the signal might get diluted. In practice, both regional and individual CpG site analyses are performed in parallel.

Transcriptome profiling with RNA sequencing

With the advent of Next Generation Sequencing (NGS), transcriptome researchers shifted from using microarrays to RNA sequencing (RNA-seq) (McGettigan, 2013). Compared to microarrays, RNA-seq technique can detect a larger number of transcripts and determine the transcript sequence at the same time (Marioni, et al., 2008; Wang, et al., 2009). By design, microarrays are limited to known transcripts only, while RNA-seq is not.

The RNA-seq technique relies on sequencing of the complementary DNA (cDNA) molecules synthesized from mRNA using the reverse transcriptase enzyme. To maximize yields of biologically relevant transcripts, samples are either depleted from ribosomal and bacterial RNAs or poly-A purified. One of the popular variants of this technique involves preparing paired-end libraries (as in case of BRCA dataset studied in Manuscript 1 and 2, and generally used at MOMA), which facilitates detection of novel transcripts (if desired) and improves estimation of transcript abundances (Illumina, 2015).

Once the library is sequenced, reads are mapped to the reference genome using a splicing-aware aligner such as *TopHat* (Trapnell, et al., 2009) or *RSEM* (Li and Dewey, 2011). After that, mapped reads are summarized by the feature of interest, often by gene or by gene splice variants. This could be done using stand-alone tools such as *HTSeq* (Anders, et al., 2015), but can also be done from within bigger frameworks such as *RSEM* or *cufflinks* (Trapnell, et al., 2012). Worth noting at this point is the importance of the library depth and its relation to our ability to perform differential analysis. Generally, if we are interested in researching the differential splice variant usage, we should prepare bigger libraries (with more input RNA) and sequence them deeper (achieve higher average coverage) than those meant for differential gene expression.

Differential expression analysis

Once summarized, reads are ready for the differential expression analysis. At this point, RNA-seq can be viewed as sampling of transcripts. The sampling will yield more accurate estimates of expression levels if the read count is high and if the sum of all read counts (size of library) is large. Given the discrete nature of reads, the sampling of transcripts can be modelled using Poisson distribution (as in case of our own method described in Manuscript 1). However, to compare samples with different library sizes, one has to normalize estimates by the sum of all reads, and hence the whole process can be modelled using the Binomial model, since the normalization is conveniently taken care of. Variants of this model were proposed by authors of *edgeR* (Robinson, et al., 2010), *DESeq* (Anders and Huber, 2010) or *PoissonSeq* (Li, et al., 2012). In particular, these methods introduce a Gamma prior on the variance, since many transcripts are biologically over-dispersed, compared to the assumed technical variance. Our own approaches to

RNA-seq data (Manuscript 1 and 2) also deal with the over-dispersion by introducing empirical priors using Gaussian kernel or Beta distribution, in Manuscript 1 and 2, respectively.

Once the parameters of the above models are determined, the differential testing can proceed. In case of the *DeSeq* and *edgeR*, a test with strong parallels to the Fisher's exact test is applied, except assuming the negative binomial distribution of the data. In case of our method for differential expression (Manuscript 1), we evaluate it non-parametrically using a random expectation of the likelihood ratio statistic (Neyman, 1933).

At last, similarly to the 450k differential analyses, we control for the False Discovery Rate (FDR) of performed statistical analyses using the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995). The same considerations regarding type II errors as in case of the 450k analysis apply.

Integrative analyses

Data integration is a broad term that encompasses many different techniques. The focus of this thesis is, however, the integration of multiple data types for which the dependency structure is known, and specifically integration of DNA methylation with gene expression. The current literature contains many examples of active mechanisms through which gene expression can be modulated by changes to CpG methylation of promoters and gene bodies (Gelfman, et al., 2013; Raynal, et al., 2012; Sati, et al., 2012; Yang, et al., 2014) or even enhancers (Ong and Corces, 2012). In some genomic contexts, DNA methylation can be simply correlated with gene expression, without any active involvement in the transcription rate.

Typical analysis of these data types is performed separately for each type and then combination of results is performed, for instance by filtering of the top candidates. This natural yet suboptimal step, however, requires that given gene's data types are statistically significant when individually analysed. A better strategy involves combination of p-values, for instance using Fisher's method (Fisher, 1938), which, however, assumes independence between each individual tests. In this strategy, a statistic based on a product of p-values is computed and evaluated in the chi-squared distribution with two times the number of combined tests as degrees of freedom (Equation 2). The alternative hypothesis for this test is that at least one of the integrated tests is significant and hence could be suboptimal to use as in some cases it may under- or over-emphasize the significance of findings, especially when dependencies exist.

$$X^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$$

Equation 2 Fisher's method: p_i is the value of the i^{th} hypothesis test, k is the number of combined tests. The value of the test statistic is evaluated in X^2 distribution with $2k$ degrees of freedom.

Throughout this thesis, we used the Fisher's method for combination of *edgeR* differential gene expression and Welch's t-test differential methylation analyses, and contrasted this independence-assuming approach to our own integrative method (Manuscript 1). Ours is a different approach which is computationally and statistically more demanding and involves modelling the underlying dependencies between the multiple integrated data types. This task can be tractable when the dependency structure is known from the domain knowledge. Below, I will present a general framework in which we define integrative methods used in all manuscripts that this thesis is comprised of.

Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are a framework that allow dependent data types to be modelled while dealing with two important problems: uncertainty and complexity (Koller and Friedman, 2009). PGMs essentially comprise of two key ingredients: probability theory and graph theory. The graph theoretic side provides an appealing interface through which we can encode the independence assumptions between integrated data types. The probability theory, on the other hand, provides the interface for the input data as well as the glue for specifying the form of dependencies. The framework is general as many of the commonly proposed statistical models can be cast as PGMs (Koller, et al., 2007). Complicated models are built by combining simpler parts. Also, many efficient general-purpose algorithms exist for inference in PGMs and hence development of new models in this framework is very cost effective.

In terms of practicalities, we have made a factor graph library for implementing our PGMs. Factor graphs are specified as undirected bipartite graphs consisting of random variable nodes (typically represented by circles) and factor nodes (typically represented by squares). Factor nodes are associated with a potential, which is a function of the neighbouring variable nodes. In case of our implementations, we enforce that these potentials represent the conditional probability measures over the variables, so that it effectively corresponds to PGMs.

PINCAGE integrative model (Manuscript 1)

With the aim of integrating multiple levels of genomic data, we developed a gene-oriented probabilistic model of gene expression, promoter methylation, and gene body methylation. The model independently evaluates the gene expression, as well as its separate relationships with methylation of promoter and gene body regions (Fig. 4). Through these conditional probability

distributions, the model also evaluates the differences in the regional methylation levels. A model structure yields the following factorization of the joint probability distribution over data tuple D_g containing promoter methylation, gene body methylation and gene expression data for a given gene g across samples ($D_g = M_g^{P.CpG}, M_g^{GB.CpG}, R_g; r_.$) (See Fig. 4 for notation definition):

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

Equation 3 PINCAGE integrative model joint probability distribution.

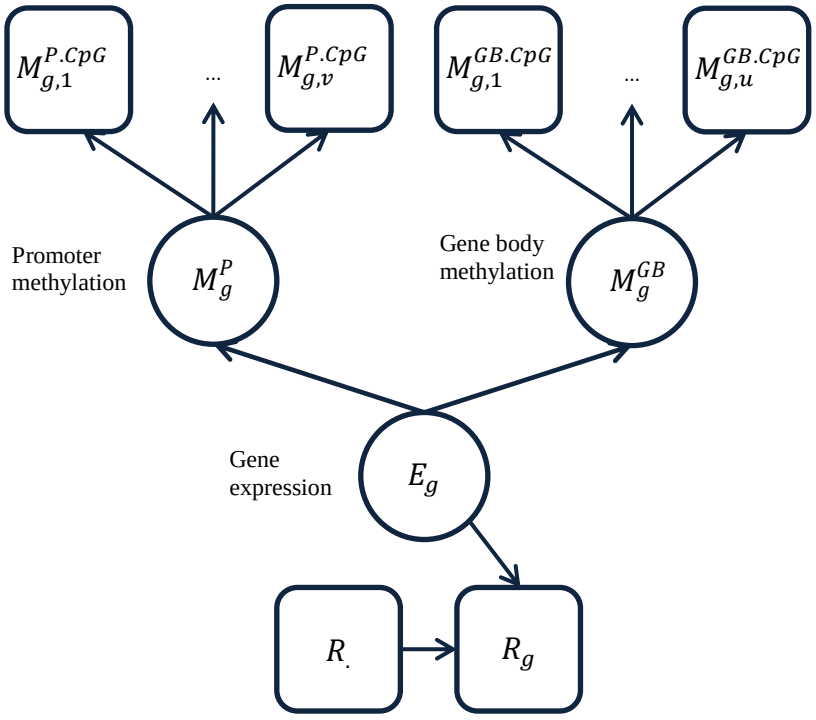


Fig. 4 Directed acyclic graph representation of PINCAGE integrative PGM. Variables in square boxes are directly observed while variables in circles are inferred: E_g is the expression of gene g, R_g is the observed read count for gene g, $R_.$ is the observed library size, $\boldsymbol{M_g^P}$ and $\boldsymbol{M_g^{GB}}$ are methylation levels of promoter and gene body regions of gene g, and $\boldsymbol{M_{g,1..v}^{P.CpG}}$ and $\boldsymbol{M_{g,1..u}^{GB.CpG}}$ are observed methylation levels of region-specific CpG sites of gene g.

In this model, the technical sampling of the data types is modelled as described for the data-type specific sub-models in respective sections of the thesis: with Poisson distribution for sampling of read counts (R_g), and with Normal distribution for sampling of CpG site methylation ($M_{g,1..v}^{P.CpG}$ and $M_{g,1..u}^{GB.CpG}$). In this setting, we model the gene expression as well as the link between the expression and methylation variables non-parametrically using Gaussian kernels (Polzehl and Spokoiny, 2006). Thanks to using these flexible specifications, we effectively tailor the method to the cancer setting, as we can capture the often-seen bimodalities in expression and methylation: for many considered genes, some cancer samples behave like normal tissue, while others are perturbed in various ways, likely reflecting on their molecular subtype (see examples in Manuscript 1). Equally important is the fact that we can also effectively capture the gene expression overdispersion seen for many transcripts. Gaussian kernels also allow us to capture the complex gene-specific and often non-linear relationships between the gene expression and methylation of the two considered regions.

Once the model parameters are fitted for each compared group of samples, we evaluate the significance of differences between them for a given gene using a variant of likelihood ratio test (Neyman, 1933). Specifically, instead of evaluating the test statistic in the chi-squared distribution, we calculate its random expectation by permuting sample labels between groups and evaluate it using an upper-tailed Z-test (Sprinthall, 2012). To predict the class label using PINCAGE, a posterior probability is normally calculated based on sample likelihoods under both fitted models. Alternatively, the sample likelihood ratio could be used as the discriminant function, as in case of Manuscript 2.

Among the weaknesses of this approach is the relative parameter richness of the integrative model, what makes it less suitable for analysis of smaller sample sets. Also, the calculation of the random expectation of the likelihood ratio statistic is CPU-time consuming and makes the genome-wide evaluation expensive to run (requiring a computer cluster to run on). Classification with the model, on the other hand, is fast. The runtime generally is strongly correlated with the variable number of CpG sites included in the model.

Sparse probabilistic model (Manuscript 2)

With the aim of facilitating robust analyses of smaller cohorts using our integrative PINCAGE model, we have developed a parameter-sparser alternative. The essential structure of the model remains the same (Fig. 4). What changes is the parametrization through which we encode the probability distributions over variables and their relationships. In particular, we sacrificed the possibility to model bimodalities of gene expression as well as of methylation. Also, we introduced a simplifying linear assumption about the relationship between gene expression and

methylation levels of the two modelled regions. In return, we achieve a sparser parameterization of the integrative model that could be reliably inferred from smaller sample sets.

In this implementation, we model the gene expression with beta-binomial model, capturing the technical variance with binomial distribution, and the overdispersion with beta distribution. The technical and biological variance of methylation variables are jointly modelled with normal distribution. Finally, the conditional relationship between gene expression and methylation of promoter and gene bodies is modelled by linear regression. This model specification induces the following joint probability distribution over a data tuple D, containing promoter methylation, gene body methylation and gene expression data for a given gene in a sample ($D = M^{P.CpG}, M^{GB.CpG}, R; r$):

$$P(D) = Binom(R; r, E)\, Beta(E; \alpha, \beta)\, \emptyset\left(M^{P}; a^{P}E + b^{P}, \sigma^{E,P^2}\right)\, \emptyset\left(M^{GB}; a^{GB}E + b^{GB}, \sigma^{E,GB^2}\right) \prod_{v=1}^{n^{P}} \emptyset\left(M_v^{P.CpG}; M_P, \sigma^{P^2}\right) \prod_{v=1}^{n^{GB}} \emptyset\left(M_v^{GB.CpG}; M_{GB}, \sigma^{GB^2}\right)$$

Equation 4 Sparse integrative model joint probability distribution for a single gene.

This time, the parameters of the model $\theta = (\alpha, \beta, \sigma^{E,P^2}, \sigma^{P^2}, a^{P}, b^{P}, \sigma^{E,GB^2}, \sigma^{GB^2}, a^{GB}, b^{GB})$ are inferred using an Expectation-Maximization (EM) algorithm (Gupta and Chen, 2010), except for the α and β parameters of beta distribution that were converging slowly and are found outside of the framework using gradient descent algorithm (Venables, et al., 2002).

Once the parameters of the models for the compared groups are found, a likelihood ratio for each sample can be calculated and used as the discriminant function. These scores are then used in a cross-validation manner to find best discriminating genes based on the training AUC. Alternatively, other strategies for identification of integrative biomarker candidates could be used, for instance based on the model dissimilarity measures like Kullback-Leibler divergence (Kullback, 1951).

To summarise, among strengths of this integrative model implementation is its parameter sparseness, which permits robust inference for relatively smaller cohorts of samples. The sparseness also correlates with faster model runtime and therefore is more cost-effective than PINCAGE. However, an undoubtful weakness is the simplistic representation of some of the probability distributions, and especially the linearity in the relationship between methylation and gene expression. This may affect the goodness of fit of the models.

ProbFold (Manuscript 3)

With the aim of integrating multiple structure-probing datasets for improved RNA secondary structure prediction, we developed a general framework using PGMs for augmenting sequence-only Stochastic Context-Free Grammars (SCFGs) (Grate, 1995).

SCFGs are probabilistic variants of Context-Free Grammars, and define the basal model describing the RNA molecule's secondary structure (Fig. 5). With the help of PGMs, we extend the basal model (defining a prior over secondary structure) to incorporate experimental data using a set of emission models (Fig. 6 A,B,C): *single*, *pair* and *stack*. *Single* models only the single sequence position, *pair* models a pair of sequence positions (nucleotide pairing), while *stack* models four sequence positions comprised of two consecutive pairs. For each of the three emission models, the joint probability distribution is given in (Fig. 6 D). The parameters of the basal and emission models are found by the EM algorithm (Gupta and Chen, 2010), using the Inside-Outside (Durbin, 1998) and the Sum-Product (Bishop, 2006) algorithms at the E step for SCFGs and PGMs, respectively.

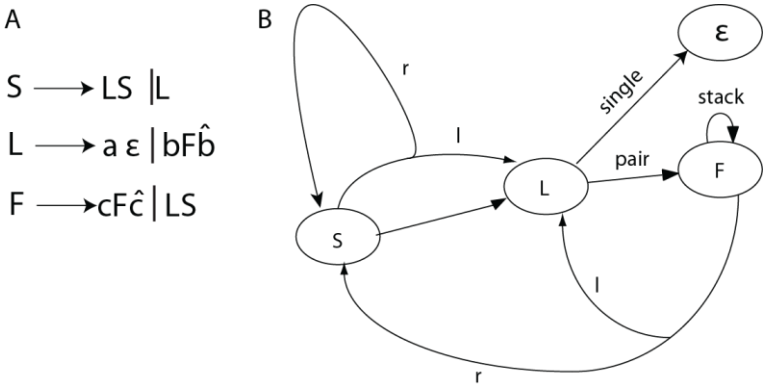


Fig. 5 (A) Grammar rules (see (B) for variable definitions). (B) Pictorial representation of the stacking grammar. The grammar has six production rules involving the four non-terminals S, L, F, and ε, which cannot be derived further. Three of the rules emit terminals, named *single*, *pair* and *stack*, and three are non-emitting, including two bifurcation rules. Each bifurcation rule splits into two parts, consisting of a left (l) non-terminal and right (r) nonterminal. The derivation starts in S. S can use either a bifurcation rule, which transits to L (l-part) as well as back to itself (r-part), or a non-emitting rule, which transits to L. L can use either the *single* emitting rule, which transits to ε and emits unpaired terminals (a), or use the *pair* rule, which transits to F and emits paired terminals (aâ). Finally, F can use the stack emitting rule, which transits back to F and emits (stacked) paired terminals (b**b̂**) dependent on the previous base pair, or a bifurcation rule, which transits to L (l-part) as well as to S (r-part).

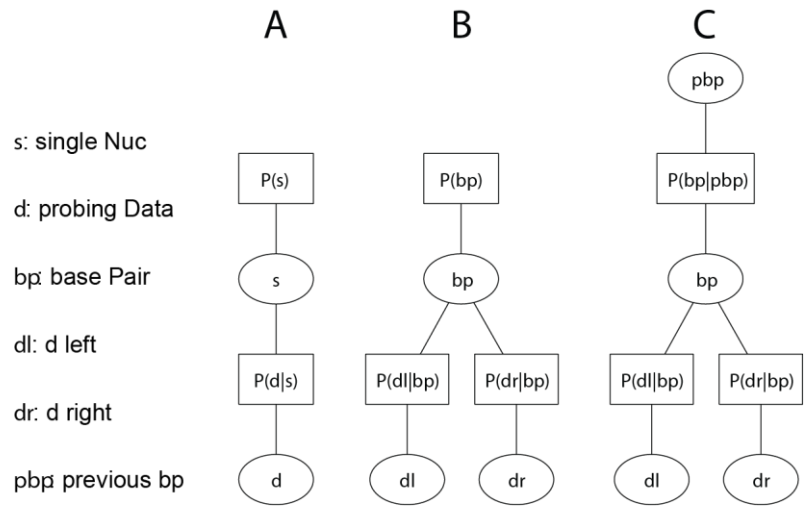
As we implemented ProbFold using our factor graph library, which, at the time, strictly operated on discrete variables only, the efficiency of signal extraction from the probing data depended on the differences between the inferred discrete distributions. The latter were correlated with the

number of discrete bins and their boundaries. Hence, an optimal bin positioning, maximizing the differences between compared distributions, was desired. An alternative approach would have been to increase the overall number of bins, but it was not preferred as the amount of training data was relatively small and would likely lead to overfitting. Hence we implemented a greedy bin boundary search approach based on the Kullback-Leibler divergence (Kullback, 1951). The Kullback-Leibler (KL) divergence (D_{KL}) optimality criterion can be defined in our case as:

$$D_{KL}\left(p^{single}||p^{pair}\right) = \sum_{i}^{k} p_i^{single} ln \frac{p_i^{single}}{p_i^{pair}}.$$

Equation 5 Kullback-Leibler divergence definition for discrete probability distributions.

The KL divergence measures the expected information content and, being closely related to likelihood ratio test, can be also thought of as the expected information to discriminate between the alternative hypotheses specified by the two distributions. Hence, we used this criterion in a greedy search to find the break points that optimize the $D_{KL}\left(p^{single}||p^{pair}\right)$ expression for desired number of bins. This procedure became a part of an automated pre-processing of the data that also included ranking and normalizing of the data from different experimental conditions.



D

	single	pair	stack
Joint prob.	P(s,d)	P(bp,dl,dr)	P(bp,pbp,dl,dr)
Factorization	P(s)P(d |s)	P(bp)P(dl|bp) P(dr|bp)	P(dl|bp)P(dr|bp) P(bp|pbp)
RNA Probing data	ACGUCACGU 001232454	ACGUCACG U 00123245 4	ACGUCAC GU 0012324 54
Mapping	s←A, d←0	bp← AU dl← 0, dr← 4	bp←CG , pbp←AU dl←0, dr ← 5

Fig. 6 Probabilistic graphical models defining the (A) single, (B) pair, and (C) stack emission models. The PGMs are shown as (bipartite) factor graphs, with variable nodes (circles) connected to factors (squares) defining local probability distribution. The variable abbreviations are given to the left. (D) For each emission model, the table gives (i) the joint probability distribution; (ii) its factorization specified by the PGM; (iii) example of short input data sequence with potential input positions highlighted. Note that the probing data has been discretized into six bins (0-5); (iv) mapping of data from highlighted sequence positions to relevant random variables of PGM.

Measuring binary classifier performance

In most cases presented throughout this thesis, we assessed the predictive performance of binary classifiers using the Receiver Operating Characteristic (ROC) analysis which is performed by drawing a plot of sensitivity versus false positive rate (one minus specificity) for a range of threshold values of the discriminant function (Fawcett, 2004). Following, the Area Under ROC Curve (AUC) is calculated for each classifier, which takes values between 0 and 1, where 0.5 corresponds to a random guess while 1 corresponds to a perfectly correct classification. The AUC of 0 corresponds to a perfect misclassification. In particular, we calculate the empirical AUCs (Fawcett, 2006) of analysed classifiers using the *pROC* R package (Robin, et al., 2011). It is straight-forward to calculate the empirical AUC as it can be interpreted as the probability of ranking a randomly drawn positive element higher than randomly drawn negative element

(Hanley and McNeil, 1982). This interpretation also corresponds to the Wilcoxon (Wilcoxon, 1945) and the Mann-Whitney (Mann and Whitney, 1947) test statistics.

Further, we compare two empirical ROC curves using a test proposed by (Delong, et al., 1988). Other alternatives for testing exist based on bootstrapping (Carpenter and Bithell, 2000) or smoothing of the curves (Venkatraman, 2000; Venkatraman and Begg, 1996).

The use of AUC was advocated for as being more discriminative then metrics such as accuracy, F-measure, positive predictive value and specificity (Ling, et al., 2003; Ling, et al., 2003). On the other hand, it was criticized for various reasons, including 1) summarizing performance over ROC regions in which one rarely operates, 2) weighing equally the errors of commission and omission, or 3) ignoring the goodness of fit of the models (Hand, 2009; Lobo, et al., 2008), amongst others. Also, it was demonstrated that AUC can be unreliable for small sample sizes (Hanczar, et al., 2010), so the classification metrics based on ROC should be treated with caution when dealing with small cohorts, for instance.

References

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

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data, *Bioinformatics*, **31**, 166-169.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing, *J Roy Stat Soc B Met*, **57**, 289-300.

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

Bishop, C.M. (2006) *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, *Statistics in Medicine*, **19**, 1141-1164.

Dedeurwaerder, S., *et al.* (2011) Evaluation of the Infinium Methylation 450K technology, *Epigenomics*, **3**, 771-784.

Delong, E.R., Delong, D.M. and Clarkepearson, D.I. (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.

Durbin, R. (1998) *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK New York.

Fawcett, T. (2004) ROC graphs: Notes and practical considerations for researchers, *ReCALL*, **31**, 1--38.

Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861-874.

Fisher, R.A. (1938) *Statistical methods for research workers*. Biological monographs and manuals,. Oliver and Boyd, Edinburgh,.

Gelfman, S., *et al.* (2013) DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure, *Genome Res*, **23**, 789-799.

Gower, J.C. (1966) Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis, *Biometrika*, **53**, 325-&.

Grate, L. (1995) Automatic RNA secondary structure determination with stochastic context-free grammars, *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **3**, 136-144.

Gupta, M.R. and Chen, Y. (2010) Theory and Use of the EM Algorithm, *Foundations and Trends® in Signal Processing*, **4**, 223-296.

Hanczar, B., *et al.* (2010) Small-sample precision of ROC-related estimates, *Bioinformatics*, **26**, 822-830.

Hand, D. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Mach Learn*, **77**, 103-123.

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29-36.

Hartigan, J.A. (1975) *Clustering algorithms*. Wiley series in probability and mathematical statistics. Wiley, New York,.

Illumina, I. (2015) Paired-end sequencing.

Jeanmougin, M., *et al.* (2010) Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies, *Plos One*, **5**, e12336.

Koller, D. and Friedman, N. (2009) *Probabilistic graphical models : principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

Koller, D.F., Nir, Getoor, L. and Taskar, B. (2007) Graphical Models in a Nutshell. In, *Introduction to Statistical Relational Learning*. MIT Press.

Kullback, S.L., Richard A. (1951) On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC bioinformatics*, **12**.

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

Ling, C.X., Huang, J. and Zhang, H. (2003) AUC: a better measure than accuracy in comparing learning algorithms. *Proceedings of the 16th Canadian society for computational studies of intelligence conference on Advances in artificial intelligence*. Springer-Verlag, Halifax, Canada, pp. 329-341.

Ling, C.X., Huang, J. and Zhang, H. (2003) AUC: a statistically consistent and more discriminating measure than accuracy. *Proceedings of the 18th international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc., Acapulco, Mexico, pp. 519-524.

Lobo, J.M., Jiménez-Valverde, A. and Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models, *Global Ecology and Biogeography*, **17**, 145-151.

Mann, H.B. and Whitney, D.R. (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, 50-60.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate analysis*. Probability and mathematical statistics. Academic Press, London ; New York.

Marioni, J.C., *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome Res*, **18**, 1509-1517.

McGettigan, P.A. (2013) Transcriptomics in the RNA-seq era, *Current opinion in chemical biology*, **17**, 4-11.

Neyman, J. (1933) On the problem of the most efficient tests of statistical hypotheses, *Philos T R Soc Lond*, **231**, 289-337.

Ong, C.T. and Corces, V.G. (2012) Enhancers: emerging roles in cell fate specification, *EMBO reports*, **13**, 423-430.

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

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 research*, **72**, 1170-1181.

Ritchie, M.E., *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Res*, **43**, e47.

Robin, X., *et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC bioinformatics*, **12**, 77.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (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.

Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical applications in genetics and molecular biology*, **3**, Article3.

Sprinthall, R.C. (2012) *Basic statistical analysis*. Pearson Allyn & Bacon, Boston.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105-1111.

Trapnell, C., *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature protocols*, **7**, 562-578.

Venables, W.N., Ripley, B.D. and Venables, W.N. (2002) *Modern applied statistics with S*. Statistics and computing. Springer, New York.

Venkatraman, E.S. (2000) A Permutation Test to Compare Receiver Operating Characteristic Curves, *Biometrics*, **56**, 1134-1138.

Venkatraman, E.S. and Begg, C.B. (1996) A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment, *Biometrika*, **83**, 835-848.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews. Genetics*, **10**, 57-63.

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

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, **1**, 80-83.

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.

Chapter 4: Summary of main results

Manuscript 1

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

Michał P. Świtnicki, Malene Juul, Tobias Madsen, Karina D. Sørensen, and Jakob S. Pedersen

Manuscript in review at *Bioinformatics*

In this publication we set out to study the genome-wide methylation-expression relationship in the cancer setting and define a probabilistic model for identification of integrative biomarker candidates based on 450k methylation and RNA-seq gene expression data. The study was performed using Breast Invasive Carcinoma (BRCA) data set from The Cancer Genome Atlas Network (TCGA) (Cancer Genome Atlas, 2012), consisting of 730 tumour and 82 adjacent normal samples.

To motivate our integrative model designs, we first performed a detailed evaluation of the variability and correlations between methylation of promoter and gene body regions and the gene expression (Fig. 7). We found significant tumour heterogeneity when comparing with adjacent normal tissue (Fig. 7 B) that supported our further use of flexible Gaussian kernels for modelling population distributions. We found that permitting bimodality in the specification of probability distributions is especially important for retaining the statistical power, as modelling mere overdispersion does not properly reflect the bimodal phenomena of cancer samples seen for some of the genes. We also evaluated the correlations between gene expression and the methylation (Fig. 7 C,D) and saw significant changes in its degree when comparing tumours and adjacent normal samples, which supported building separate models for each group. Additionally, some gene case studies revealed the often non-linear nature of relationship between gene expression and the methylation.

An important result of this manuscript was the definition of the probabilistic graphical model integrating the methylation and gene expression data described in the Chapter 3. Using this model to evaluate the studied BRCA data set, however, gave rise to another set of results. At first, we applied our method to genome-wide comparison between tumour and adjacent normal samples and found very high percentage of all genes being significantly perturbed (>91%). The same high perturbation was seen even when analysing this dataset using established methods described in this thesis, combined with the Fisher's method. The top-ranking list (Table 1) contains genes both unlinked and previously linked to breast and other cancer types. Overall, discrimination between tumour and normal samples is generally not difficult, for instance based on histopathological

analysis, however, identification of extreme perturbation of genes previously not linked to cancer, and especially of large intergenic non-coding RNA (lincRNA) LOC148145, point to interesting leads of cancer development.

We further applied our integrative model to comparison between progressing and non-progressing tumours, defined based on the recurrence of cancer after initial treatment. In this comparison, a much smaller number of genes was found significantly perturbed by our method (n=95), as well as by established methods combined with Fisher's (n=234). The most robust genes in the cross-validation procedure were the Zinc Finger Protein 706 (ZNF706) and Serpin Peptidase Inhibitor Member 3 (SERPINE3). ZNF706 was previously linked to Laryngeal Squamous Cancer (Colombo, et al., 2009), while SERPINE3, although belonging to a family of proteins playing role in brain localization of breast cancer metastases (Valiente, et al., 2014), was previously not directly associated with breast cancer. Comparison of the predictive performance of our model with corresponding Logistic Regression (LR) classifiers (assuming independence) for these two integrative biomarker candidates, showed significant improvement in predictive power (0.8358 vs 0.7895 AUC). Further LR classifiers showed erratic AUCs ranging from 0.4091 at the fifth gene to 0.8860 at the ninth gene, while our model combinations remained relatively stable, suggesting that the LR combined classifiers were less robust.

Table 1 Integrative PINCAGE model top-10 most significantly perturbed genes in BRCA and their ability to classify tumour and normal samples. For comparison, the right-most column contains top-10 most significant genes according to Fisher's method applied to established methods * signifies known role in cancer. ** signifies known role in breast cancer.

Significance evaluation of BRCA data set (55 AN's vs 487 T's)				**Classification performance on BRCA validation subset (27 AN's and 243 T's)**			
	Integrative PINCAGE		**Established methods combined**	**Integrative PINCAGE**		**Logistic regression using PINCAGE-identified genes**	
Gene ID	Z-score	Rank (k)	Rank	AUC of single gene model	AUC using running combination of genes (1-k)	AUC of single gene model	AUC using running combination of genes (1-k)
RAG1AP1*	115.70	1	773	0.9311	0.9311	0.9813	0.9813
CPA1*	114.92	2	96	0.9297	0.9747	**0.9960**	**0.9989**
NEK2**	112.56	3	446	0.9291	0.9927	0.9720	0.9986
RNASEH2A **	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	0.9962
TIMM17A**	102.79	7	1664	0.9576	0.9977	0.9497	0.9198
PLK1**	99.95	8	496	0.9427	0.9970	0.9709	0.9290
RABIF*	98.58	9	1441	0.9531	0.9988	0.9694	0.9156
PTF1A*	98.45	10	1577	**0.9806**	0.9988	0.9561	0.9070

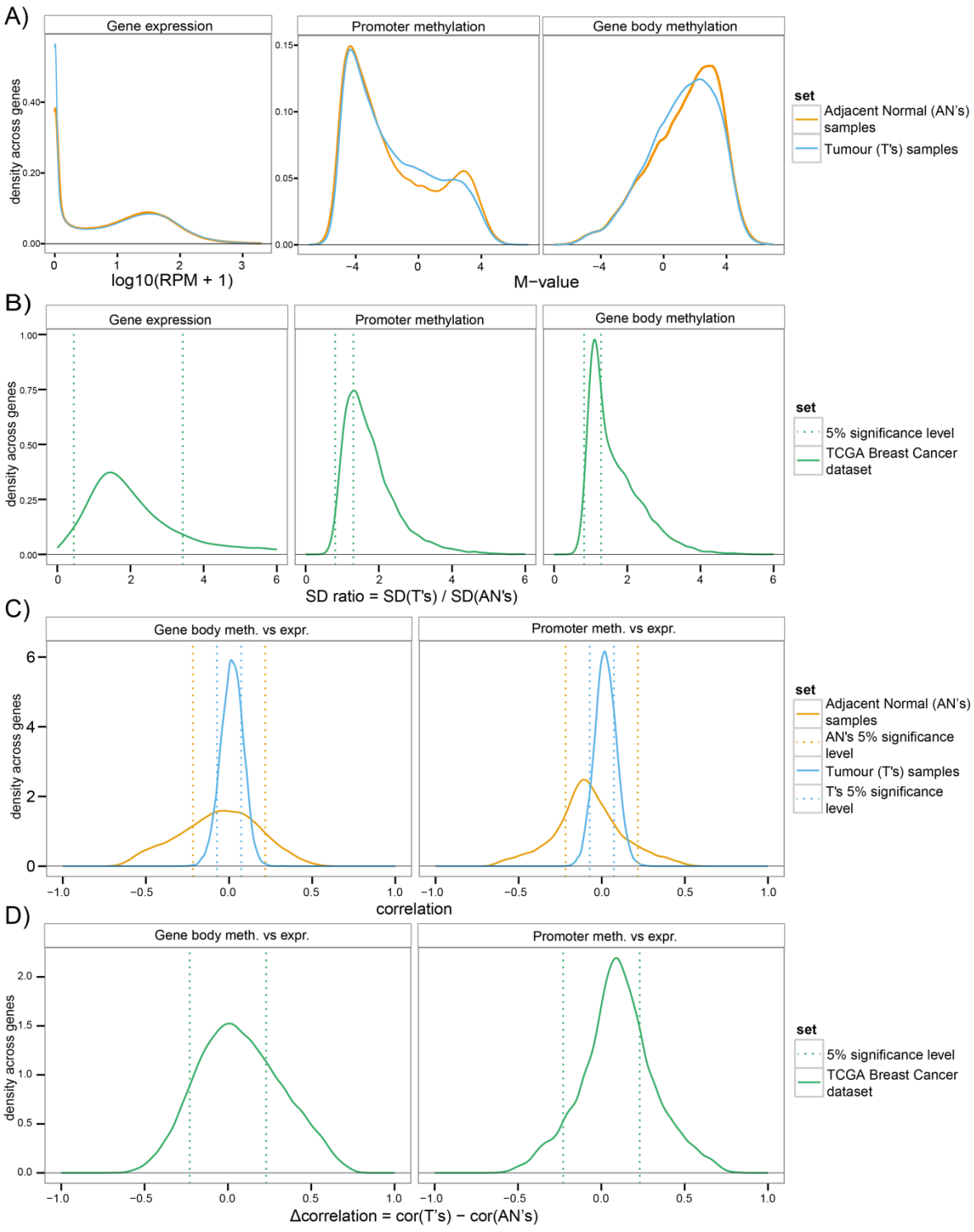


Fig. 7 Expression and methylation profiles in BRCA. A) Global distributions of expression levels, measured in reads per million (RPM), and mean methylation levels (M-value) across promoter and gene body regions for both groups across samples. B) Distribution of gene-wise standard deviation ratios between 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 data set for AN's and T's. D) Gene-wise changes of correlations observed between the AN's and T's.

Table 2 Left: Top-10 ranked genes in the BRCA progression data set. **Right**: Comparison of classification performance for integrative PINCAGE, logistic regression on PINCAGE-identified genes, and logistic regression on genes found by combination of established methods with Fisher's method.

Significance evaluation of progression set (14 progressing and 57 non-progressing tumours)			
	Integrative PINCAGE		**Established methods combined**
Gene ID	Z-score	Rank	Rank
SERPINE3	11.46	1	251
ZNF706	8.75	2	752
ACTN2	6.90	3	1518
AKR1B15	6.75	4	714
AGBL3	6.47	5	5645
LOC100240734	6.19	6	931
MYL10	6.13	7	5869
NDUFA9	6.04	8	9953
HIGD1B	5.84	9	311
ARG1	5.74	10	614

Classification performance on progression set using 14-fold cross-validation						
	Integrative PINCAGE		**Logistic regression using genes found by integrative PINCAGE**		**Logistic regression using genes found by combination of established methods**	
Rank at each fold (k)	AUC of single gene model	AUC using running combination of genes (1-k)	AUC of single gene model	AUC using running combination of genes (1-k)	AUC of single gene model	AUC using running combination of genes (1-k)
1	**0.8008**	0.8008	**0.7431**	0.7431	**0.7055**	**0.7055**
2	0.6316	**0.8358**	0.7043	0.7895	0.4624	0.6291
3	0.6629	0.6742	0.5990	0.7143	0.4912	0.6253
4	0.4818	0.7055	0.5689	0.7406	0.5564	0.5376
5	0.6216	0.6491	0.6654	0.4091	0.4950	0.4787
6	0.6685	0.6805	0.6967	0.7105	0.6190	0.5526
7	0.6291	0.6366	0.5338	0.7375	0.5714	0.5764
8	0.4524	0.6479	0.5426	0.8296	0.5175	0.6109
9	0.5188	0.6378	0.5927	**0.8860**	0.5815	0.5013
10	0.5188	0.6253	0.5025	0.7162	0.5414	0.5263

References

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

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

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

Manuscript 2

"Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data"

Michał P. Świtnicki, Tobias Madsen, and Jakob S. Pedersen

Manuscript in preparation

In this publication we set out to define an alternative and sparser parameterization of the integrative model proposed in Manuscript 1 (Świtnicki, et al., 2015) and evaluate it on the same Breast Cancer Adenocarcinoma (BRCA) (Cancer Genome Atlas, 2012) dataset, focusing on identifying new integrative biomarker candidates and improving the inference on smaller sample sets.

Having defined the parameter sparse implementation of the integrative model described in Chapter 3 of this thesis, we applied it in an 8-fold cross-validation setting for identification of most discriminating genes between adjacent normal (n=82) and tumour (n=730) samples. We used the training AUC as the candidate selection criteria at each fold. The top-10 list (Table 3) included four genes previously found implicated in cancer (TMEM132C, ULBP1, SLC6A2, A2BP1). Another four identified genes, despite being well characterized, were not previously associated with any cancer type (TMEM132D, CACNG3, FXYD1 and NRSN1). Interestingly, the final two genes, KIR3DX1 and LOC388692, were poorly characterized but encoded a pseudogene and a long non-coding RNA (lncRNA), respectively. Their identification points at the importance of these non-coding transcripts in the development of the cancer disease. Despite being expressed in low quantities, their methylation patterns were highly diagnostic, far greater than gene expression alone. Such aberrant methylation patterns could signify their differential splicing or, if unexpressed, their promoters could act in an enhancer-like trans-acting mechanism of regulation of other genes.

Table 3 Top-10 ranked genes in the evaluation of 82 normal and 730 tumour BRCA samples.

8-fold cross-validation analysis 82 normal and 730 tumour BRCA samples						
Top genes across folds			**Classification performance (AUC)**			
			Sparse integrative model		**Logistic Regression**	
Mean rank	**Gene ID**	**Rank (k)**	**Single rank**	**Combined (1-k)**	**Single rank**	**Combined (1-k)**
1.5	TMEM132D	**1**	0.9827	0.9827	**0.9916**	0.9916
2.1	TMEM132C	**2**	0.9900	0.9903	**0.9934**	**0.9956**
3.5	ULBP1	**3**	**0.9850**	0.9938	0.9786	**0.9944**
4.6	KIR3DX1	**4**	**0.9867**	0.9943	0.9637	0.9943
5.4	CACNG3	**5**	0.9705	**0.9914**	**0.9766**	0.9722
5.9	LOC388692	**6**	**0.9892**	**0.9923**	0.9617	0.9288

6.8	FXYD1	**7**	0.9565	**0.9925**	**0.9833**	0.9616
8.9	SLC6A2	**8**	**0.9815**	**0.9939**	0.9674	0.9524
10.6	NRSN1	**9**	**0.9629**	**0.9942**	0.8676	0.9639
10.8	A2BP1	**10**	**0.9717**	**0.9944**	0.9379	0.9782
		Average	**0.9777**	**0.9930**	0.9622	0.9713

We next applied the model in a challenging setting for comparison between progressing (n=14) and non-progressing (n=57) tumours. To maximize the number of training samples for the progression set at each fold, we applied our sparse integrative model in a 14-fold cross-validation procedure (Table 4). Again, we used the training AUC as the candidate selection criteria at each fold. The top-3 genes in this analysis consistently reappeared in the top-20 at each fold, suggesting them to be robust biomarker candidates. Interestingly, the list included the previously identified SERPINE3 gene at the top-2, validating our previous analysis using the initial parameter-rich implementation. The other candidate, KAAG1, was found implicated in many tumour types including breast cancer (Van Den Eynde, et al., 1999). ZFATAS, the final candidate biomarker, is a poorly characterized gene. However, it was classified as a lncRNA and again points at the potential importance of these types of transcripts for the progression of the cancer disease.

Table 4 Top-10 ranked genes in the evaluation of 14 progressing and 57 non-progressing BRCA tumour samples.

14-fold cross-validation analysis 14 progressing and 57 non-progressing BRCA tumours						
Top genes across folds		**Rank (k)**	**Classification performance (AUC)**			
			Sparse integrative model		**Logistic Regression**	
Mean rank	**Gene ID**		**Single rank**	**Combined (1-k)**	**Single rank**	**Combined (1-k)**
2.4	ZFATAS	**1**	**0.6165**	0.6165	0.5677	0.5677
4.6	SERPINE3	**2**	0.6391	**0.6867**	**0.7055**	0.6654
6.2	KAAG1	**3**	**0.6341**	**0.6591**	0.6165	0.6541
8.1	SFRS8	**4**	0.6278	**0.6842**	**0.6717**	0.6491
14.7	DPY19L3	**5**	**0.6880**	**0.6404**	0.5113	0.5959
14.9	LOC149620	**6**	0.4612	**0.6591**	**0.5564**	0.5426
17.8	ATP9A	**7**	0.6980	**0.6692**	**0.7406**	0.5213
18.4	IQGAP2	**8**	**0.5815**	**0.6692**	0.5689	0.6028
19.8	GPBAR1	**9**	0.5313	**0.6504**	**0.5677**	0.5915
21.4	TMEM198	**10**	**0.6253**	**0.6604**	0.5852	0.5946
		Average	**0.6103**	**0.6643**	0.6091	0.6019

References

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

Świtnicki, M.P., *et al.* (2015) PINCAGE: Probabilistic integration of cancer genomics data for perturbed gene identification and sample classification.

Van Den Eynde, B.J., *et al.* (1999) A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription, *The Journal of experimental medicine*, **190**, 1793-1800.

Manuscript 3

"ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction"

Sudhakar Sahoo, Michał P. Świtnicki, and Jakob S. Pedersen

Manuscript in review at *Bioinformatics*

In this publication we set out to define and evaluate a probabilistic graphical model integrating diverse probing data sets with the stochastic context-free grammars (SCFGs), a probabilistic formal language capable of capturing and modelling the nested interactions of secondary RNA structure. The integrated probing sets are based on the Selective 2' Hydroxyl acylation Analysed by Primer Extension (SHAPE) method and extended to other types of probing data. In particular, we analysed E. coli 16S and 23S ribosomal RNA (rRNA) SHAPE data from (Deigan, et al., 2009; Weeks, 2012), and augmented it with several small RNA structures from dimethyl sulphate (DMS) and 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene (CMCT) probing variants (Cordero, et al., 2012; Rice, et al., 2014).

At first, we evaluated if the SHAPE reactivities were correlated with the primary sequence of nucleotides (Fig. 8). Looking at both single and paired regions separately (Fig. 8 A,B), we found significant differences in distributions of the SHAPE values across different nucleotides (p-value=4.4e-03 for single and p-value=8.6e-06 for pair; Kruskal-Wallis rank sum test (Kruskal and Wallis, 1952)). The correlation of reactivity within these base pairs was surprisingly not very significant (Fig. 8 C, Pearson's correlation coefficient cc=-0.042, p-value=0.075), what could be explained by the significant experimental noise for this data set, or by the fact that SHAPE reactivities are generally low for base-paired positions and correlations are therefore hard to detect. On the other hand, correlations of SHAPE reactivities between neighbouring single and paired nucleotides (Fig. 8 D,E) were high (cc=0.559, cc=0.397, respectively) and significant (p-values < 1.0e-05, Pearson's test). Based on these findings, we extended the emission models incorporating the probing sets to capture these sequential correlations in the SHAPE data.

We further applied a series of models, gradually integrating more data sets from different SHAPE variants, and recorded the improvements over structure-only models using ProbFold, and compared the performance gains against a standard in the field, the *RNAstructure* (Mathews, et al., 2004). We assessed the changes in performance using the F-measure, which is a harmonic mean of the sensitivity and positive predictive value (Hand, 2012). While the *RNAstucture* performs much better using the sequence-only information than ProbFold's basal model based on SCFGs, and achieves the highest overall F-values (Table 5), the gradual incorporation of the

different probing datasets is done best by ProbFold, showing the highest ΔF values, improving significantly with each addition. This shows that ProbFold emission models defined for each incorporated dataset make good use of the available structure signal.

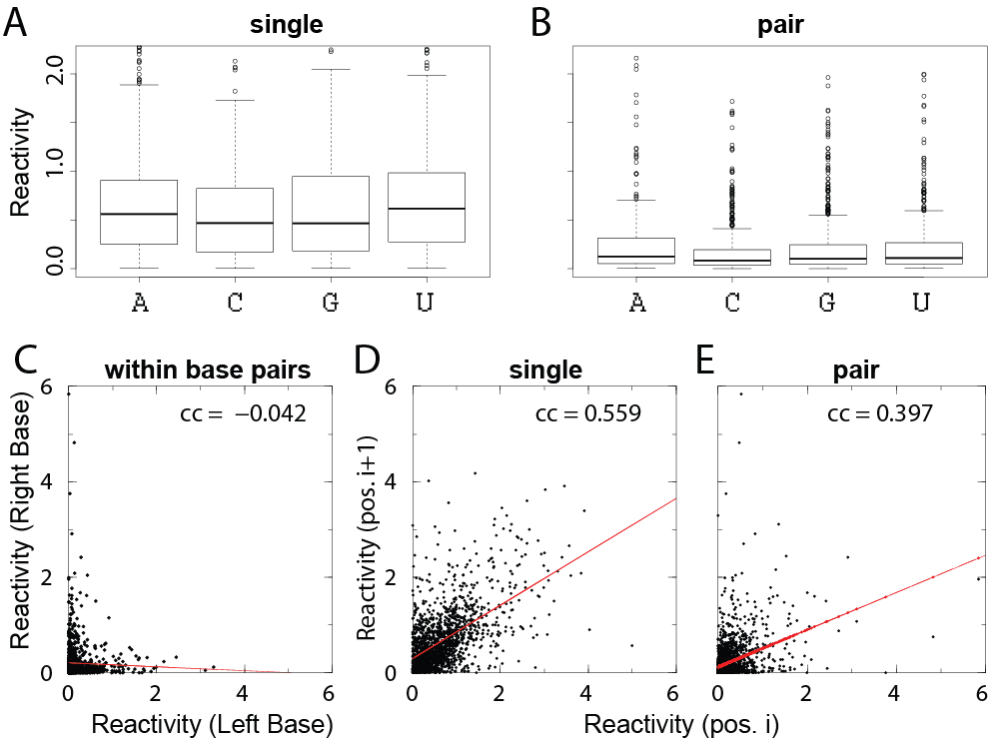


Fig. 8 Correlations in SHAPE data. Box-plots showing distribution of SHAPE reactivities for individual nucleotides for (A) single (unpaired) and (B) paired regions. Scatterplots showing (C) lack of correlation between left and right side of base pairs; (D) positive correlation along the sequence for both unpaired bases; and (E) positive correlation along the sequence for paired bases in stems. The regression line (red dashed line) summarizes the trend in the data.

Table 5 Average performance on six small structural RNAs of the Multidata versions of ProbFold and *RNAstructure* (Mathews, et al., 2004) with step-wise inclusion of CMCT, DMS and SHAPE structure probing data. Both the F-value and the change in F-value (**Δ**F) relative to the sequence-only (seq-only) predictions are shown.

	ProbFold		RNAstructure	
Data	*F-value*	ΔF	*F-value*	ΔF
Seq-only	0.40	NA	**0.73**	NA
Seq, CMCT	0.48	0.08	**0.85**	0.12
Seq, CMCT, DMS	0.54	**0.14**	**0.85**	0.12
Seq, CMCT, DMS, SHAPE	0.71	**0.31**	**0.82**	0.09

References

Cordero, P., et al. (2012) Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference, Biochemistry-Us, 51, 7037-7039.

Deigan, K.E., et al. (2009) Accurate SHAPE-directed RNA structure determination, Proceedings of the National Academy of Sciences of the United States of America, 106, 97-102.

Hand, D.J. (2012) Assessing the Performance of Classification Methods, International Statistical Review, 80, 400-414.

Kruskal, W.H. and Wallis, W.A. (1952) Use of Ranks in One-Criterion Variance Analysis, Journal of the American Statistical Association, 47, 583-621.

Mathews, D.H., et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, Proceedings of the National Academy of Sciences of the United States of America, 101, 7287-7292.

Rice, G.M., Leonard, C.W. and Weeks, K.M. (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE, Rna, 20, 846-854.

Weeks, K.M. (2012) 16S ans 23S E. coli data. Personal Communication.

Ongoing methylation studies

Genome-wide profiling of the prostate cancer methylome for biomarker discovery

Siri H. Strand, Michał Świtnicki, Philippe Lamy, Søren Høyer, Michael Borre, Jakob S. Pedersen, Torben Ørntoft, and Karina D. Sørensen

Introduction

Prostate cancer (PC) is the most commonly diagnosed malignancy and the third leading cause of cancer-related death in males in the Western world. Upon diagnosis of PC, the prognostic indicators available today (e.g. Gleason score) often have limited value for the individual patient (Felgueiras, et al., 2014), since many are mid-range. Thus, a major challenge in PC management is to distinguish between PC that will progress rapidly and become life-threatening, and PC that will remain latent and not affect the health of the patient. The latter group is very large and, theoretically, all men will develop PC if they live a long life, but less than 5% will die from it (Haas, et al., 2008). Overtreatment of clinically insignificant tumours, often identified by opportunistic PSA (prostate-specific antigen) testing (Borza, et al., 2013), remains a major problem due to the lack of accurate tools to distinguish aggressive from non-aggressive prostate cancer. Accordingly, there is a need for novel biomarkers that will help clinicians manage PC patients. Hence, by genome-wide profiling of the DNA methylome, we aim to identify new molecular markers that can improve the accuracy of diagnosis and prognosis of PC.

Methods

21 PC, 9 normal (N), and 12 adjacent normal (AN) prostate tissue samples were subjected to DNA methylation profiling using the Infinium HumanMethylation450 BeadChip®. Standard handling of the 450k methylation data described in this thesis was applied. Following the data preparation, we performed a hypothesis-free exploration of the data (using MDS) as well as a differential methylation analysis between PC and N+AN groups to identify biomarker candidates.

The diagnostic and prognostic potentials of 8 selected biomarkers were assessed by methylation specific qPCR (qMSP) in a new patient/control cohort consisting of 250 samples of localized PC and 29 benign specimens. Methylation levels were normalized to aluC4. The samples in this cohort were collected in Denmark and Switzerland, and the mean follow-up was 44.3 months (range 2-170 months).

The diagnostic potential was assessed using ROC analysis and rank sum-test. The prognostic potential was investigated by means of univariate and multivariate Cox regression analysis (Cox and Oakes, 1984), with methylation as continuous as well as dichotomized variables. Kaplan-

Meier survival analysis (Rich, et al., 2010) was also performed using the biochemical recurrence as the end point. The cut-off points for dichotomized analyses were found by ROC analysis.

Results & Discussion

At first, MDS analysis was performed on various subsets of most variable probes/regions. The highlight of most MDS plots was the significant PC heterogeneity in comparison with N and AN samples, as exemplified by 10,000 most variable CpG sites (Fig. 9). Also, multi-dimensional scaling, using the 10,000 most variable CpG sites, showed that N and AN samples clustered very tightly together, whereas PC samples showed great heterogeneity. The heatmap visualization of this set revealed that these sites in general are highly methylated in PC samples and unmethylated in N and AN samples (Fig. 10).

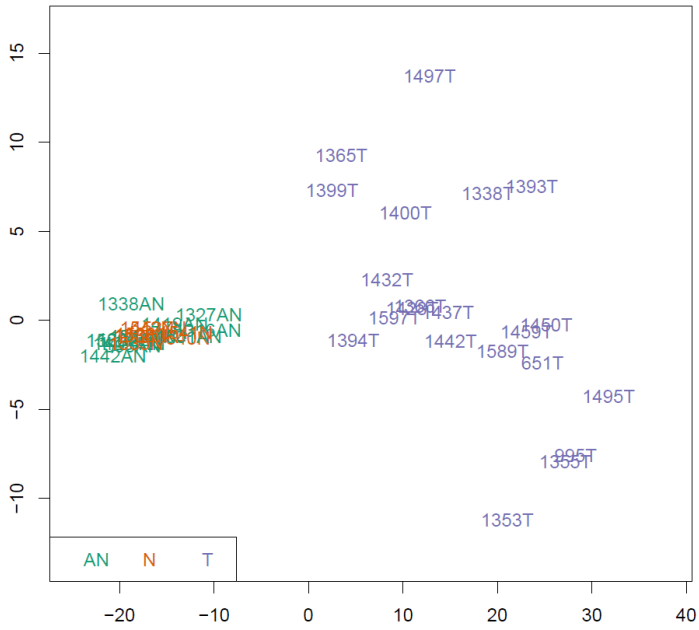


Fig. 9 Multidimensional scaling analysis of 10,000 most variable CpG sites showing significant PC heterogeneity in comparison with AN and N controls. Colouring of samples according to the group membership: normal (orange), adjacent normal (green) and PC (light purple).

Seeing close resemblance of N and AN samples, we pooled them into one control group for further analyses. Although identification of differential methylation between N and AN samples could potentially aid in correct diagnosis of patients with false negative biopsies, only 16 probes showed significant differential methylation between N and AN samples (FDR <0.05, Δbeta≥0.2), none of which corresponded to the same genomic locus. Due to this similarity in methylation between N and AN samples, pooling seemed to be the right choice.

We also looked at global distribution of methylation in different gene elements (Fig. 11). It revealed typical methylation patterns in control tissue: low (TSS1500 and 5'UTRs) and very low (TSS200 and 1st exons' probes) methylation status in promoters and high levels in gene bodies

and 3'UTRs. In all interrogated gene elements, we observed overall hypermethylation for tumours in comparison with control samples. The overall similarity between methylation levels of CpG sites located in TSS1500, TSS200, 5' UTR and 1st EXON led us to consider these as promoter CpGs, while the similarity of GENE BODY and 3' UTR methylation led us to consider these sites as gene body CpGs in our future developments.

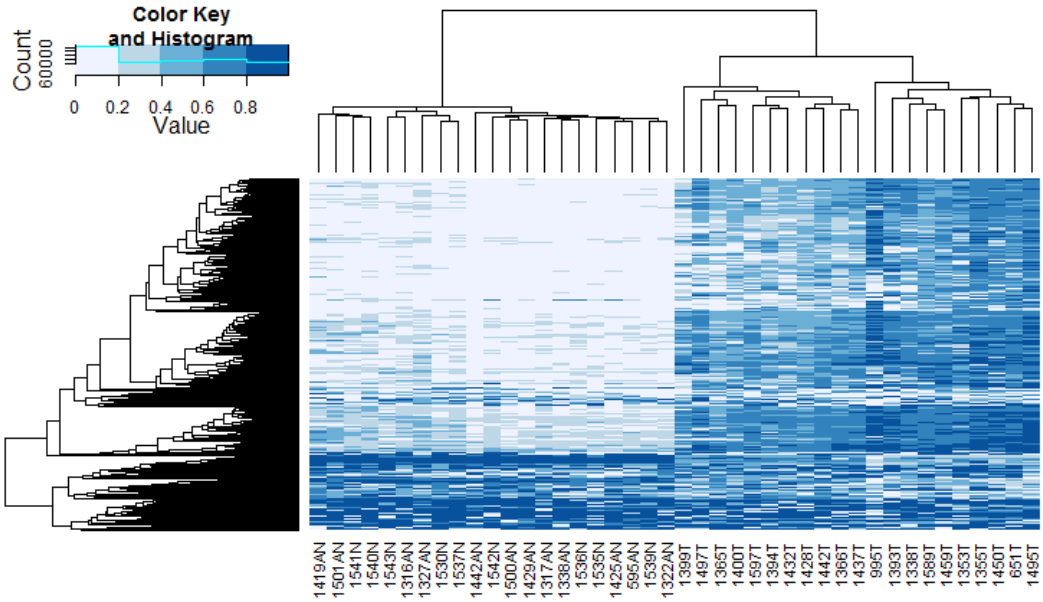


Fig. 10 Heatmap with double clustering visualizing the 10,000 most variable CpG sites, showing predominant high methylation of these loci in PC, compared to AN and N samples. Colouring of cells according to beta methylation level: the darker the colour, the higher the methylation.

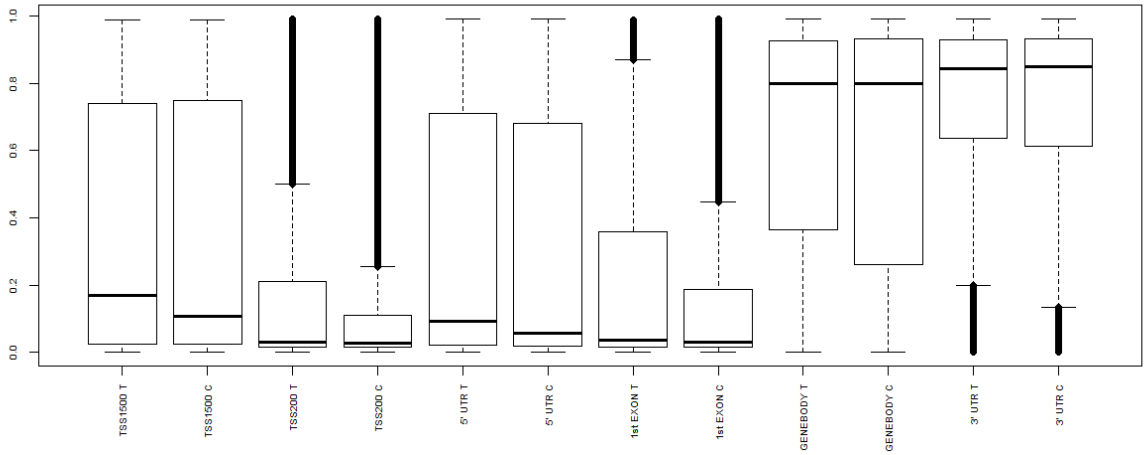


Fig. 11 Boxplot presenting the group's (T: tumours, C: controls) mean beta-value at individual CpG sites split between gene elements, as defined for the 450k platform by Illumina (Bibikova, et al., 2011).

In the differential methylation analysis between PC and control groups, 138,634 (29.76%) CpG-sites were found significant. To identify biomarker candidates, we applied a strict Δbeta cut-off value of 0.55, and identified 324 significantly differentially methylated CpG sites. Of these, 259 sites were associated with a number of different genes (163) and primarily hypermethylated in PC samples. Following initial screening, additional criteria were applied for the selection of final set of 8 candidate genes (Table 6) that included novelty for PC and consistency of the methylation change among neighbouring CpGs. Due to patenting constraints, identity of selected candidates must remain concealed. To exemplify our biomarker choices, we present methylation across Candidate 1 CpG sites (Fig. 12).

Table 6 Eight candidates selected for further validation. Univariate Cox regression analysis of candidate methylation was performed as continuous variable, showing highly significant prognostic potential for 4 of the candidates. The end-point was time to PSA recurrence.

Gene	Mann-Whitney p-value	AUC	HR (95% CI)	Cox regression p-value
Cand. 1	<0.001	0.9236	2.23 (1.48 – 3.37)	<0.001
Cand. 2	<0.001	0.9561	14.2 (4.88 – 41.2)	<0.001
Cand. 3	<0.001	0.8971	0.625 (0.019 – 20.5)	0.792
Cand. 4	<0.001	0.876	0.754 (0.193 – 2.94)	0.684
Cand. 5	<0.001	0.9612	6.37 (2.51 – 16.2)	<0.001
Cand. 6	<0.001	0.9397	12.5 (3.64 – 43.0)	<0.001
Cand. 7	<0.001	0.9356	1.32 (0.570 – 3.04)	0.52
Cand. 8	<0.001	0.8514	2.09 (0.998 – 4.37)	0.051

The final list contained candidates showing great diagnostic potential with AUCs ranging from 0.8514 to 0.9616 (Table 6). In terms of prognostic potential, univariate Cox regression analysis showed that methylation of four of the candidates was significant when analysed as continuous variables, with hazard ratios (HRs) ranging from 2.23 to 14.2 ($p<0.001$).

One candidate (Cand. 2) was significant in multivariate Cox regression analysis when analysed as continuous (HR=5.18, p=0.007), as well as dichotomized variable (HR=2.40, p=0.005). A second candidate (Cand. 5) was significant in multivariate analysis as a dichotomized variable only (HR=2.32, p=0.002).

Thus, two of our novel methylation biomarker candidates seem to provide added prognostic value to the currently used parameters tumour stage, Gleason score, preoperative PSA and surgical margin status. The findings are to be validated in an independent cohort including ~400 samples with long clinical follow-up, and the diagnostic potential will be investigated in needle biopsy specimens.

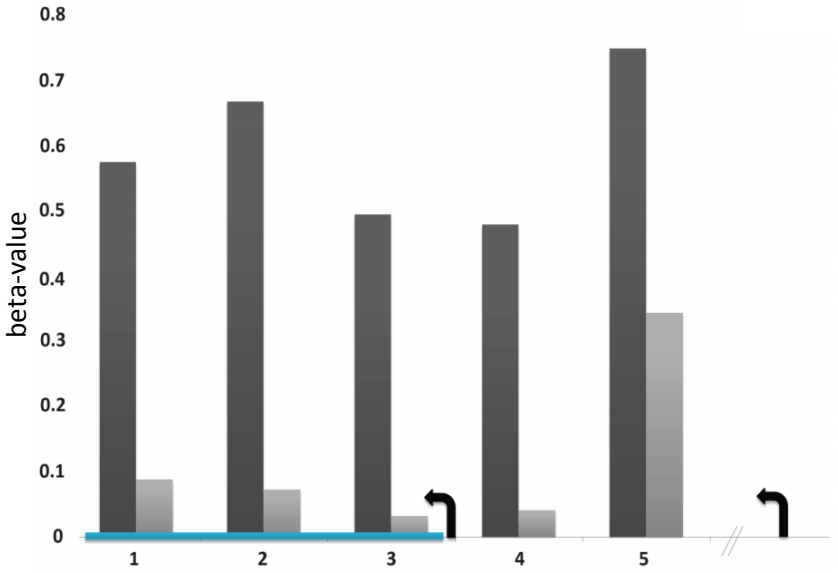


Fig. 12 Mean PC and control beta-values for the 5 differentially methylated CpG sites associated with Candidate 1, as measured by the Illumina 450K array. Colouring of bars according to group membership: PC (dark grey), control (light grey).

References

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

Borza, T., Konijeti, R. and Kibel, A.S. (2013) Early detection, PSA screening, and management of overdiagnosis, *Hematology/oncology clinics of North America*, **27**, 1091-1110, vii.

Cox, D.R. and Oakes, D. (1984) *Analysis of survival data*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York.

Felgueiras, J., Silva, J.V. and Fardilha, M. (2014) Prostate cancer: the need for biomarkers and new therapeutic targets, *Journal of Zhejiang University. Science. B*, **15**, 16-42.

Haas, G.P., *et al.* (2008) The worldwide epidemiology of prostate cancer: perspectives from autopsy studies, *The Canadian journal of urology*, **15**, 3866-3871.

Rich, J.T., *et al.* (2010) A practical guide to understanding Kaplan-Meier curves, *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, **143**, 331-336.

Genome-wide methylation analysis in Klinefelter syndrome

Anne Skakkebæk, Michał Świtnicki, Anders Bojesen, Jens M. Hertz, John Østergaard, Anders Degn, Mikkel Wallentin, Karina D. Sørensen, and Claus H. Gravholt

Introduction

Klinefelter syndrome (KS) is a set of symptoms that result from presence of two or more X chromosomes in male karyotypes (Nieschlag, et al., 2014). Epigenetic changes such as DNA methylation have been proposed to play a role in human illnesses such as psychiatric diseases, autoimmune disorders and metabolic diseases such as obesity and diabetes. While KS is associated with an increased risk of these disorders, no study to date investigated genome-wide methylation patterns in patients with KS.

Methods

Blood samples from 73 patients with KS and 73 age- and gender-matched controls were subjected to DNA methylation profiling using the Infinium HumanMethylation450 BeadChip®. Apart from standard handling of the 450k methylation data described in this thesis, a number of CpG probes (n=67461) were excluded from further analysis based on presence of common SNPs, missing values or high detection p-values (signifying technical problems).

Following the data preparation, we performed a hypothesis-free exploration of the data (using MDS) as well as a differential methylation analysis of all remaining CpG sites.

Results & Discussion

At first, MDS analysis was performed on all CpG sites to inspect the consistency of sample groupings (Fig. 13). It revealed one of the samples was approx. 50% mosaic (with ID=”31”). Seeing such a large difference between KS and control groups, we suspected that the pattern is driven by the second X-chromosome inactivation by methylation (Ahn, 2008) and attempted to see whether the differential pattern between patients is retained when CpG sites on X chromosome were excluded. Focusing on 10,000 most variable CpG sites in this reduced set, MDS revealed similar separation of KS patients from controls (Fig. 14). However, the effect size was smaller than when X-chromosome sites were considered.

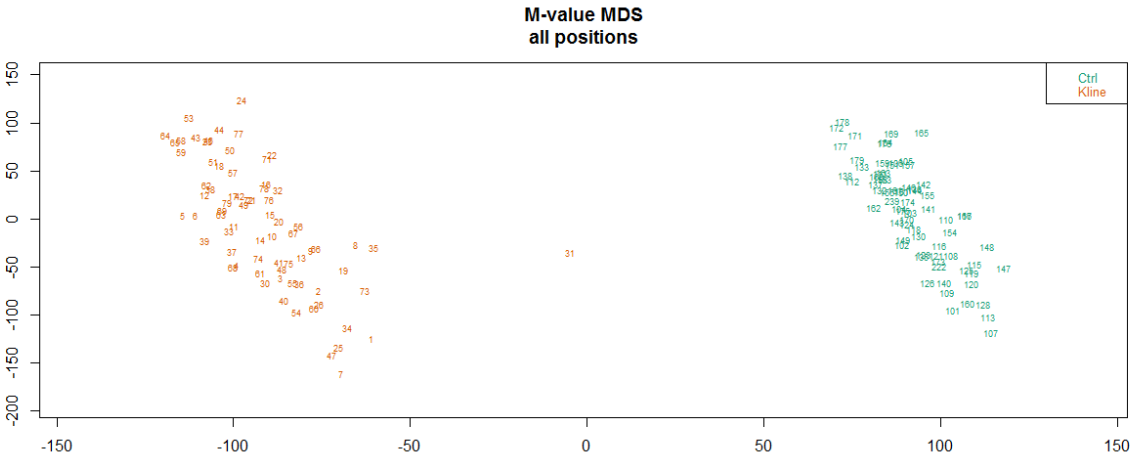


Fig. 13 Multidimensional scaling analysis of all CpG sites after initial filtering, showing a single patient with mosaicism. Colouring of samples according to the group membership: control (green) and KS (orange).

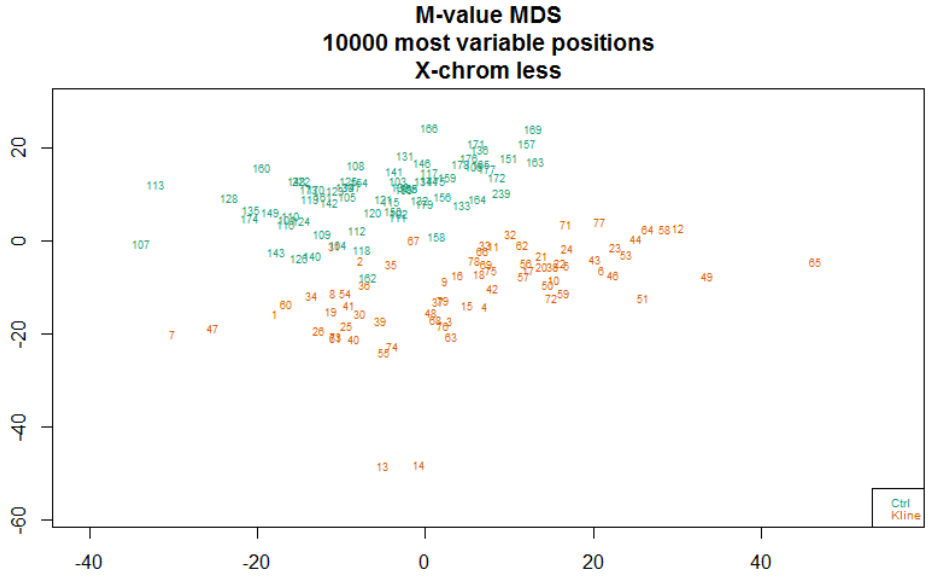


Fig. 14 Multidimensional scaling analysis of 10,000 most variable CpG sites revealing group separation. Colouring of samples according to the group membership: control (green) and KS (orange).

Based on the finding about mosaicism of patient "31", we decided to exclude him from further consideration in the differential testing. In the differential methylation analysis between KS and control groups, 70,525 CpG-sites covering over 15,000 genes were found significant. Among these 61.567 were on autosomal chromosomes (Fig. 15), 8903 were on the X-chromosome (Fig. 16) and 55 were on Y-chromosome (Fig. 17).

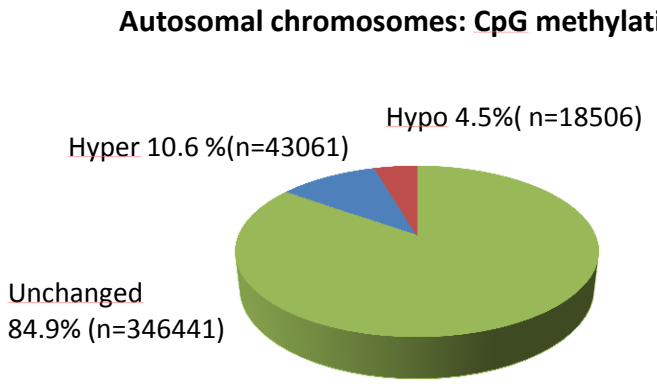


Fig. 15 Distribution of CpG sites located on autosomal chromosomes according to differential methylation status.

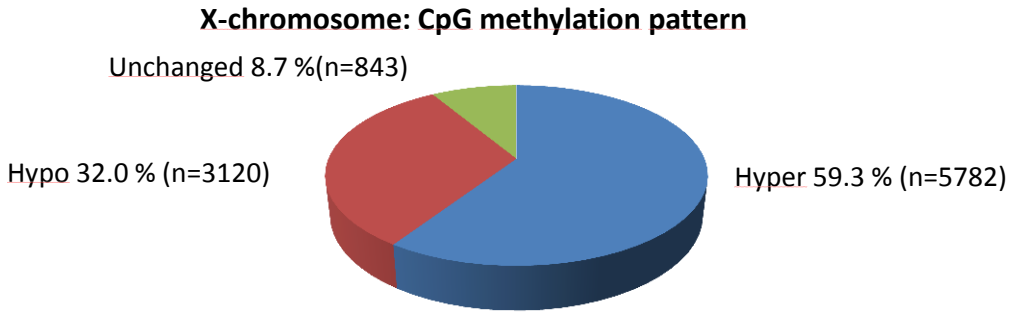


Fig. 16 Distribution of CpG sites located on X chromosome according to differential methylation status.

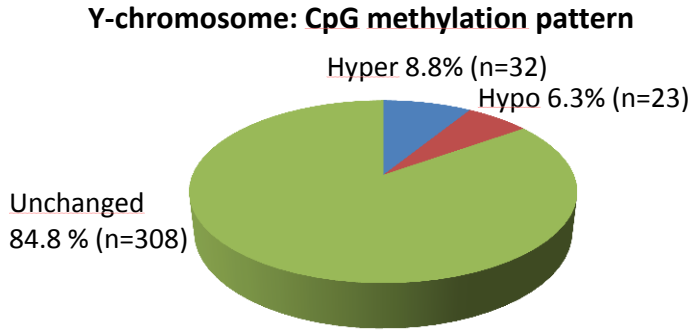


Fig. 17 Distribution of CpG sites located on Y chromosome according to differential methylation status.

One of the genes (NSD1) had its promoter differentially methylated in 3 CpG sites. NSD1, Nuclear receptor SET-domain protein 1**,** is involved in the androgen receptor (AR) transactivation (Chan, et al., 2013). Deletion or mutation in NSD1 causes Sotos syndrome (cerebral gigantism) (Kurotaki, et al., 2005) which is characterized by intellectual impairment, behavioral problems, attention deficit hyperactivity disorder (ADHD), phobias, problems with speech and language and

hypotonia – the state of low muscle tone. Changes in the IGF system were also observed in Sotos syndrome patients.

Other genes found differentially methylated in our study could possibly be involved in the phenotype of KS. These include ABI3BP, APOB, C1orf59, CACYBP, DPPA5, GABRG1, HOXA4, LRRC61, NLRP2, PEX10, RPLP1, RFPL2, SDHAF1 and SPEG. Several of these candidates (RPLP1, NLRP2, SDHAF1) additionally exhibited reduced expression in comparison with controls in our separate RNA-seq study on the subset of the cohort analysed here.

In summary, this is the first time anyone showed that KS is associated with pervasive genome-wide methylation changes. These changes are believed to play a role in the clinical phenotype seen with KS and may suggest that a hitherto unknown mechanisms may be involved in Klinefelter syndrome.

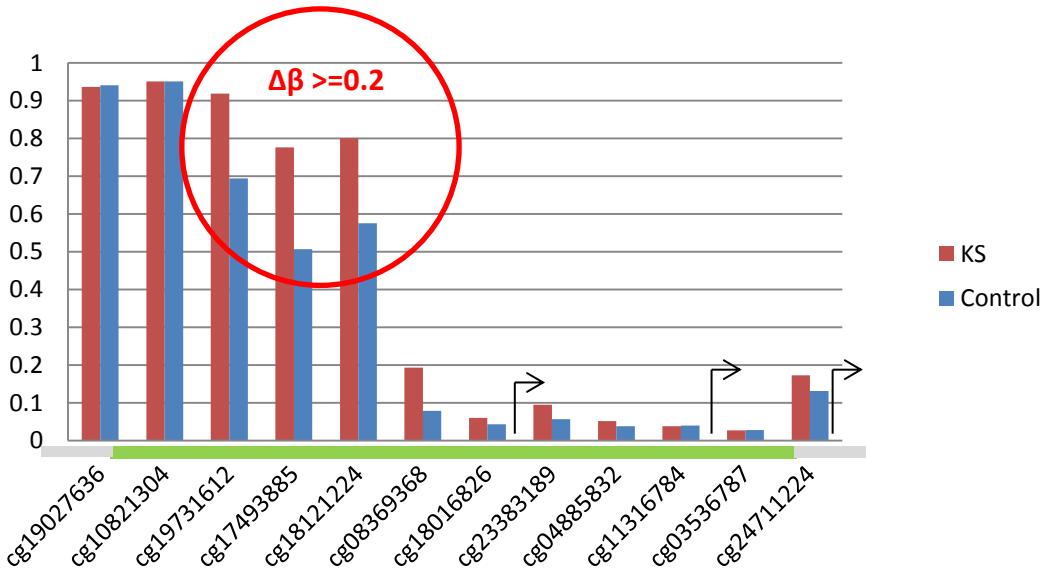


Fig. 18 Review of the CpG sites measured by the 450k platform for the NSD1 gene. 3 sites had differences in methylation levels between KS and control group larger than 0.2, signifying biological relevance.

References

Ahn, J.Y.L., J. T. (2008) X Chromosome: X Inactivation. *Nature Education*.

Chan, C.M., *et al.* (2013) A signature motif mediating selective interactions of BCL11A with the NR2E/F subfamily of orphan nuclear receptors, *Nucleic Acids Res*, **41**, 9663-9679.

Kurotaki, N., *et al.* (2005) Sotos syndrome common deletion is mediated by directly oriented subunits within inverted Sos-REP low-copy repeats, *Hum Mol Genet*.

Nieschlag, E., *et al.* (2014) New approaches to the Klinefelter syndrome, *Annales d'endocrinologie*, **75**, 88-97.

Regulation of Growth hormone target genes by DNA methylation and its relation to in vivo GH signalling in skeletal muscle of adult human subjects: a pilot study.

Morten H. Pedersen, Michał Świtnicki, Poul F. Vestergaard, Niels Jessen, and Jens O.L. Jørgensen

Introduction

Growth hormone (GH) secretion decreases with age in humans. GH plays an important role in substrate metabolism and hence its effect in the human body weakens in aging subjects. Targeted disruption of the GH receptor in mice extends longevity, which is associated with decreased expression of apoptosis-related genes including caspase-9 (CASP9) in skeletal muscle (Gesing, et al., 2011). This pilot study was performed to research the DNA methylation of putative GH target genes in skeletal muscle of adult male subjects in relation to body composition, physical fitness, serum IGF-I levels and in vivo GH signalling.

Methods

12 healthy adult subjects (10 males and 2 females) were divided into a 'young' (n= 5) and 'old' (n=7) groups with mean age of 25 (20-27) and 66 (63-69) for young and old, respectively. The subjects' skeletal muscle tissues were subjected to DNA methylation profiling using the Infinium HumanMethylation450 BeadChip®. Apart from standard handling of the 450k methylation data described in this thesis, additional filtering of CpG sites was performed due to mixed male-female setup: X and Y chromosome sites were filtered out to equalize male and female probe sets and to ensure sound comparison (methylation of second female X-chromosome is a mechanism for silencing the duplicated genes located on this chromosome (Ahn, 2008)).

Following the data preparation, we performed a hypothesis-free exploration of the data (using MDS) as well as a differential methylation analysis of proximal promoter sites between old and young subjects. Specifically, we focused on CpG loci located 1500 and 200 bases upstream of transcription start sites (Illumina's TSS200 and TSS1500), increasing the possibility that the CpG site is regulating the gene transcription.

Results & Discussion

Analysis of the 50 (Fig. 19) and 1000 (Fig. 20) most-variable probes within the set in the hypothesis-free exploration of the data did not reveal any patterns that could be explained by the known clinical variables such as gender and age.

In a differential methylation analysis between old and young groups, only a single site was significant after multiple-testing correction (Fig. 21). The identified site, "cg16706559", is located within 1500 bases upstream from the CASP9 gene start site. It was always fully unmethylated

amongst the young subjects, with small, yet consistent methylation upward shifts in the muscle tissue for old subjects.

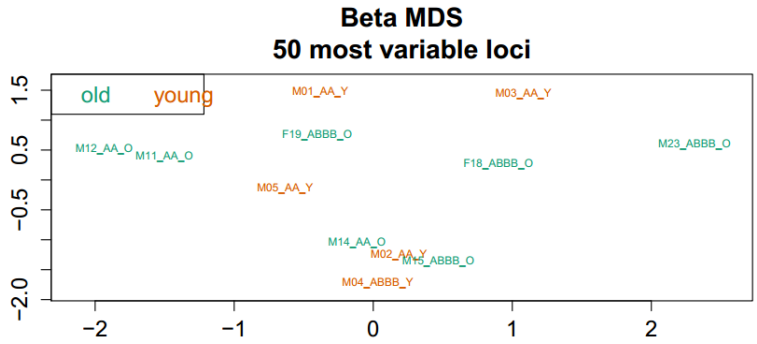


Fig. 19 Multidimensional scaling analysis of 50 most variable CpG sites. Colouring of samples according to the age group.

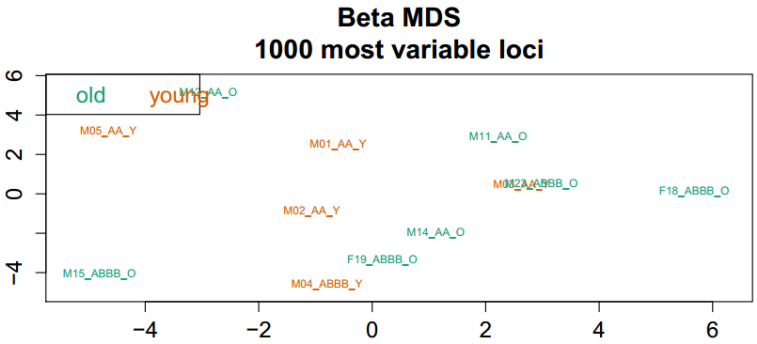


Fig. 20 Multidimensional scaling analysis of 1000 most variable CpG sites. Colouring of samples according to the age group.

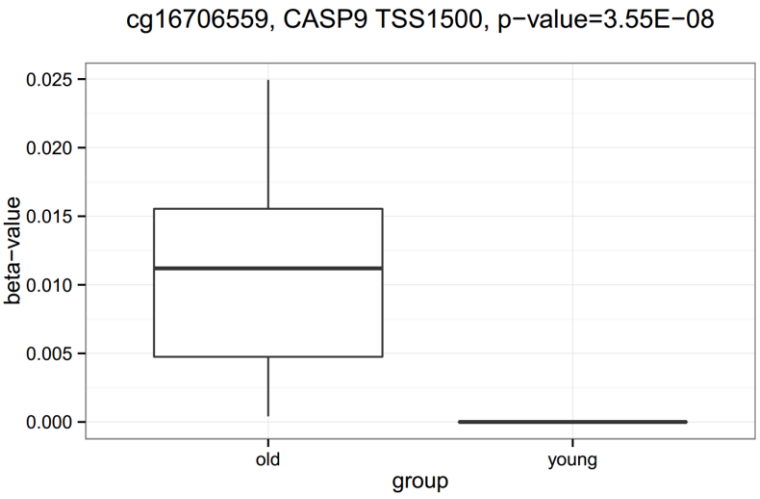


Fig. 21 Old versus young boxplot of CASP9 TSS1500 CpG site showing marginal methylation of the locus in old patients.

To conclude, no consistent changes in methylation of CpG sites between the age groups were observed. This could be partly explained by the small sample size which limits our power to detect methylation changes (many genes were nominally significant). However, only a single

CpG site in the promoter of CASP9 was found differentially methylated, meeting the stringent false discovery criteria (<0.05). CASP9 is a pro-apoptotic initiator of the intrinsic pathway-activating effector CASP3. Decreased expression of CASP9 was previously found in skeletal muscle of GHRKO mice (Gesing, et al., 2011). Since reduced levels of apoptosis are considered beneficial for longevity, we hypothesize that reduced GH levels during senescence alters methylation status of the promoter region of pro-apoptotic factors, including CASP9. This in vivo model holds promises to disclose hitherto unrecognized regulatory mechanism of GH activity.

References

Ahn, J.Y.L., J. T. (2008) X Chromosome: X Inactivation. *Nature Education*.

Gesing, A., *et al.* (2011) Decreased expression level of apoptosis-related genes and/or proteins in skeletal muscles, but not in hearts, of growth hormone receptor knockout mice, *Experimental biology and medicine*, **236**, 156-168.

DNA-methylation profile in laryngeal spinocellular carcinoma and the impact of HPV

Claes H. Karstensen, Michał Świtnicki, Jakob S. Pedersen, and Thomas Kjærgaard

Introduction

Clinical and epidemiological studies highlight a significant heterogeneity for head and neck spinocellular carcinoma (HNSCC) regarding aetiology, cellular, and molecular features, as well as clinical behaviour. Prognosticators such as disease stage and etiological and demographic factors do not sufficiently predict patient outcome, and knowledge of underlying molecular mechanisms responsible for differences in clinical behaviour remains limited. Better understanding of HNSCC tumour biology and identification of prognostic molecular biomarkers is needed to improve prediction of patient survival and treatment decision making. This is particularly relevant for laryngeal spinocellular carcinoma (LSCC), one of the most prevalent HNSCC-types, with survival rates being only modest, and largely unchanged during the last decade, despite novel treatment algorithms.

Within other fields of cancer research, aberrant DNA methylation is considered an attractive novel molecular biomarker for staging and prognosis, and a possible potent druggable target. For HNSCC, however, DNA methylation is still an unexplored concept to explain development and prognosis of this malignancy, and, to date, the majority of published data show limited ability to detect strong overall survival associations. Only a few studies have considered Human Papilloma Virus (HPV) status, but recent findings indicate differences in methylation patterns of HPV+ and HPV- oropharyngeal spinocellular carcinoma. This remains unexplored in LSCC.

Here, we focus our attention on the molecular characteristics of LSCC. Our aim is to describe DNA-methylation profiles in LSCC and uncover signatures for HPV and non-HPV (tobacco) related gene promoter methylation.

Methods

A total of 24 LSCC and 12 normal larynx Formalin-Fixed Paraffin-Embed (FFPE) samples were subjected to DNA methylation profiling using the Infinium HumanMethylation450 BeadChip®. To determine the HPV positivity, a surrogate marker correlated with HPV status was used in immunohistochemistry staining, the p16INK4A (Stephen, et al., 2013) (p16, in short). Among the LSCC samples, 12 were p16-positive and 12 were p16-negative. X and Y chromosome CpG probes were also removed as the cohort contains both male and female patients.

Standard handling of the 450k revealed that 5 samples had erroneous global methylation profiles for one type of probes as signified by lack of bimodality of methylation, and due to that, they were excluded from further consideration. It could be that these bad samples were especially old or were not handled properly when subjected to FFPE procedure. Failed samples were evenly spread between controls and p16-postivie and –negative tumours and hence did not greatly affect downstream analyses.

As a special focus of this study, methylation of gene promoters was calculated as mean of CpG sites belonging to regions extending from 1,500 bases upstream of the transcription start site (TSS) to the end of the first exon as defined by Illumina's categories (TSS1500, TSS200, 5' UTR and 1st Exon) (Bibikova, et al., 2011).

Results & Discussion

At first, MDS analysis was performed on various subsets of most variable probes/regions. The highlight of most MDS plots was the significant LSCC heterogeneity, as exemplified by 1000 most variable promoter regions (Fig. 22).

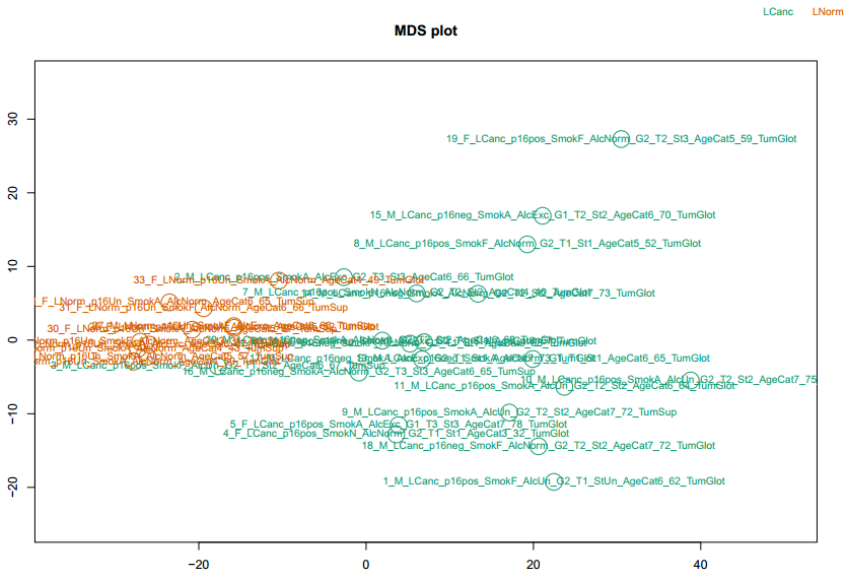


Fig. 22 Multidimensional scaling cohort analysis using 1000 most variable promoters reveals significant cancer heterogeneity. Colouring of samples according to the group membership: LSCC (green) and normal larynx (orange).

Further, we performed differential methylation analysis between normal larynx and LSCC samples. We identified many significant CpG sites (n=60,175) at 5% FDR. Differential analysis between p16-positive and –negative samples did not reveal any CpG sites that were significant after multiple testing correction, despite many of them being nominally significant (n=22,739). The increased variance among tumour samples makes it difficult to detect signal when sample counts are small. To strengthen the p16-positive and –negative differential analysis, additional sample inclusion is recommended in this case.

Regardless, visualizing the p16-positive and –negative most significant sites/regions, revealed interesting clustering of normal larynx samples with p16-negative tumours, as exemplified by 100 most significant promoters (Fig. 23), suggesting different LSCC aetiology from the p16-positive ones. Despite the clustering analysis consistently clustered normal larynx samples with p16-negative tumours in variants of the heatmap analysis, that trend was not that apparent in most equivalent MDS analyses (Fig. 24).

Further work is required to interpret comparisons made between defined groups in this dataset in a comprehensive way, as the research is still in an early stage.

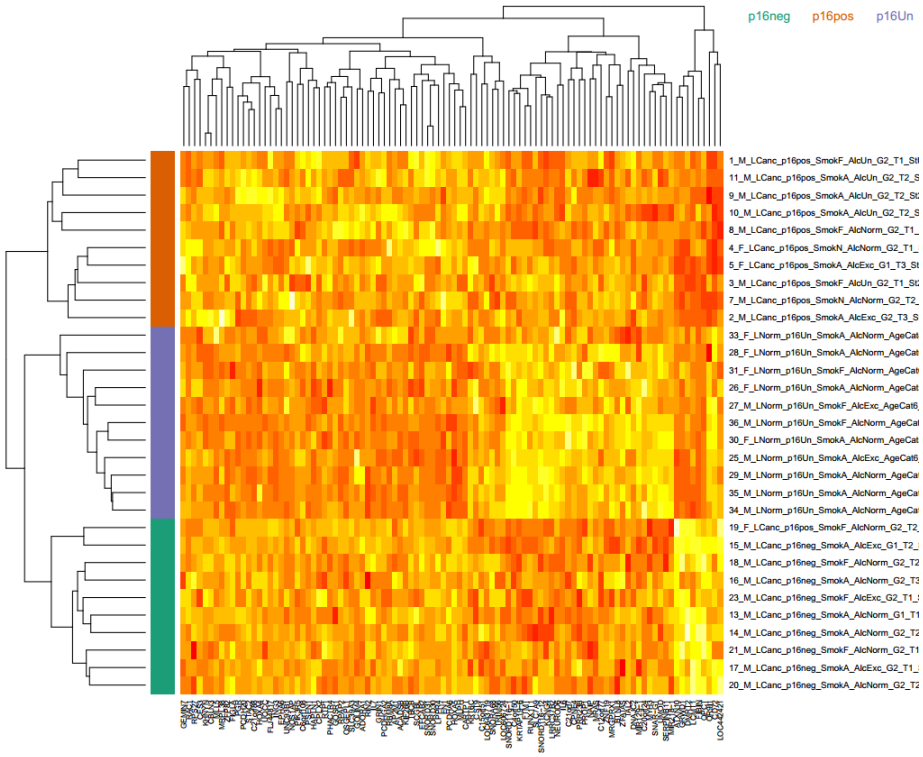


Fig. 23 Heatmap analysis of 100 most significant promoters in the comparison between p16-positive and –negative tumours. Colouring of samples according to the group membership: p16-positive (orange), p16-negative (green) and normal larynx (light blue).

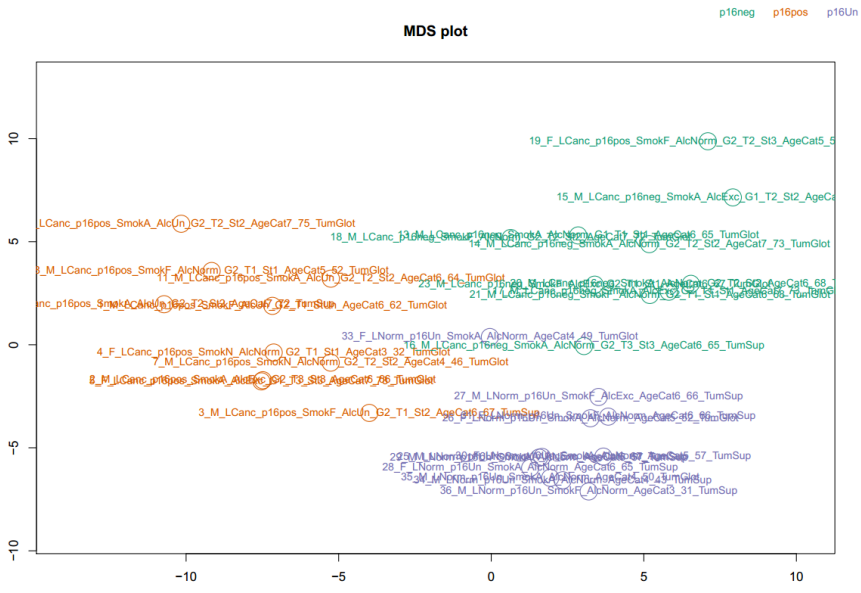


Fig. 24 MDS analysis of 100 most significant promoters in the comparison between p16-positive and –negative tumours. Colouring of samples according to the group membership: p16-positive (orange), p16-negative (green) and normal larynx (light purple).

References

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

Stephen, J.K., *et al.* (2013) Significance of p16 in Site-specific HPV Positive and HPV Negative Head and Neck Squamous Cell Carcinoma, *Cancer and clinical oncology*, **2**, 51-61.

Chapter 5: Discussion

Integrative data analyses are likely to increase in importance as researchers and clinicians begin to routinely generate multiple data types from the same sets of samples. However, to date only a few international consortia, such as TCGA and ICGC, could afford such large-scale analyses for a sufficient number of samples. On the other hand, integrative analyses should become relevant for smaller sample sizes too, to advance the patient diagnosis and prognosis for the less studied diseases.

Methods for integrating multiple molecular levels can be efficiently designed thanks to the use of Probabilistic Graphical Models (PGMs). However, one has to be aware of the limitations of building such multi-data models. First of all, generally the more comprehensive the proposed model, the more parameters there will be to learn from the data, to the point of being prohibitive in comparison with the number of available data points (bias-variance trade-off (Hastie, et al., 2009)). This might seem counter-intuitive at first, as the purpose of integrating data is to share information across the data types about the studied disease. However, the extra parameters really describe the relationships between the integrated data types and hence we end up learning about the system as a whole too. Using the domain knowledge we can, however, greatly reduce the parameter cost. For instance, we have presented two variants of an integrative model of gene expression and DNA methylation of promoter and gene body regions (each region summarizing a number of CpG loci). The variants differ by the parameterization: the first one offering very comprehensive reflection of the data types and the relationships among them, suitable for large cohort analyses, while the second one imposing some additional constraints and thus reducing the number of parameters and tailoring the method to smaller sample sizes.

Comparing our integrative methods for cancer analysis with off the shelf machine learning approaches, we can immediately spot the great advantage of using the structured integration: interpretability. In particular, the genes identified by our methods require not any further importance analyses, as would be required in case of most machine learning approaches. Also, most other integrative models aim to identify clusters of features stratifying the patients according to the outcome (Kristensen, et al., 2014), which is suboptimal from the biomarker discovery perspective, as it is the individual genes that are adopted into the clinical panels.

Finally, it is essential to stress the importance of the domain knowledge. Unsupervised integration often leads to unnecessary spikes in the number of parameters and the contribution from new information could be rendered useless in such cases. It may also fail to contribute new information whatsoever, and hence artificially reduce the variance (if the same information was integrated over) or introduce additional noise to the predictions made (if the integrated

information was neither complementary, nor supplementary to the already included set of features). Last but not least, PGMs are modular in nature, so integration of additional data types can be done again in an optimal way, ensuring that a system is comprehensive as a whole.

In case of the secondary structure prediction using ProbFold, we presented a framework capable of combination of SHAPE and other chemical or enzymatic probing agents such as CMCT and DMS, bearing both complementary and supplementary information contributing positively to the folding performance. It is preferred to include many such data sets as the noise is typically a problem for each individual one. Moreover, this noise at individual sites is typically correlated between different probing agents and hence it is important to capture this correlation in the model, to retain the specificity. One way to better address this would be to include some tertiary structure aspects into the model. What is a noise to the secondary structure modelling can be a signal in the tertiary structure prediction (Kopeikin and Chen, 2005; Lorenz, et al., 2013). As more NGS-based high-throughput probing datasets emerge, we expect a continuous improvement in the quality and uniformity of these sets. This will likely improve our ability to correctly predict novel structures, and specifically, build transcriptome-wide RNA structure maps.

Chapter 6: Future perspectives

Future works could be focused on five aspects of the research presented in this thesis:

1) The feasibility of identified integrative biomarkers. Initially, a significant amount of work would be required to design robust PCR-based validation strategies for both gene expression and methylation of the identified biomarkers. Thereafter these assays could be used for the actual validation in independent patient cohorts.

2) The expansion of the model to additional data types, for example to copy number status. Interestingly, it was demonstrated that copy number aberrations could be directly inferred from the 450k data (Morris, et al., 2014) so no additional data generation constraints would be needed in place when integrating this type of data. More generally, the integrative data analysis field is expected to follow with the advancement of the experimental techniques, and especially with the decrease in the cost of these procedures. Due to that, more data types will be generated for cohorts larger than today. This represents a huge potential for development and use of future integrative data analysis methods.

3) Analysing other cancer types, either using local MOMA data sets, or a pan-cancer analysis using other publically available data sets from TCGA and ICGC. The latter is not preferred as analysis of local data carries a potential to validate the findings experimentally in an independent cohort.

4) As the sequence-only ProbFold model based on SCFGs had poor performance in comparison with the field standards like *RNAStructure*, it could be worthwhile to exchange it for better performing model. The data-specific emission sub-models integrated the experimental data well so such an upgrade would be desired before realising the final aim and 5) predicting the RNA secondary structures genome-wide for a number of organisms.

Chapter 7: Lay summaries

English lay summary

The goal of most cancer studies is to improve our understanding of development and progression of the disease and define biomarkers that aid clinicians in their practice. Both gene expression and DNA methylation have been extensively studied as cancer biomarker candidates. In recent years, DNA methylation was found to be one of the main epigenetic regulators of gene transcription. Different mechanisms of action were proposed for DNA methylation, depending on the functional role of the DNA sequence that the methylation occurs at. In this thesis, we describe some insights into the degree and the role of gene body and promoter DNA methylation in regulation of expression, and its relevance to cancer and other pathologies. Subsequently, we apply this knowledge for integration of methylation with expression data.

Generally, it is thought that performance of predictive models can be improved by integrating multiple types of data. Additionally, model fits can be enhanced by incorporating inter-relationships between integrated data types, especially when dependency structure is informed by expert knowledge. Predictive power can also be improved by including data types that bear independent, yet informative signal. DNA methylation and gene expression data types fulfil both complementarity and supplementarity of signal. In this thesis I describe the use of factor graphs, a family of probabilistic graphical models, for integrative genomics analysis of gene expression and DNA methylation in the cancer setting.

We applied knowledge of DNA methylation for building integrative models with gene expression. The primary aims were to identify integrative biomarkers of tumour development and progression and to better understand the cancer genome turbulence. The proposed models can exploit the dependency structure in the correct assignment of samples to trained groups and in evaluation of gene perturbation. The thesis comprises of two primary manuscripts: 1) PINCAGE: Probabilistic integration of cancer genomics data for perturbed gene identification and sample classification; 2) Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data. It also contains an example of how factor graphs can be used to integrate genomics data in the field of RNA structure analysis: ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction.

The first manuscript defines and evaluates an integrative model of gene expression and methylation of promoters and gene bodies in cancer. It learns the pairwise relationships between the data types and exploits these in group evaluations and classification. It also contains quantitative evaluation of the expression-methylation relationship in the cancer context.

The second manuscript describes a parameter-sparser implementation of the integrative model from the first publication that facilitates robust analyses of smaller cancer cohorts. It also discusses the relative merits of using both the parameter-sparser and the initial implementation of the model.

The third manuscript describes an application of probabilistic graphical models for integration of sequence data with different probing data sets to improve secondary structure prediction using Stochastic Context-Free Grammars. The goal here is to make better use of experimental structure probing data for RNA secondary structure prediction.

Apart from the integrative genomics work, a number of ongoing collaborative studies into DNA methylation in various pathologies are outlined in the thesis. They include: a characterization of methylation landscape in prostate cancer and identification of methylation-based biomarkers of diagnostic and prognostic value; an analysis of methylation in laryngeal cancer and differential methylation characterization of HPV-positive and HPV-negative cancer samples; characterization of the muscle methylation patterns in aging human subjects; and differential methylation analysis of Klinefelter syndrome patients. They provided valuable insights into the nature of DNA methylation phenomena. Additionally, the prostate cancer methylation studies informed our integrative modelling choices and designs.

As more and more types of genomic data types are being routinely produced, their integration becomes increasingly important. We show how expert knowledge helps producing powerful, yet interpretable models that can identify integrative biomarkers of cancer development and progression. Generally, we show that combining the evidence from multiple complementary sources, using factor graphs to encode the existing knowledge about interactions between data types, aids in predictive tasks in the fields of cancer and RNA secondary structure prediction. However, further works are needed both to 1) establish the robustness of integrative biomarker candidates and to 2) predict and validate the RNA secondary structures genome-wide.

Danish lay summary

Målet med de fleste cancer studier er at forbedre vores forståelse af hvordan sygdommen udvikler sig, samt at definere biomarkører til hjælp for klinikere i deres praksis. Både genekspression og DNA methylering er nøje studerede som mulige biomarkører for cancer. I de seneste år er det blevet kendt, at DNA methylering er en vigtig epigenetisk regulerende faktor for gentranskription. Forskellige mekanismer er blevet undersøgt i forbindelse med DNA methylering, afhængigt af funktionaliteten af den methylerede DNA sekvens. I denne afhandling beskriver vi nye indsigter i graden og funktionen af methylering af promoter- og kodende gensekvenser i reguleringen af ekspression, samt dennes relevans for cancer og andre patologier. Dernæst anvender vi denne viden til integration af methylerings- og ekspressionsdata.

Det er en generel antagelse, at resultater opnået ved hjælp af prediktive modeller kan forbedres ved at integrere forskellige typer af data. Derudover kan modeller forstærkes ved at indarbejde interne afhængigheder mellem integrerede data typer, særligt når strukturen af afhængighederne er fagkyndigt bestemt. Styrken af predikteringer kan ydermere forbedres ved at inkludere data typer med uafhængige, men stadig informative signaler. Signalerne fra data typerne DNA methylering og genekspression er både komplementære og supplerende. I denne afhandling beskriver jeg brugen af faktor grafer, en familie af probabilistiske grafiske modeller, til integrationsanalyse af genekspression og DNA methylering i sammenhæng med cancer.

Vi har anvendt viden om DNA methylering til at opbygge integrative modeller med genekspression. De primære mål var at identificere integrative biomarkører for udviklingen af cancer tumorer samt bedre at forstå hvordan cancer genomet er forstyrret. De foreslåede modeller kan udnytte afhængighedsstrukturen til korrekt at tilknytte samples til bestemte grupper samt at evaluere genfortyrrelser. Denne afhandling består af to primære manuskripter: 1) *PINCAGE: Probabilistic integration of cancer genomics data for perturbed gene identification and sample classification;* 2) *Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data*. Desuden indeholder afhandlingen et eksempel på hvordan faktor grafer kan anvendes til at integrere gen data til RNA struktur analyse: *ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction*.

I det første manuskript defineres og evalueres en integrativ model til genekspression og methylering af promotorer og kodende gensekvenser i cancer. Modellen lærer de parvise sammenhænge mellem data typerne, og udnytter disse i gruppeevalueringer og klassifikation. Den indeholder desuden kvantitativ evaluering af sammenhængen mellem ekspression og methylering i forhold til cancer.

I det andet manuskript beskrives en implementering af den integrative model fra den første publikation, med færre parametre, som muliggør robust analyse af mindre cancer kohorter. Her

diskuteres desuden de relative fordele ved at benytte implementeringen med færre parametre og den oprindelige implementering af modellen.

I det tredje manuskript beskrives en anvendelse a probabilistiske grafiske modeller til integrering af sekvensdata med forskellige probe datasæt for at forbedre predikteringen af den sekundære struktur ved hjælp af Stochastic Context-Free Grammars. Formålet med dette er en mere hensigtsmæssig anvendelse af eksperimentel struktur probing data til prediktering af sekundær RNA struktur.

Udover de integrative gen analyser er et antal igangværende kollaborative analyser af DNA methylering i forskellige patalogier skitseret i afhandlingen. Disse inkluderer: en karakterisering a methyleringslandskabet i prostatakræft samt identifikation a methyleringsbaserede biomarkører til gavn for diagnostik og prognostik; en analyse af methylering i strubekræft samt karakterisering af forskelle i methylering for HPV positive og HPV negative cancer prøver; en karakterisering af mønstrene i muskelmethylering hos patienter med type II diabetes af forskellig grad; en analyse af forskellige i methylering hos patienter med Klinefelter syndrom. Disse analyser leverede værdifulde indblik i DNA methylering. Derudover gav analysen af prostatakræft indsigter til hjælp af valg og design i vores integrative modellering.

Som flere og flere gen datatyper rutinemæssigt frembringes, bliver integrationen af disse vigtig. Vi viser hvordan ekspert viden hjælper til med at fremstille stærke, men stadig fortolkelige modeller, som kan identificere integrative biomarkører for udviklingen af cancer. Vi viser, at kombinationen af information fra flere komplementære kilder, hvor faktor grafer anvendes til at indarbejde eksisterende viden omkring sammenhænge mellem data typer, er brugbar til predikteringsopgaver i cancer samt til prediktering af sekundære RNA strukture. Dog er yderligere tiltag nødvendige for at 1) fastslå robustheden af kandidater til integrative biomarkører, samt 2) prediktere og validere den sekundære RNA struktur langs hele genomet.

Chapter 8: Manuscript 1

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

Manuscript in review at *Bioinformatics*

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

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Brendstrupgårdsvej 21, 8200 Aarhus, Denmark and [2]Bioinformatics Research Centre (BiRC), Aarhus University, C.F.Møllers Allé 8, 8000 Aarhus, Denmark

Running head: PINCAGE

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], Jakob S. Pedersen[1,2*]

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Brendstrupgårdsvej 21, 8200 Aarhus, Denmark and [2]Bioinformatics Research Centre (BiRC), Aarhus University, C.F.Møllers Allé 8, 8000 Aarhus, Denmark

Received on 11-06-2015; revised on 18-07-2015; accepted on XXXXX

Associate Editor: Inanc Birol

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, 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 data set 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: R scripts available at http://moma.ki.au.dk/prj/pincage/

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

Supplementary information: 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 sub-types 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 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; 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 data sets 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: (1) naïve combination of individual methods, (2) use of general-purpose machine-learning methods, and (3) 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 machine learning 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 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

*To whom correspondence should be addressed.

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 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. 1 A). The resulting data set consisted of 730 tumour (T) samples and 82 Adjacent Normal (AN) samples (Table S7).

The methylation array data was processed using the statistical language R (R 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 1,500 bases upstream of the transcription start site (TSS) to the end of the first exon, as defined by Illumina's categories (TSS1500, TSS200, 5' UTR and 1st Exon; 450k Manifest File v1.2 (Bibikova, et al., 2011); 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; 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 was already summarized per gene and no further processing was needed, except for calculation of original library sizes. For plotting and logistic regression analysis, we normalized gene expression read counts by library size and reported reads per million (RPM).

The data was summarized and organized by disease groups (T vs 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. 1 B). The data types, their distribution across samples, and their pairwise correlations are exemplified by the PLK1 gene (Fig. 2).

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

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)

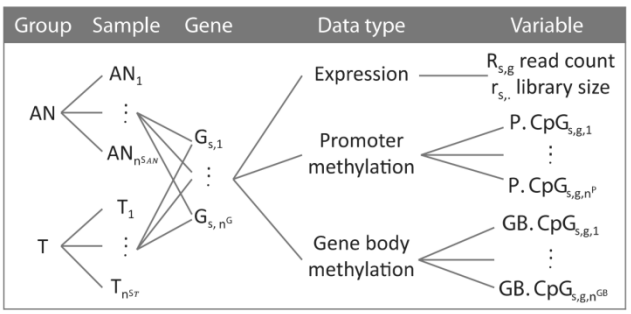


Fig. 1. Data summary. **A)** Definition of data sets and their sizes. **B)** Data structure schema: samples were divided into two groups: adjacent normal (AN), and tumour (T). Within each sample (indexed by s), genes (indexed by g) were profiled for mRNA expression levels and DNA methylation, yielding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

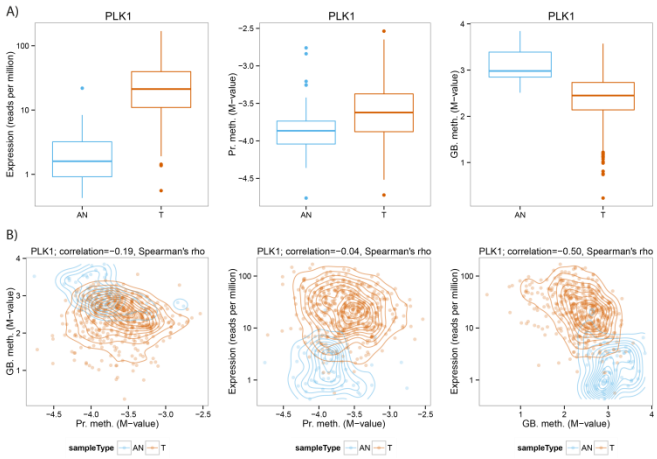


Fig. 2. Marginal and pairwise distribution of gene expression, promoter methylation, and gene body methylation for the PLK1 gene. **A)** Marginal distribution of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value across BRCA Tumour (T) and Adjacent 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).

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).

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. 4 A), 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 factorisation

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):

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

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

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:

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

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

where σ is an experimentally determined standard deviation; σ = 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 SUPPLEMENT: *Implementation* section).

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 reads per million (RPM). The uncertainty in the measured expression level depends on the library size: the smaller the library the larger the sampling variance. To capture it, 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:

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

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:

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

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 SUPPLEMENT: *Implementation* section).

PINCAGE integrative model

The integrative model combines the submodels 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 submodels 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.**).

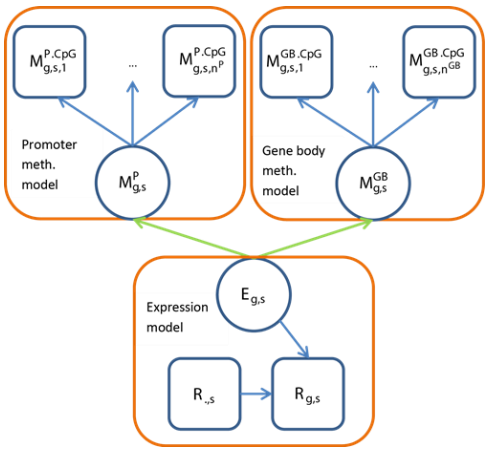


Fig. 3 Directed acyclic graph representation of PINCAGE probabilistic graphical model. Individual submodels are sub-set using orange boxes. The dependencies highlighted in green are present only in the integrative model.

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:

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

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 two-dimensional Gaussian kernels (see SUPPLEMENT: *Implementation* section). The integrative model can learn the joint distribution of expression and methylation in promoter as well as gene body regions.

2.3 Applications

Significance evaluations

We evaluate the significance of expression-, methylation- or joint expression and methylation gene perturbations using a variant of the Likelihood Ratio Test (LRT) (Neyman, 1933). Consider a calculation of the D statistic in a comparison between adjacent normal and tumour groups (Gr.):

$$D = -2\ln\left(\frac{P(D_g = d_{g,[T\cup AN]} \mid \text{null}_g)}{P(D_g = d_{g,[T]} \mid \text{T model}_g) * P(\text{Gr.} = T) * P(D_g = d_{g,[AN]} \mid \text{AN model}_g) * P(\text{Gr.} = AN)}\right). \quad (8)$$

The T and AN gene models are trained using only tumour or only adjacent normal samples, respectively. The null model is trained using samples from both groups. The significance of the D statistic is evaluated based on its random expectation, as obtained by permuting sample labels rather than using the standard chi-squared distribution. We use an upper-tailed Z-test for final significance evaluation in which we compute the Z statistics as follows:

$$Z = \frac{D - E[D]}{\sigma(D)}. \quad (9)$$

We control the false discovery rate using the Benjamini & Hochberg procedure (Benjamini and Hochberg, 1995) in the differential analysis of expression, methylation or joint expression and methylation data across all genes.

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 adjacent normal, 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: $P(\mathrm{Gr.} = \mathrm{T}|D_{g,s} = d_{g,s})$ and $P(\mathrm{Gr.} = \mathrm{AN}\,|D_{g,s} = d_{g,s})$ (Eqs. 10 and 11).

$$P(\mathrm{Gr.} = \mathrm{T}|D_{g,s} = d_{g,s}) = \frac{P(D_{g,s} = d_{g,s}|\mathrm{T\ model}_g) * P(C = T)}{P(D_{g,s} = d_{g,s}|\mathrm{T\ model}_g) * P(\mathrm{Gr.} = \mathrm{T}) + P(D_{g,s} = d_{g,s}|\mathrm{AN\ model}_g) * P(\mathrm{Gr.} = \mathrm{AN})} \quad (10)$$

$$P(\mathrm{Gr.} = \mathrm{AN}\,|D_{g,s} = d_{g,s}) = 1 - P(\mathrm{Gr.} = \mathrm{T}|D_{g,s} = d_{g,s}) \quad (11)$$

The prior probability would typically reflect the expected proportion of normal samples $P(\mathrm{Gr.} = \mathrm{AN})$ versus the proportion of tumour ples $P(\mathrm{Gr.} = \mathrm{T})$. 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:

$$P(\mathrm{Gr.} = \mathrm{T}|D_{[G],s} = d_{[G],s}) = \frac{\prod_{g \in G} P(D_{g,s} = d_{g,s}|\mathrm{T\ model}_g) * P(C = T)}{\prod_{g \in G} P(D_{g,s} = d_{g,s}|\mathrm{T\ model}_g) * P(\mathrm{Gr.} = \mathrm{T}) + \prod_{g \in G} P(D_{g,s} = d_{g,s}|\mathrm{AN\ model}_g) * P(\mathrm{Gr.} = \mathrm{AN})} \quad (12)$$

$$P(\mathrm{Gr.} = \mathrm{AN}\,|D_{[G],s} = d_{[G],s}) = 1 - P(\mathrm{Gr.} = \mathrm{T}|D_{[G],s} = d_{[G],s}) \quad (13)$$

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

2.4 Comparison to existing methods

We compare PINCAGE's performance with established methods within differential methylation and 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 data set 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.

For sample classification, we compare against the Logistic Regression (LR). We use the normalized gene expression (RPM), and overall regional gene body and promoter methylation levels as predictors, without any interaction terms. Logistic regression classifiers involving multiple genes independently include the set of corresponding expression and methylation predictors.

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 data set 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 to normals.

We first looked at the global distributions of the three data types. The overall expression profiles of tumours and adjacent normal samples were

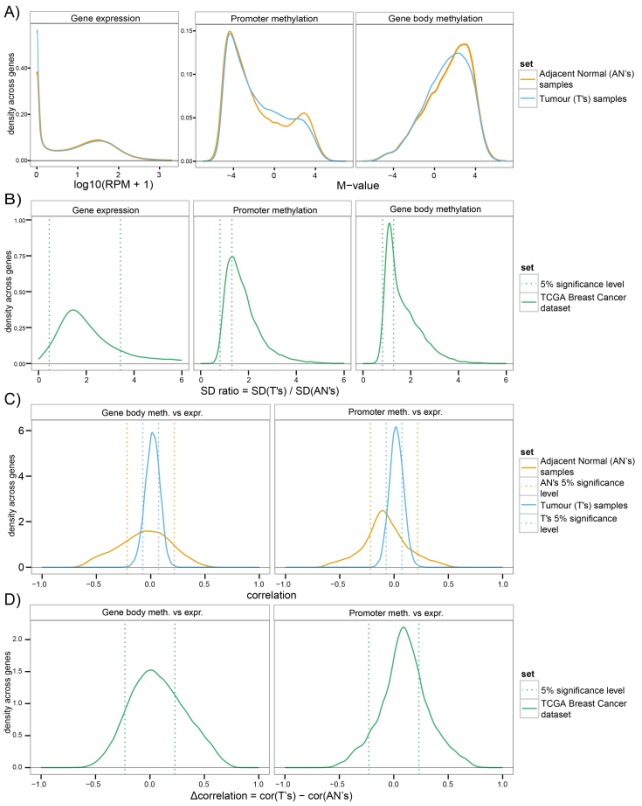


Fig. 4 Expression and methylation profiles in BRCA. **A)** Global distributions of expression levels, measured in reads per million (RPM), and mean methylation levels (M-value) across promoter and gene body regions for both groups across samples. **B)** Distribution of gene-wise standard deviation ratios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

similar, though tumours showed a relative increase in the number of lowly expressed genes (Fig. 4 A, 1.25x 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 adjacent normals (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 adjacent normals, 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 adjacent normal 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 adjacent normal samples (Fig. 4 B). We defined 5% significance levels using genome-wide random expectation (Fig. S2 A) by repeatedly (n=10) permuting sample labels genome-wide. Significantly increased variability in tumours compared to normals was more often seen for the methylation data types (71.3% of gene bodies and 58.5% of promoters) than expression (12.9%).

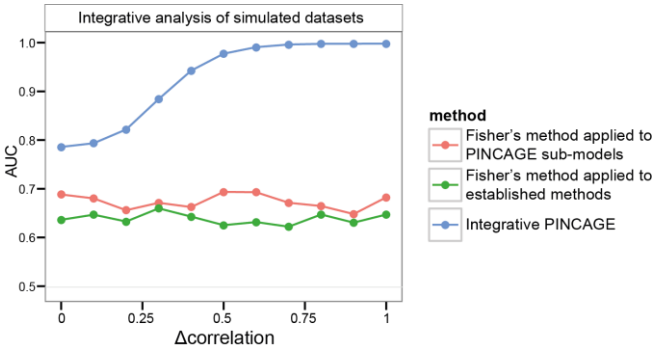


Fig. 5 The performance effect on simulated data sets of tumour correlation perturbation. The effect on performance (AUC) of changes in the correlation level between methylation and expression between tumour and normal samples (Δcorrelation).

We next evaluated the gene-specific correlation of promoter and gene body methylation with expression (Fig. 4 C) to further motivate separation of these relationships. Gene body methylation was primarily negatively correlated with expression in the adjacent normals (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 correla tion existed, with no general rule, though tumours generally showed less extreme levels of correlation between methylation and expression than adjacent normal samples.

We further quantified the significant fraction of gene-specific expression-methylation correlations at 5% significance levels (Fig. 4 C) using the group-specific random expectations (Fig. S2 B). The significant fractions of gene expression correlation with promoter and gene body methylation were generally larger in adjacent normal 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 adjacent normal groups (Fig. 4 D). More genes show significant positive correlation changes than expected by random (Fig. S2 C) 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. 4 C). 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 multimodal 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 data sets as follows. We first simulated data sets 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 & 9). The overall performance was quantified using the area under the receiver operating characteristic curve (AUC). Each data set consisted of an equal number (n=100) of tumour and normal samples with values for all three data types simulated for 2,000 genes (SUPPLEMENT: *Simulation procedure description*). The parameters of the simulation (Table S1) were chosen to resemble the values observed for the BRCA data set (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 (SUPPLEMENT: *Evaluation of heterogeneity simulated data sets*). The performance was both better initially when the signal is the purest and degraded more slowly as the fraction of perturbed tumour samples decreases (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 data sets with constant levels of correlation between expression and methylation in the normal samples and varying levels in the tumours, as seen in the BRCA data set (Fig. 4 C 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 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 normal samples

We next used PINCAGE to detect perturbed genes across all genes of the BRCA data set. 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 adjacent normal 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, with nearly the same number of perturbed genes detected by the integrative PINCAGE model (n=16,276) and 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 PINCAGE and the combination of established methods (Fig. S6). We evaluated the set of candidate genes from the original TCGA study of the BRCA set (Cancer Genome Atlas, 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 genes by either method (Table S2). Also, the differences in the gene ranking between methods were insignificant (Table S2). 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 to 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 (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; Fig. S6) and the combination of individual PINCAGE sub-models (cor=0.747; Fig. S7 A). The analysis using integrative PINCAGE correlated strongly with the combination of PINCAGE sub-models (cor=0. 868; Fig. S7 B). 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.

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 (Refs in Table S3). Marginal distributions of the PLK1 (Figs. 2 and S8, PLK1) show pronounced changes between disease groups. Also, patterns of differential correlation were seen between groups.

Table 1 Integrative PINCAGE model top-10 most significantly perturbed genes in BRCA and their ability to classify tumour and normal samples. * signifies known role in cancer. ** signifies known role in breast cancer.

Significance evaluation of BRCA data set (55 AN's vs 487 T's)				Classification performance on BRCA validation subset (27 AN's and 243 T's)			
	Integrative PINCAGE		Established methods combined	Integrative PINCAGE		Logistic regression using PINCAGE-identified genes	
Gene ID	Z-score	Rank (k)	Rank	AUC of single gene model	AUC using running combination of genes (1-k)	AUC of single gene model	AUC using running combination of genes (1-k)
RAG1AP1*	115.70	1	773	0.9311	0.9311	0.9813	0.9813
CPA1*	114.92	2	96	0.9297	0.9747	**0.9960**	**0.9989**
NEK2**	112.56	3	446	0.9291	0.9927	0.9720	0.9986
RNASEH2A**	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	0.9962
TIMM17A**	102.79	7	1664	0.9576	0.9977	0.9497	0.9198
PLK1**	99.95	8	496	0.9427	0.9970	0.9709	0.9290
RABIF*	98.58	9	1441	0.9531	0.9988	0.9694	0.9156
PTF1A*	98.45	10	1577	**0.9806**	0.9988	0.9561	0.9070

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. The PTF1A encodes the subunit alpha of pancreas transcription factor 1, which is involved in cell fate determination in various organs and is causally implicated in exocrine pancreatic cancer. Although the expression of PTF1A is lost in exocrine pancreatic cancer and is normally unexpressed in breast tissue, we observed an activation of transcription of this gene in some of the BRCA tumour samples (Fig. S8, PTF1A Expression). It is highly differentially methylated in both promoter and gene body 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 (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 gene body, Welch's t-test).

Classification of tumour versus normal

We explored PINCAGE's classification performance on the top-10 most significant genes and compared it against that of logistic regression (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 (Table S4). Several genes among the top-10 are of relevance to breast cancer (Refs in Table S5), however, their ranking is primarily driven by changes in gene expression between cancers and adjacent normal samples, rather than by joint expression-methylation gene perturbation (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 (Table S8). 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 integrative PINCAGE model as well as the combination of established methods (Table 2, left-hand side). PINCAGE identified fewer (n=95) statistically significant genes at 1% FDR than established methods (n=234). The reason could be the low sample count, which limits the power of the parameter-rich PINCAGE models. Among the top-10 most significantly perturbed genes, the distributions of observations are complex for both groups (Figure S10) and classification based on individual data types appears difficult.

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 (a) rank genes according to significance evaluation and (b) 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=7.8e-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.

ZNF706 and SERPINE3

The most consistently top-ranked genes (22 of 28 possible positions in the top-2; Table S6) in the 14-fold cross-validation procedure were the Serpin Peptidase Inhibitor Member 3 (SERPINE3) and the Zinc Finger Protein 706 (ZNF706). Neither has previously been linked to breast cancer. ZNF706 is a zinc finger transcription factor with limited characterization in the literature; however, it was found up-regulated in Laryngeal Squamous

Table 2 Left: Top-10 ranked genes in the BRCA progression data set**. Right**: Comparison of classification performance for integrative PINCAGE, logistic regression on PINCAGE-identified genes, and logistic regression on genes found by combination of established methods with Fisher's method.

Significance evaluation of progression set (14 progressing and 57 non-progressing tumours)			
	Integrative PINCAGE		Established methods combined
Gene ID	Z-score	Rank	Rank
SERPINE3	11.46	1	251
ZNF706	8.75	2	752
ACTN2	6.90	3	1518
AKR1B15	6.75	4	714
AGBL3	6.47	5	5645
LOC100240734	6.19	6	931
MYL10	6.13	7	5869
NDUFA9	6.04	8	9953
HIGD1B	5.84	9	311
ARG1	5.74	10	614

Classification performance on progression set using 14-fold cross-validation						
	Integrative PINCAGE		Logistic regression using genes found by integrative PINCAGE		Logistic regression using genes found by combination of established methods	
Rank at each fold (k)	AUC of single gene model	AUC using running combination of genes (1-k)	AUC of single gene model	AUC using running combination of genes (1-k)	AUC of single gene model	AUC using running combination of genes (1-k)
1	**0.8008**	0.8008	**0.7431**	0.7431	**0.7055**	**0.7055**
2	0.6316	**0.8358**	0.7043	0.7895	0.4624	0.6291
3	0.6629	0.6742	0.5990	0.7143	0.4912	0.6253
4	0.4818	0.7055	0.5689	0.7406	0.5564	0.5376
5	0.6216	0.6491	0.6654	0.4091	0.4950	0.4787
6	0.6685	0.6805	0.6967	0.7105	0.6190	0.5526
7	0.6291	0.6366	0.5338	0.7375	0.5714	0.5764
8	0.4524	0.6479	0.5426	0.8296	0.5175	0.6109
9	0.5188	0.6378	0.5927	**0.8860**	0.5815	0.5013
10	0.5188	0.6253	0.5025	0.7162	0.5414	0.5263

Cancer (Colombo, et al., 2009). We also found it consistently upregulated in tumours versus normals in the BRCA set (2.02e-08 nominal p-value; edgeR). In the progression data set, its gene body and promoter methylation levels were significantly correlated with gene expression (Spearman's coefficients of 0.24 and -0.25, respectively). 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; 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., 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 (3.15e-5 nominal p-value, Welch's t-test), but remains very lowly expressed in both groups (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 nominal p-values; edgeR), though the functional interpretation and possible relation to SERPINE3 is not known.

Though 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.

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 adjacent normal samples in the Breast Invasive Carcinoma (BRCA) data set 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 methods for read counts data analysis (Anders and Huber, 2010; Robinson and Smyth, 2008) that introduce a gamma prior to account for the overdispersion of Poisson-distributed read counts, we also model the overdispersion, however, using the gene-specific empirical priors instead. This improves model fits in the analysis of cancer data sets known for high variance. To our knowledge, no other method models the overdispersion in the integrative context. Benefits of such integrated data analysis are twofold. First, it enables detection of subtle simultaneous deviations of all three variables that would be too weak to become significant if analysed separately. Also, 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 the assignment of samples to trained groups.

Our use of Gaussian kernels to specify the joint 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 to parameter-sparse models. In this respect, our BRCA progression data set could greatly benefit from more patients being 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 stratifying 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.

Integrative cancer genomics analysis has received growing attention over the last years, but much work remains. With the advent of large publically

available data sets, 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 more molecular data has to be met with increased quality of clinical data as well to facilitate discovery of clinically-actionable findings from such studies. This will facilitate molecular clinical research oriented towards better diagnosis, prognosis and treatment for patients. Further studies are required to confirm the robustness of integrative biomarker candidates, and to test how well they generalize across cohorts.

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.

ACKNOWLEDGEMENTS

Funding: This work was supported by The Danish Strategic Research Council (Innovation Fund Denmark) and the Sapere Aude program of the the Danish Council for Independent Research | Medical Sciences.
Conflicts of Interest: none declared.

REFERENCES

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 biology*, **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 Roy Stat Soc B Met*, **57**, 289-300.

Berse, B. and Lynch, J.A. (2015) Molecular diagnostic testing in breast cancer, *Seminars in oncology nursing*, **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, N. (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., Delong, D.M. and Clarkepearson, D.I. (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 fron independent experiments., *Journal of Psychology*, **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*. Biological monographs and manuals,. Oliver and Boyd, Edinburgh,.

Forbes, S.A., *et al.* (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC), *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, **Chapter 10**, Unit 10 11.

Gelfman, S., *et al.* (2013) DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure, *Genome Res*, **23**, 789-799.

Hamid, J.S., *et al.* (2009) Data integration in genetics and genomics: methods and challenges, *Human genomics and proteomics : HGP*, **2009**.

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

Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) The elements of statistical learning : data mining, inference, and prediction. In, *Springer series in statistics,*. Springer, New York, NY, pp. xxii, 745 p.

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, *Nature reviews. Genetics*, **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 research*, **66**, 8927-8930.

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

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

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

Neyman, J. (1933) On the problem of the most efficient tests of statistical hypotheses, *Philos T R Soc Lond*, **231**, 289-337.

Parrella, P. (2010) Epigenetic Signatures in Breast Cancer: Clinical Perspective, *Breast care*, **5**, 66-73.

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

Ray, B., *et al.* (2014) Information content and analysis methods for multi-modal high-throughput biomedical data, *Scientific reports*, **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 research*, **72**, 1170-1181.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139-140.

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

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.

Shen, R.L., Olshen, A.B. and Ladanyi, M. (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 Huber, R.G.a.V.C.a.S.D.a.R.I.a.W. (ed), *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 review of molecular diagnostics*, **10**, 49-64.

Strand, S.H., Orntoft, T.F. and Sorensen, K.D. (2014) Prognostic DNA methylation markers for prostate cancer, *International journal of molecular sciences*, **15**, 16544-16576.

Tibshirani, R. (2011) Regression shrinkage and selection via the lasso: a retrospective, *J R Stat Soc B*, **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**, i237-245.

Venables, W.N., Ripley, B.D. and Venables, W.N. (2002) *Modern applied statistics with S*. Statistics and computing. 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.

Weiss, R. (2005) NIH Launches Cancer Genome Project. *Washington Post.*

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 biology*, **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, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 3700-3705.

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.

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

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 : the journal of biological databases and curation*, **2011**, bar026.

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

SUPPLEMENT for:

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

Michał P. Świtnicki, Malene Juul, Tobias Madsen, Karina D. Sørensen , Jakob S. Pedersen

Table of Contents

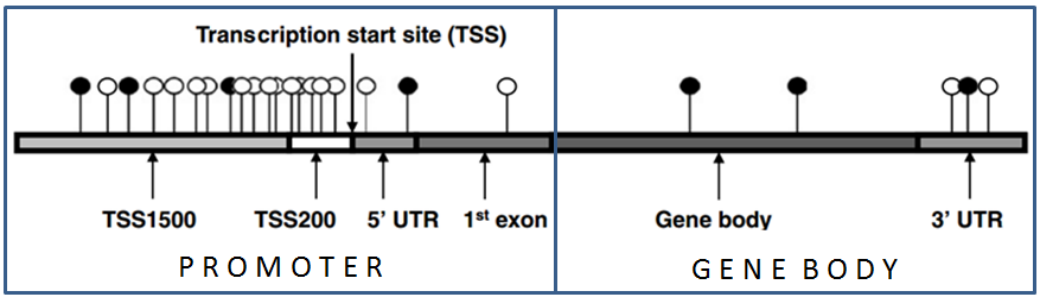


Figure S1 Division of 450k platform probe annotations between 2 functional groups. Adapted and reprinted from (Bibikova, et al., 2011) with permission from Elsevier.

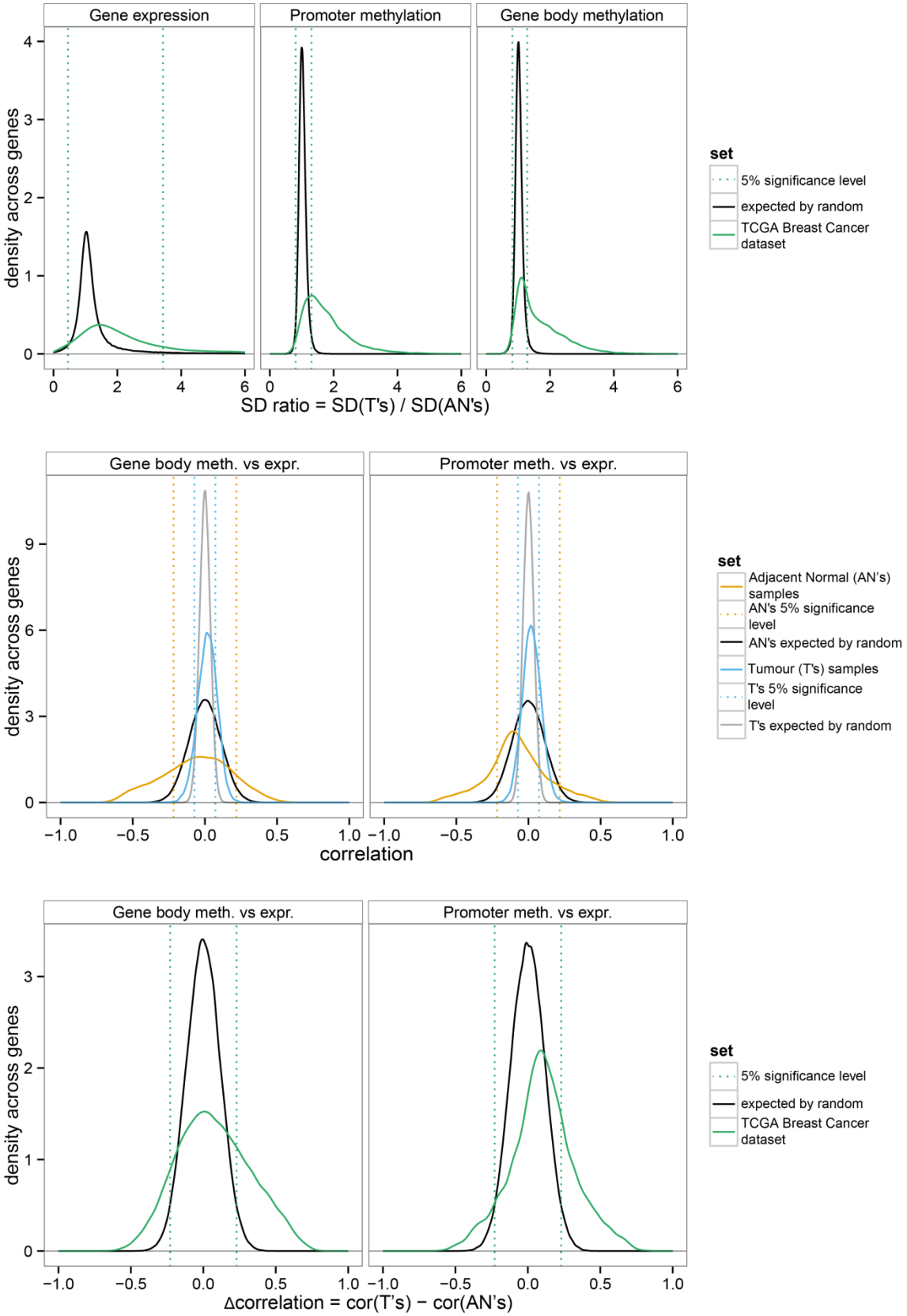


Figure S2 A) Distribution of gene-wise standard deviation ratios between T's and AN's of the expression (RPM), gene body and promoter methylation (M-value) variables. The random expectation was obtained by repeating the analysis 10 times on a null set of samples, which had their sample labels permuted. Same procedure was used in the panels below. **B**) Correlations between promoter and gene body methylation and gene expression for each gene across the entire BRCA data set for AN's and T's. **C**) Gene-wise changes of correlations observed between the AN's and T's.

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 conditional methylation given expression distributions become multinomial distributions with parameters specified using grid Gaussian kernels as implemented in the AWS R package (Polzehl and Spokoiny, 2006) (see below: *PLK1 example of training process*, Figure S3). The discretized versions of the joint probability distributions sum, rather than integrate, out the unobserved random variables (SUPPLEMENT, Eqs. 14, 15, 16, and 17). 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. 4 B). For a given gene, cancer samples often show much heterogeneity, with some behaving like normal tissue while others are perturbed in various ways (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 (Fig. S8, RAG1AP1, CPA1, PLK1).

Discretized version of sub-model and integrative model factorizations:

$$\mathrm{P}\left(\mathrm{M}_{\mathrm{g}}^{\mathrm{P.CpG}}\right) = \prod_{\mathrm{s}=1}^{\mathrm{n}} \left(\sum_{\mathrm{k}=1}^{\mathrm{d^P}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g}}^{\mathrm{P}} = \mathrm{m}_{\mathrm{g,s,k}}^{\mathrm{P}}\right) \prod_{\mathrm{v}=1}^{\mathrm{n^P}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g,s,v}}^{\mathrm{P.CpG}} = \mathrm{m}_{\mathrm{g,s,k,v}}^{\mathrm{P.CpG}} \middle| \mathrm{m}_{\mathrm{g,s,k}}^{\mathrm{P}}\right) \right), \quad (14)$$

$$\mathrm{P}\left(\mathrm{M}_{\mathrm{g}}^{\mathrm{GB.CpG}}\right) = \prod_{\mathrm{s}=1}^{\mathrm{n}} \left(\sum_{\mathrm{l}=1}^{\mathrm{d^{GB}}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g}}^{\mathrm{GB}} = \mathrm{m}_{\mathrm{g,s,l}}^{\mathrm{GB}}\right) \prod_{\mathrm{v}=1}^{\mathrm{n^{GB}}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g,s,v}}^{\mathrm{GB.CpG}} = \mathrm{m}_{\mathrm{g,s,l,u}}^{\mathrm{GB.CpG}} \middle| \mathrm{m}_{\mathrm{g,s,l}}^{\mathrm{GB}}\right) \right), \quad (15)$$

$$\mathrm{P}\left(\mathrm{E_g}, \mathrm{R_g}; \mathrm{r}\right) = \prod_{\mathrm{s}=1}^{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{d^E}} \mathrm{P}\left(\mathrm{E_{s,g}} = \mathrm{e_{s,g,j}}\right) \mathrm{P}\left(\mathrm{R_{s,g}} \middle| \mathrm{e_{s,g,j}}, \mathrm{r_{s,.}}\right), \quad (16)$$

$$\mathrm{P(D)} = \prod_{\mathrm{s}=1}^{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{d^E}} \left(\begin{array}{c} \mathrm{P}\left(\mathrm{E_{s,g}} = \mathrm{e_{s,g,j}}\right) \mathrm{P}\left(\mathrm{R_{s,g}} \middle| \mathrm{e_{s,g,j}}, \mathrm{r_{s,.}}\right) \\ \sum_{\mathrm{k}=1}^{\mathrm{d^P}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g}}^{\mathrm{P}} = \mathrm{m}_{\mathrm{g,s,k}}^{\mathrm{P}} \middle| \mathrm{e_{s,g,j}}\right) \prod_{\mathrm{v}=1}^{\mathrm{n^P}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g,s,v,k}}^{\mathrm{P.CpG}} = \mathrm{m}_{\mathrm{g,s,v,k}}^{\mathrm{P.CpG}} \middle| \mathrm{m}_{\mathrm{g,s,k}}^{\mathrm{P}}\right) \\ \sum_{\mathrm{l}=1}^{\mathrm{d^{GB}}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g}}^{\mathrm{GB}} = \mathrm{m}_{\mathrm{g,s,l}}^{\mathrm{GB}} \middle| \mathrm{e_{s,g,j}}\right) \prod_{\mathrm{v}=1}^{\mathrm{n^{GB}}} \mathrm{P}\left(\mathrm{M}_{\mathrm{g,s,v,l}}^{\mathrm{GB.CpG}} = \mathrm{m}_{\mathrm{g,s,v,l}}^{\mathrm{GB.CpG}} \middle| \mathrm{m}_{\mathrm{g,s,l}}^{\mathrm{GB}}\right) \end{array} \right), \quad (17)$$

where *s* denotes the sample, *g* denotes the gene under consideration, $\mathrm{E_g}$ is a random variable denoting the normalized expression state, d^E is the number of bins used to discretize normalized expression, $r_{s,g}$ is the observed read count for the current gene and $r_{s,.}$ is the total number of reads in the library of sample *s*, $\mathrm{M}_{\mathrm{g}}^{\mathrm{P.CpG}}$ is a random variable denoting the individual promoter-

belonging probe, $\mathrm{M}_{\mathrm{g}}^{\mathrm{GB.CpG}}$ is a random variable denoting the individual gene body-belonging probe, $\mathrm{M}_{\mathrm{g}}^{\mathrm{P}}$ is a random variable denoting promoter methylation state, $\mathrm{M}_{\mathrm{g}}^{\mathrm{GB}}$ is random variable denoting gene body methylation state with d^P and d^{GB} being numbers of bins used to discretize promoter and gene body methylation states, respectively.

PLK1 example of training process

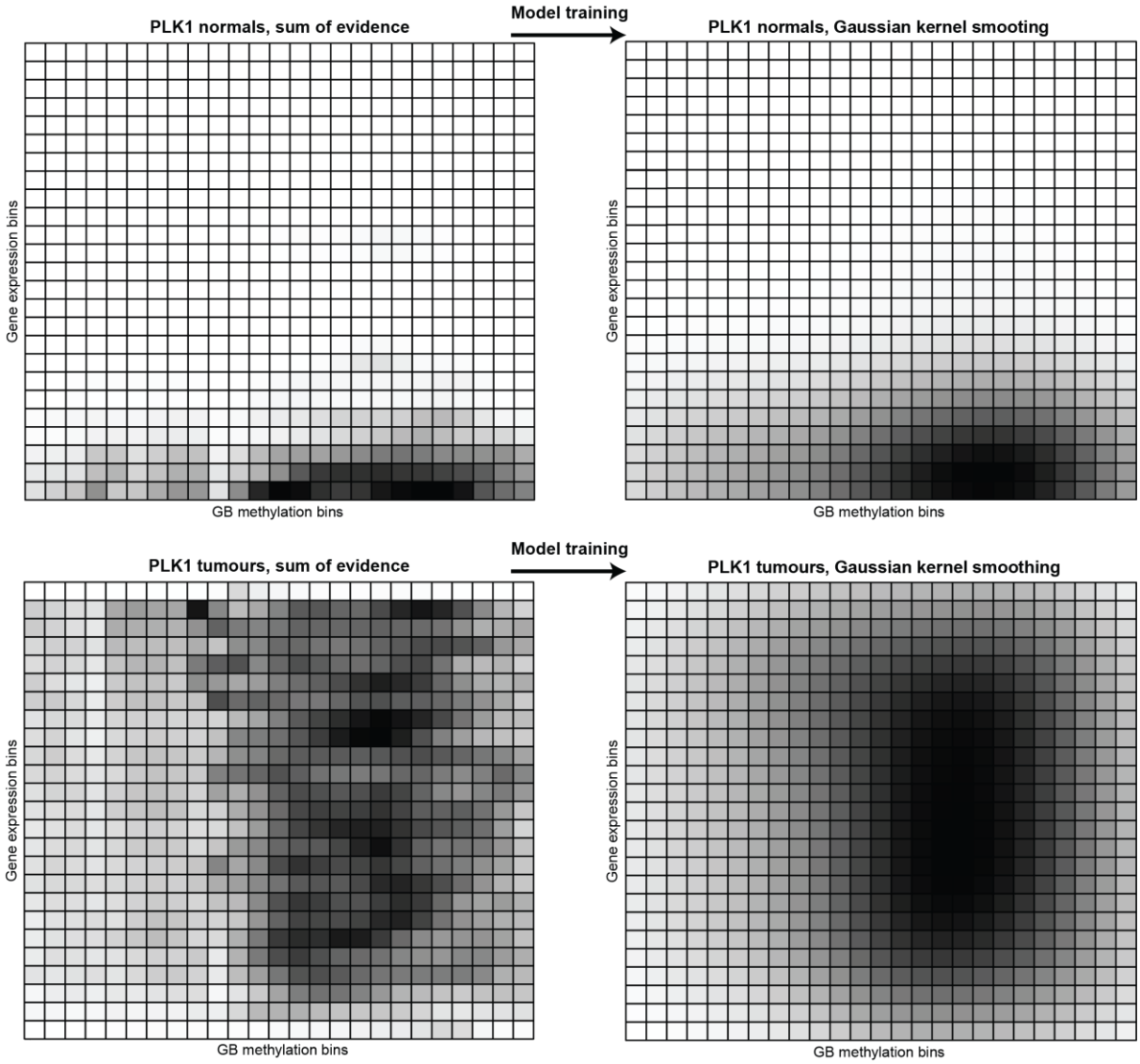


Figure S3 Illustration of the training process using joint distributions of discretized gene expression and gene body methylation data types. The darker the cell, the higher the accumulated evidence/probability in the distribution. For illustrative purposes, we show the joint distribution of the two data types, rather than the conditional distribution, which is required for the PINCAGE model and is calculated by normalizing rows to obtain $\sum_{\mathrm{l}=1}^{\mathrm{d^{GB}}} \mathbf{P}\left(\mathbf{M}_{\mathrm{g}}^{\mathbf{GB}} = \mathbf{m}_{\mathrm{g,s,l}}^{\mathbf{GB}} \mid \mathbf{e}_{\mathrm{s,g,j}}\right)$.

Simulation procedure description

We compare the performance of PINCAGE and established methods by applying them to simulated data sets with known characteristics. We generate these data sets using the following approach. We repeatedly simulate 100 “tumour” and 100 “normal” samples per gene in each investigated setting. The overall number of simulated genes in each condition is 2000: half of which are true cases and half of which are negative cases. This enables the calculation of AUCs.

In this overall scheme, we vary one parameter of the simulation to generate a series of conditions stratified by that parameter.

The data types of each gene (expression, promoter and gene body methylations) are at first simulated according to the Multivariate Gaussian Distribution specification of the *mvrnom* function from the R MASS package (Venables, et al., 2002). This step enforces the desired correlations between the data types for each simulated gene and sample group. Following that, we transform each variable to the given mean and standard deviation, reflecting the desired group-wise changes. Finally, Gaussian noise is added to the expression and the 11 CpGs (same number as the average per region in the 450k platform) that are simulated independently for each methylation type from its point estimates.

The procedure is varied for generating the two data sets that are investigated in this publication. The first data set is stratified by the fraction of truly affected samples in the "tumours" group – here referred to as the dilution data set. In this data set, only a fraction of samples in the "tumours" group is simulated according to the characteristics of that group. The remaining samples are simulated according to the "normals" group specification. The second data set is stratified by the group-wise change of degree of correlation between expression and the two inversely correlated methylation types. This is referred to as the Δcorrelation data set.

In both data sets, the true samples' group-wise changes of expression and methylation are fixed. What varies between data sets is the amount of added noise (Table S1). This ensures that we can observe the differential behaviour of methods as parameters are stratified in each data set. Simulation parameters are selected so that differential behaviour can be observed across stratifications. Standard deviations for group dispersion and Gaussian noise were kept relative to and consistently larger than what our submodels assume to ensure the generated data sets do not favour them.

Table S1 Parameters used to simulate genes in the two data sets.

Parameter / Data set /group	Mean unnormalized Expr.	Group St. dev. of Expr.	Noise St. dev. of Expr.	Mean Pr. meth.	Mean GB meth	Group St. dev. of GB and Pr. meth.	11 CpG St. dev.	Noise St. dev. of GB and Pr. 11CpGs	Fraction of true samples	Pr./GB meth. correlation with Expr.
Fraction perturbed /"normals"	500	2*√500	0.5*√500	0	0	0.28	0.28	0.093	1	-0.5 / 0.5
Fraction perturbed /"tumours"	502	2.5*√500	0.5*√500	-0.015	0.015	0.35	0.35	0.093	[1, 0]	0 / 0
Δcorrelation /"normals"	500	2*√500	√500	0	0	0.28	0.28	0.14	1	-0.5 / 0.5
Δcorrelation /"tumours	502	2.5*√500	√500	-0.015	0.015	0.35	0.35	0.14	1	[-0.5,0.5] / [0.5,-0.5]

Evaluation of heterogeneity simulated data sets

We evaluated how the detection of perturbed genes changed if only a fraction of the tumour samples were truly perturbed, to evaluate the effect of intertumour heterogeneity. We

simulated a series of data sets where a gradually smaller fraction of tumour samples had perturbed genes, with the rest resembling the normal samples. As expected, the performance consistently decreased with decreasing fraction of truly perturbed tumour samples (Figure S4). Compared to the established methods (edgeR for gene expression, Welch's t-test for methylation), the individual data type models had better performance throughout in this scenario (Figure S4 A). The performance was both better initially when the signal is the purest and degraded more slowly as the fraction of perturbed tumour samples decreases. We attribute PINCAGE's greater robustness to tumour sample heterogeneity in this setting to its ability to model the resulting multi-modal distributions.

We next asked how the integration of the three different data types affected the overall performance in the tumour heterogeneity setting. We used Fisher's method to combine p-values both for the individual data type models and for the established methods. We also applied the integrative PINCAGE model to the joint data sets, which performed similarly to the individual data type models combined in this setting. The integration of data types improved performance throughout (Figure S4 B). Both the PINCAGE models and the established methods showed much greater robustness to a decrease in the fraction of perturbed samples than when only a single data type was included. Integration of the three data types thus consistently improves performance.

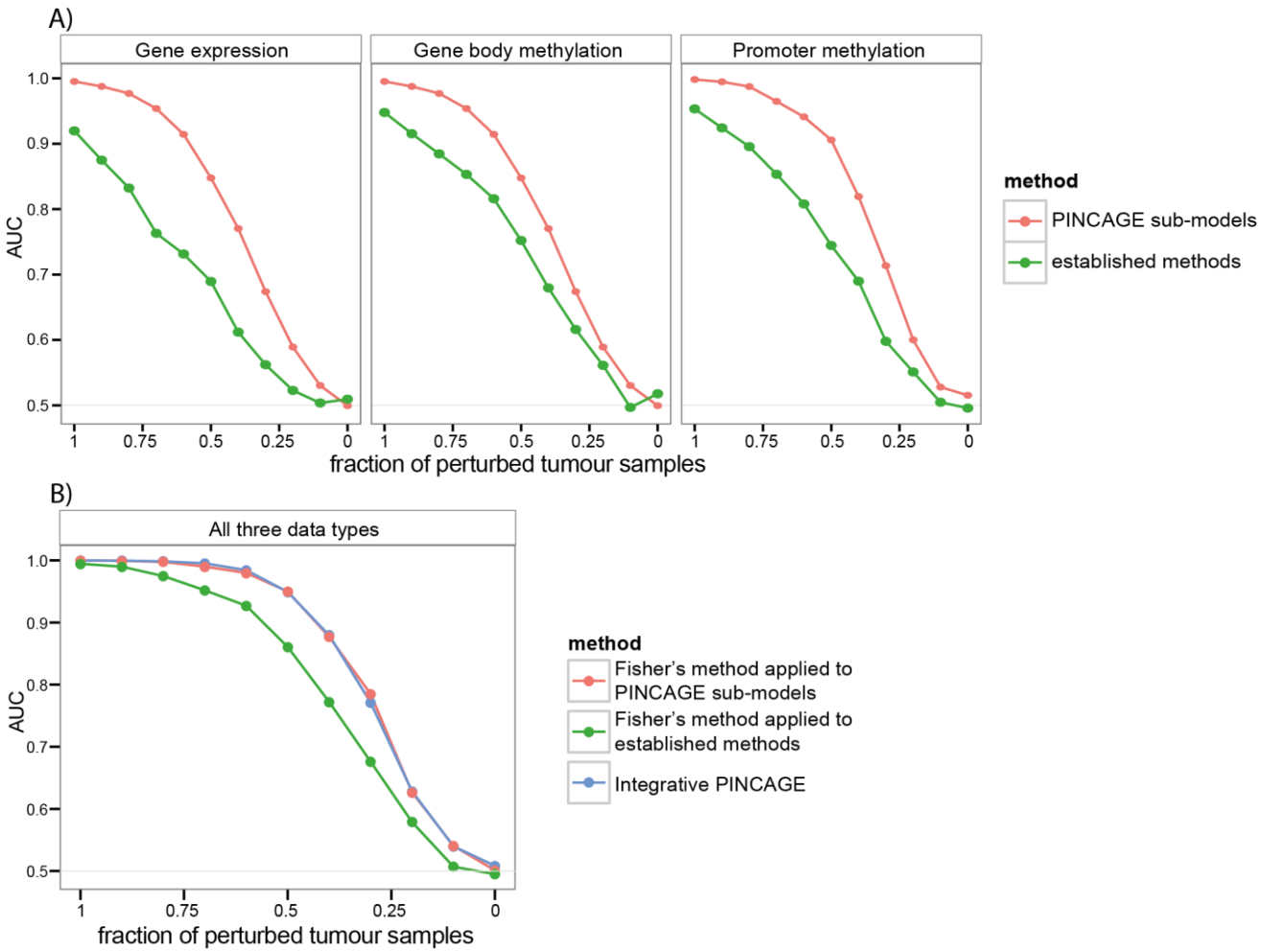


Figure S4 The performance effect on simulated data sets of tumour cohort heterogeneity and correlation perturbation. **A)** Performance measured by area under the receiver operating characteristic curve (AUC) as a function

of the fraction of perturbed tumour samples evaluated for both the individual data-specific PINCAGE sub-models and for the established methods. **B)** Integrated analysis of all three data types shown in (A) using Fisher's method on individual PINCAGE sub-models and established methods as well as results from applying the integrative PINCAGE model.

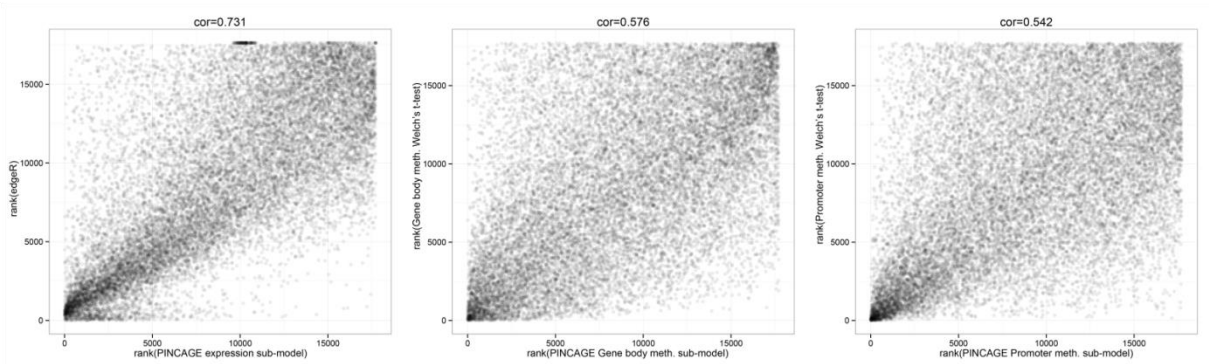


Figure S5 Scatterplots of ranks of genes in the BRCA data set as analysed by different established approaches for expression data (first panel), and gene body and promoter methylation data (second and third, respectively) and corresponding PINCAGE sub-models.

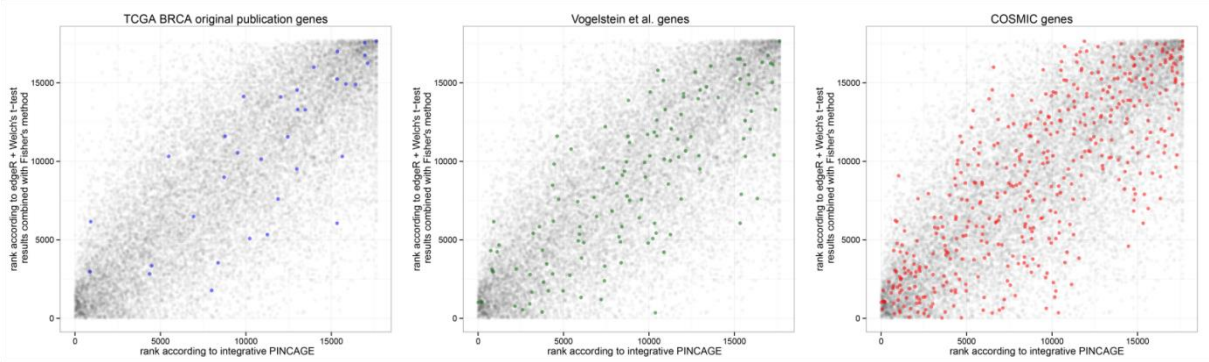


Figure S6 Scatterplots of gene ranks in the BRCA data set obtained with integrative PINCAGE and with Fisher's method combining established methods. Spearman's correlation coefficient between the two approaches is 0.742. The 3 plots contain colouring of genes according to their membership to the panels defined by (Cancer Genome Atlas, 2012) in the first plot, by (Vogelstein, et al., 2013) in the second and by (Forbes, et al., 2008) in the third one.

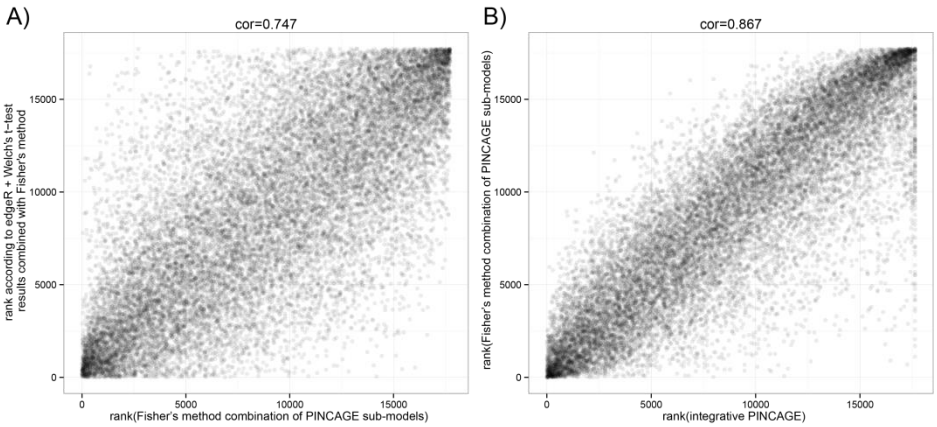


Figure S7 Scatterplot of ranks of genes in the BRCA data set. **A)** Fisher's method combination of PINCAGE sub-models and the Fisher's method combination of established methods. Spearman's correlation coefficient between the two is 0.747. **B)** Integrative PINCAGE and the Fisher's method combination of individual PINCAGE sub-models. Spearman's correlation coefficient between the two is 0.867.

Table S2 Significance evaluations of ranks produced by the integrative PINCAGE and the Fisher's method combining established methods on different sets of cancer-related genes (Wilcoxon rank sum test). **Right-hand side:** Comparison of ranks produced by both methods (Mann-Whitney test).

Evaluation / Gene set	PINCAGE	Combination of established methods	PINCAGE vs combination of established methods
COSMIC	0.7358	0.3294	0.7661
Vogelstein et al.	0.8493	0.3296	0.8425
TCGA BRCA	0.9977	0.9758	0.7125

Table S3 Top-10 genes according to integrative PINCAGE evaluation of BRCA tumours vs normals. * signifies known role in cancer. ** signifies known role in breast cancer.

Gene ID	References
RAG1AP1*	(Eeles, et al., 2013)
CPA1*	(Matsugi, et al., 2007)
NEK2**	(Brendle, et al., 2009; Liu, et al., 2012; Pitner and Saavedra, 2013; Tsunoda, et al., 2009; Wang, et al., 2012)
RNASEH2A**	(Shah, et al., 2009)
LOC148145	NA
TMEM63B	NA
TIMM17A**	(Salhab, et al., 2012; Xu, et al., 2010)
PLK1**	(Maire, et al., 2013; Uckun, et al., 2007; Valsasina, et al., 2012)
RABIF*	(Tang and Ng, 2009)
PTF1A*	(Adell, et al., 2000; Sellick, et al., 2004)

Table S4 Logistic regression using genes found by combination of established methods in the course of comparison between 55 AN's vs 487 T's in the BRCA data set.

Classification performance on BRCA validation subset (27 AN's and 243 T's)		
Gene ID	AUC of single gene model	AUC using running combination of genes (1-k)
FHL1	0.9718	0.9718
GPAM	0.9547	0.9710
LYVE1	0.9893	0.9840
SORBS1	0.9867	**0.9907**
TNS1	0.9913	0.9278
GYG2	0.9346	0.9342
KLB	0.9570	0.9474
ACVR1C	0.9435	0.9262
KCNIP2	0.9547	0.9477
CAV1	**0.9947**	0.9509

Table S5 Top-10 genes according to the established methods combined with Fisher's method. * signifies known role in cancer. ** signifies known role in breast cancer.

Gene ID	References
FHL1**	(Zhang, et al.)
GPAM**	(Brockmoller, et al.)

LYVE1**	(Timoshenko, et al., 2006; V, et al.; Van der Auwera, et al.)
SORBS1	NA
TNS1*	(Martuszewska, et al.)
GYG2**	(Harris, et al.)
KLB*	(Poh, et al.)
ACVR1C**	(Zeng, et al.)
KCNIP2	NA
CAV1**	(Pinilla, et al.; Van den Eynden, et al.)

Table S6 Top-10 genes according to the integrative PGM at each fold of the cross-validation procedure. In bold: genes discussed in main text.

Top \ Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	**SERPINE3**	**SERPINE3**	**SERPINE3**	DFFA	**ZNF706**	**SERPINE3**	**SERPINE3**	**SERPINE3**	**SERPINE3**	**SERPINE3**	**SERPINE3**	**SERPINE3**	**SERPINE3**	**ZNF706**
2	**ZNF706**	AGBL3	ARG1	**SERPINE3**	**SERPINE3**	**ZNF706**	AGBL3	PSG1	HIGD1B	**ZNF706**	**ZNF706**	**ZNF706**	**ZNF706**	**SERPINE3**
3	FBXO15	AKR1B15	**ZNF706**	DPY30	CEP78	COX7A2L	HIST1H1B	TSGA10IP	SULT1A3	DCT	HIGD1B	HIGD1B	PPARGC1A	MYL10
4	NDUFA9	HIST1H1B	FAM27L	MYL10	MRPS2	AGBL3	**ZNF706**	**ZNF706**	**ZNF706**	BGLAP	NDUFA9	NUP62	DPY30	LOC347376
5	LSM10	MYL10	COX7A2L	COX7A2L	JPH4	DPY30	ARG1	IDH3B	WNK3	ZNF641	AKR1B15	NUDCD3	EPHX2	IDH3B
6	AKR1B15	**ZNF706**	AGBL3	**ZNF706**	AKR1B15	HIGD1B	PSG1	AKR1B15	AGBL3	EPHX2	MYL10	CEP78	LONP2	TNIK
7	REM2	ACTN2	NDUFA9	AKR1B15	RPS13	NUDCD3	SNORA9	AGBL3	EPHX2	TBL3	ZSCAN4	FBXO15	COX7A2L	AKR1B15
8	AGBL3	HIGD1B	IL3	SLC28A2	CAPN12	CEP78	NTSR1	ARG1	REM2	AMOTL1	CEP78	RICH2	LOC347376	SLC45A4
9	SULT1A3	CADM3	TECTA	NME2	EHBP1L1	HIST1H1B	PKD1L2	HIST1H1B	MYL10	AKR1B15	ACTN2	EPHX2	REM2	AGBL3
10	EPHX2	LOC283070	DFFA	REM2	VAT1	SLC28A2	AKR1B15	JPH4	CEP78	CTBP2	TMC5	RNASE13	PKD1L2	LZTFL1

Table S7

Sample type	List with TCGA sample IDs
Adjacent Normal samples n=82	TCGA-A7-A0CE-11A TCGA-A7-A0CH-11A TCGA-A7-A0D9-11A TCGA-A7-A0DB-11A TCGA-A7-A13E-11A TCGA-A7-A13F-11A TCGA-A7-A13G-11A TCGA-AC-A23H-11A TCGA-AC-A2FB-11A TCGA-AC-A2FF-11A TCGA-AC-A2FM-11B TCGA-BH-A0AU-11A TCGA-BH-A0AY-11A TCGA-BH-A0AZ-11A TCGA-BH-A0B3-11B TCGA-BH-A0B8-11A TCGA-BH-A0BA-11A TCGA-BH-A0BC-11A TCGA-BH-A0BJ-11A TCGA-BH-A0BM-11A TCGA-BH-A0BS-11A TCGA-BH-A0BT-11A TCGA-BH-A0BV-11A TCGA-BH-A0BZ-11A TCGA-BH-A0C0-11A TCGA-BH-A0C3-11A TCGA-BH-A0DG-11A TCGA-BH-A0DH-11A TCGA-BH-A0DK-11A TCGA-BH-A0DP-11A TCGA-BH-A0DQ-11A TCGA-BH-A0DV-11A TCGA-BH-A0DZ-11A TCGA-BH-A0E0-11A TCGA-BH-A0E1-11A TCGA-BH-A0H7-11A TCGA-BH-A0HA-11A TCGA-BH-A0HK-11A TCGA-BH-A1EN-11A TCGA-BH-A1EO-11A TCGA-BH-A1ET-11B TCGA-BH-A1EU-11A TCGA-BH-A1EV-11A TCGA-BH-A1EW-11B TCGA-BH-A1F0-11B TCGA-BH-A1F2-11A TCGA-BH-A1F6-11B TCGA-BH-A1F8-11B TCGA-BH-A1FB-11A TCGA-BH-A1FC-11A TCGA-BH-A1FD-11B TCGA-BH-A1FE-11B TCGA-BH-A1FG-11B TCGA-BH-A1FH-11B TCGA-BH-A1FJ-11B TCGA-BH-A1FM-11B TCGA-BH-A1FN-11A TCGA-BH-A1FR-11B TCGA-BH-A203-11A TCGA-BH-A204-11A TCGA-BH-A208-11A TCGA-BH-A209-11A TCGA-E2-A15I-11A TCGA-E2-A15K-11A TCGA-E2-A1BC-11A TCGA-E2-A1L7-11A TCGA-E2-A1LB-11A TCGA-E2-A1LS-11A TCGA-E9-A1N4-11A TCGA-E9-A1N5-11A TCGA-E9-A1N6-11A TCGA-E9-A1NA-11A TCGA-E9-A1ND-11A TCGA-E9-A1NF-11A TCGA-E9-A1NG-11A TCGA-E9-A1R7-11A TCGA-E9-A1RB-11A TCGA-E9-A1RC-11A TCGA-E9-A1RD-11A TCGA-E9-A1RF-11A TCGA-E9-A1RH-11A TCGA-E9-A1RI-11A
Tumour samples n=730	TCGA-A1-A0SB-01A TCGA-A1-A0SE-01A TCGA-A1-A0SF-01A TCGA-A1-A0SG-01A TCGA-A1-A0SH-01A TCGA-A1-A0SI-01A TCGA-A1-A0SJ-01A TCGA-A1-A0SK-01A TCGA-A1-A0SM-01A TCGA-A1-A0SN-01A TCGA-A1-A0SO-01A TCGA-A1-A0SP-01A TCGA-A1-A0SQ-01A TCGA-A2-A04R-01A TCGA-A2-A0CK-01A TCGA-A2-A0CO-01A TCGA-A2-A0CR-01A TCGA-A2-A0CT-01A TCGA-A2-A0EN-01A TCGA-A2-A0EP-01A TCGA-A2-A0EU-01A TCGA-A2-A0ST-01A TCGA-A2-A0SU-01A TCGA-A2-A0SV-01A TCGA-A2-A0SW-01A TCGA-A2-A0SX-01A TCGA-A2-A0SY-01A TCGA-A2-A0T0-01A TCGA-A2-A0T1-01A TCGA-A2-A0T2-01A TCGA-A2-A0T4-01A TCGA-A2-A0T5-01A TCGA-A2-A0T6-01A TCGA-A2-A0T7-01A TCGA-A2-A0YC-01A TCGA-A2-A0YD-01A TCGA-A2-A0YE-01A TCGA-A2-A0YF-01A TCGA-A2-A0YG-01A TCGA-A2-A0YH-01A TCGA-A2-A0YI-01A TCGA-A2-A0YJ-01A TCGA-A2-A0YK-01A TCGA-A2-A0YL-01A TCGA-A2-A0YM-01A TCGA-A2-A0YT-01A TCGA-A2-A1FV-01A TCGA-A2-A1FW-01A TCGA-A2-A1FX-01A TCGA-A2-A1FZ-01A TCGA-A2-A1G0-01A TCGA-A2-A1G1-01A TCGA-A2-A1G4-01A TCGA-A2-A1G6-01A TCGA-A2-A259-01A TCGA-A2-A25A-01A TCGA-A2-A25B-01A TCGA-A2-A25C-01A TCGA-A2-A25D-01A TCGA-A2-A25E-01A TCGA-A2-A25F-01A TCGA-A2-A3KC-01A TCGA-A2-A3KD-01A TCGA-A2-A3XS-01A TCGA-A2-A3XT-01A TCGA-A2-A3XU-01A TCGA-A2-A3XV-01A TCGA-A2-A3XW-01A TCGA-A2-A3XX-01A TCGA-A2-A3XY-01A TCGA-A2-A3XZ-01A TCGA-A2-A3Y0-01A TCGA-A2-A4RW-01A TCGA-A2-A4RX-01A TCGA-A2-A4RY-01A TCGA-A2-A4S0-01A TCGA-A2-A4S1-01A TCGA-A2-A4S2-01A TCGA-A2-A4S3-01A TCGA-A7-A0D9-01A TCGA-A7-A13D-01A TCGA-A7-A13E-01A TCGA-A7-A13F-01A TCGA-A7-A13G-01A TCGA-A7-A13H-01A TCGA-A7-A26E-01A TCGA-A7-A26F-01A TCGA-A7-A26G-01A TCGA-A7-A26H-01A TCGA-A7-A26I-01A TCGA-A7-A2KD-01A TCGA-A7-A3IY-01A TCGA-A7-A3IZ-01A TCGA-A7-A3J0-01A TCGA-A7-A3J1-01A TCGA-A7-A3RF-01A TCGA-A7-A425-01A TCGA-A7-A426-01A TCGA-A7-A4SA-01A TCGA-A7-A4SB-01A TCGA-A7-A4SC-01A TCGA-A7-A4SD-01A TCGA-A7-A4SE-01A TCGA-A7-A4SF-01A TCGA-A7-A5ZV-01A TCGA-A7-A5ZW-01A TCGA-A7-A5ZX-01A TCGA-A7-A6VV-01A TCGA-A7-A6VW-01A TCGA-A7-A6VX-01A TCGA-A7-A6VY-01A TCGA-A8-A075-01A TCGA-A8-A08O-01A TCGA-A8-A0A6-01A TCGA-A8-A0AD-01A TCGA-AC-A23C-01A TCGA-AC-A23E-01A TCGA-AC-A23G-01A TCGA-AC-A23H-01A TCGA-AC-A2B8-01A TCGA-AC-A2BK-01A TCGA-AC-A2BM-01A TCGA-AC-A2FB-01A TCGA-AC-A2FE-01A TCGA-AC-A2FF-01A TCGA-AC-A2FG-01A TCGA-AC-A2FK-01A TCGA-AC-A2FM-01A TCGA-AC-A2FO-01A TCGA-AC-A2QH-01A TCGA-AC-A2QI-01A TCGA-AC-A2QJ-01A TCGA-AC-A3BB-01A TCGA-AC-A3EH-01A TCGA-AC-A3HN-01A TCGA-AC-A3OD-01A TCGA-AC-A3QP-01A TCGA-AC-A3QQ-01A TCGA-AC-A3TM-01A TCGA-AC-A3TN-01A TCGA-AC-A3W5-01A TCGA-AC-A3W6-01A TCGA-AC-A3W7-01A TCGA-AC-A3YI-01A TCGA-AC-A3YJ-01A TCGA-AC-A5EH-01A TCGA-AC-A5XS-01A TCGA-AC-A5XU-01A TCGA-AC-A62V-01A TCGA-AC-A62X-01A TCGA-AC-A62Y-01A TCGA-AC-A6IV-01A TCGA-AC-A6IW-01A TCGA-AC-A6IX-01A TCGA-AC-A6NO-01A TCGA-AC-A7VB-01A TCGA-AC-A7VC-01A TCGA-AC-A8OP-01A TCGA-AN-A0XL-01A TCGA-AN-A0XN-01A TCGA-AN-A0XO-01A TCGA-AN-A0XP-01A TCGA-AN-A0XR-01A TCGA-AN-A0XS-01A TCGA-AN-A0XT-01A TCGA-AN-A0XU-01A TCGA-AN-A0XV-01A TCGA-AN-A0XW-01A TCGA-AO-A03L-01A TCGA-AO-A03M-01B TCGA-AO-A03N-01B TCGA-AO-A03U-01B TCGA-AO-A0JA-01A TCGA-AO-A0JB-01A TCGA-AO-A0JC-01A TCGA-AO-A0JD-01A TCGA-AO-A0JE-01A TCGA-AO-A0JF-01A TCGA-AO-A0JG-01A TCGA-AO-A0JI-01A TCGA-AO-A0JJ-01A TCGA-AO-A0JL-01A TCGA-AO-A0JM-01A TCGA-AO-A124-01A TCGA-AO-A125-01A TCGA-AO-A126-01A TCGA-AO-A128-01A TCGA-AO-A129-01A TCGA-AO-A12B-01A TCGA-AO-A12C-01A TCGA-AO-A12E-01A TCGA-AO-A12G-01A TCGA-AO-A1KO-01A TCGA-AO-A1KP-01A TCGA-AO-A1KQ-01A TCGA-AO-A1KR-01A TCGA-AO-A1KS-01A TCGA-AO-A1KT-01A TCGA-AQ-A04H-01B TCGA-AQ-A04L-01B TCGA-AQ-A0Y5-01A TCGA-AQ-A1H2-01A TCGA-AQ-A1H3-01A TCGA-AQ-A54N-01A TCGA-AQ-A54O-01A TCGA-AQ-A7U7-01A TCGA-AR-A0TP-01A TCGA-AR-A0TQ-01A TCGA-AR-A0TR-01A TCGA-AR-A0TT-01A TCGA-AR-A0TU-01A TCGA-AR-A0TV-01A TCGA-AR-A0TW-01A TCGA-AR-A0TX-01A TCGA-AR-A0TZ-01A TCGA-AR-A0U0-01A TCGA-AR-A0U2-01A TCGA-AR-A0U3-01A TCGA-AR-A0U4-01A TCGA-AR-A1AI-01A TCGA-AR-A1AJ-01A TCGA-AR-A1AK-01A TCGA-AR-A1AL-01A TCGA-AR-A1AM-01A TCGA-AR-A1AN-01A TCGA-AR-A1AO-01A TCGA-AR-A1AP-01A TCGA-AR-A1AQ-01A TCGA-AR-A1AR-01A TCGA-AR-

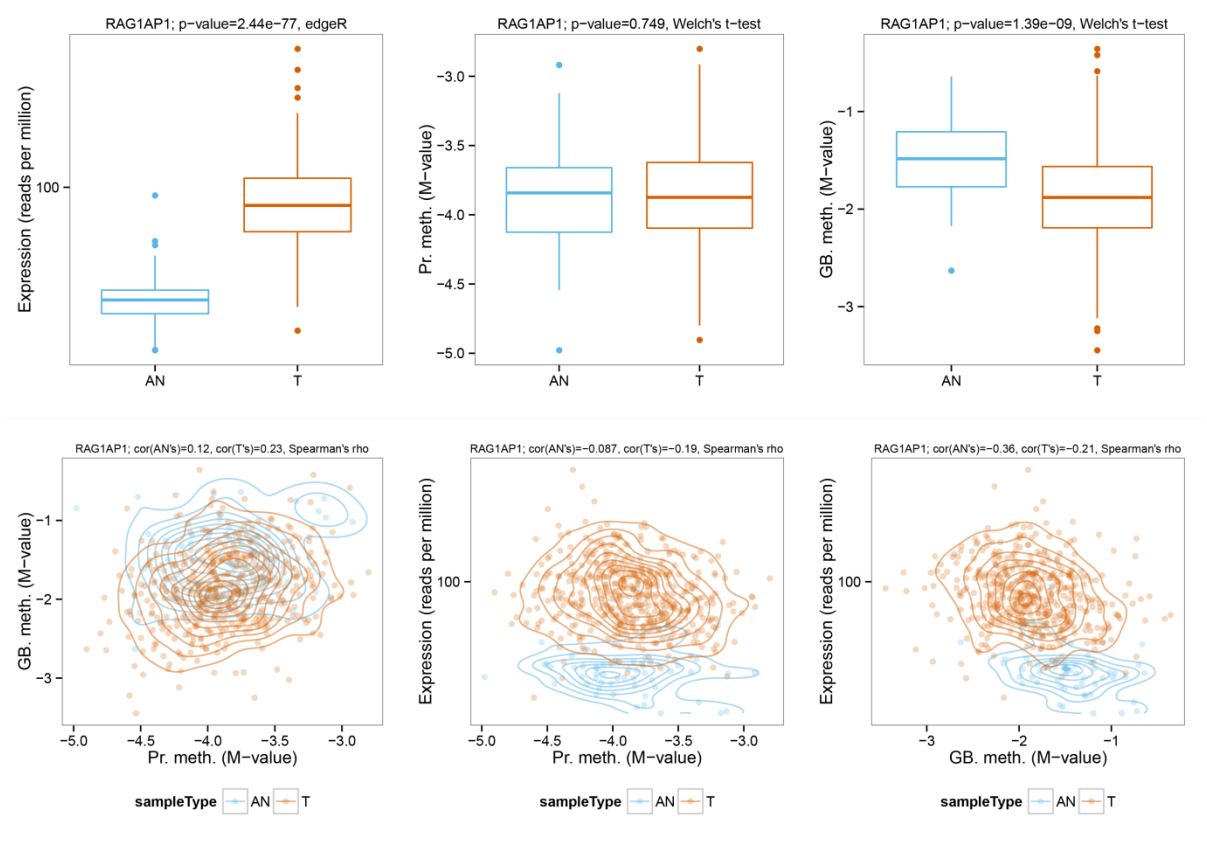
	A1AS-01A TCGA-AR-A1AT-01A TCGA-AR-A1AU-01A TCGA-AR-A1AV-01A TCGA-AR-A1AW-01A TCGA-AR-A1AX-01A TCGA-AR-A1AY-01A TCGA-AR-A24H-01A TCGA-AR-A24K-01A TCGA-AR-A24L-01A TCGA-AR-A24M-01A TCGA-AR-A24N-01A TCGA-AR-A24O-01A TCGA-AR-A24P-01A TCGA-AR-A24Q-01A TCGA-AR-A24R-01A TCGA-AR-A24S-01A TCGA-AR-A24T-01A TCGA-AR-A24U-01A TCGA-AR-A24V-01A TCGA-AR-A24W-01A TCGA-AR-A24X-01A TCGA-AR-A24Z-01A TCGA-AR-A250-01A TCGA-AR-A251-01A TCGA-AR-A252-01A TCGA-AR-A254-01A TCGA-AR-A255-01A TCGA-AR-A256-01A TCGA-AR-A2LE-01A TCGA-AR-A2LH-01A TCGA-AR-A2LJ-01A TCGA-AR-A2LK-01A TCGA-AR-A2LL-01A TCGA-AR-A2LM-01A TCGA-AR-A2LN-01A TCGA-AR-A2LO-01A TCGA-AR-A2LQ-01A TCGA-AR-A2LR-01A TCGA-AR-A5QN-01A TCGA-AR-A5QP-01A TCGA-AR-A5QQ-01A TCGA-B6-A0I1-01A TCGA-B6-A0IK-01A TCGA-B6-A0RE-01A TCGA-B6-A0RG-01A TCGA-B6-A0RI-01A TCGA-B6-A0RL-01A TCGA-B6-A0RM-01A TCGA-B6-A0RN-01A TCGA-B6-A0RO-01A TCGA-B6-A0RP-01A TCGA-B6-A0RS-01A TCGA-B6-A0RT-01A TCGA-B6-A0RU-01A TCGA-B6-A0RV-01A TCGA-B6-A0WT-01A TCGA-B6-A0WV-01A TCGA-B6-A0WW-01A TCGA-B6-A0WX-01A TCGA-B6-A0WY-01A TCGA-B6-A0WZ-01A TCGA-B6-A0X1-01A TCGA-B6-A0X4-01A TCGA-B6-A0X5-01A TCGA-B6-A0X7-01A TCGA-B6-A1KC-01B TCGA-B6-A1KF-01A TCGA-B6-A1KI-01A TCGA-B6-A1KN-01A TCGA-B6-A2IU-01A TCGA-B6-A3ZX-01A TCGA-B6-A400-01A TCGA-B6-A401-01A TCGA-B6-A402-01A TCGA-B6-A408-01A TCGA-B6-A409-01A TCGA-B6-A40B-01A TCGA-B6-A40C-01A TCGA-BH-A0AU-01A TCGA-BH-A0AW-01A TCGA-BH-A0AZ-01A TCGA-BH-A0B2-01A TCGA-BH-A0B3-01A TCGA-BH-A0B5-01A TCGA-BH-A0B6-01A TCGA-BH-A0B8-01A TCGA-BH-A0B9-01A TCGA-BH-A0BA-01A TCGA-BH-A0BC-01A TCGA-BH-A0BF-01A TCGA-BH-A0BJ-01A TCGA-BH-A0BM-01A TCGA-BH-A0BS-01A TCGA-BH-A0BT-01A TCGA-BH-A0BZ-01A TCGA-BH-A0C0-01A TCGA-BH-A0C3-01A TCGA-BH-A0DD-01A TCGA-BH-A0DG-01A TCGA-BH-A0DH-01A TCGA-BH-A0DI-01A TCGA-BH-A0DK-01A TCGA-BH-A0DP-01A TCGA-BH-A0DQ-01A TCGA-BH-A0DS-01A TCGA-BH-A0DV-01A TCGA-BH-A0E0-01A TCGA-BH-A0E1-01A TCGA-BH-A0E2-01A TCGA-BH-A0GY-01A TCGA-BH-A0GZ-01A TCGA-BH-A0H0-01A TCGA-BH-A0H3-01A TCGA-BH-A0H6-01A TCGA-BH-A0H7-01A TCGA-BH-A0H9-01A TCGA-BH-A0HA-01A TCGA-BH-A0HB-01A TCGA-BH-A0HF-01A TCGA-BH-A0HI-01A TCGA-BH-A0HK-01A TCGA-BH-A0HN-01A TCGA-BH-A0HP-01A TCGA-BH-A0HX-01A TCGA-BH-A0HY-01A TCGA-BH-A0RX-01A TCGA-BH-A0W3-01A TCGA-BH-A0W4-01A TCGA-BH-A0W5-01A TCGA-BH-A0WA-01A TCGA-BH-A1EN-01A TCGA-BH-A1EO-01A TCGA-BH-A1ES-01A TCGA-BH-A1ET-01A TCGA-BH-A1EU-01A TCGA-BH-A1EV-01A TCGA-BH-A1EW-01A TCGA-BH-A1EX-01A TCGA-BH-A1EY-01A TCGA-BH-A1F0-01A TCGA-BH-A1F2-01A TCGA-BH-A1F5-01A TCGA-BH-A1F6-01A TCGA-BH-A1F8-01A TCGA-BH-A1FB-01A TCGA-BH-A1FC-01A TCGA-BH-A1FD-01A TCGA-BH-A1FE-01A TCGA-BH-A1FG-01A TCGA-BH-A1FH-01A TCGA-BH-A1FJ-01A TCGA-BH-A1FL-01A TCGA-BH-A1FM-01A TCGA-BH-A1FN-01A TCGA-BH-A1FR-01A TCGA-BH-A1FU-01A TCGA-BH-A201-01A TCGA-BH-A202-01A TCGA-BH-A203-01A TCGA-BH-A204-01A TCGA-BH-A208-01A TCGA-BH-A209-01A TCGA-BH-A28O-01A TCGA-BH-A28Q-01A TCGA-BH-A2L8-01A TCGA-BH-A42T-01A TCGA-BH-A42U-01A TCGA-BH-A42V-01A TCGA-BH-A6R8-01A TCGA-BH-A6R9-01A TCGA-BH-A8FY-01A TCGA-BH-A8FZ-01A TCGA-BH-A8G0-01A TCGA-C8-A1HE-01A TCGA-C8-A1HF-01A TCGA-C8-A1HG-01A TCGA-C8-A1HI-01A TCGA-C8-A1HJ-01A TCGA-C8-A1HK-01A TCGA-C8-A1HL-01A TCGA-C8-A1HM-01A TCGA-C8-A1HN-01A TCGA-C8-A1HO-01A TCGA-C8-A26V-01A TCGA-C8-A26W-01A TCGA-C8-A26X-01A TCGA-C8-A26Y-01A TCGA-C8-A26Z-01A TCGA-C8-A273-01A TCGA-C8-A274-01A TCGA-C8-A275-01A TCGA-C8-A278-01A TCGA-C8-A27A-01A TCGA-C8-A27B-01A TCGA-C8-A3M7-01A TCGA-C8-A3M8-01A TCGA-C8-A8HP-01A TCGA-C8-A8HQ-01A TCGA-C8-A8HR-01A TCGA-D8-A1J8-01A TCGA-D8-A1J9-01A TCGA-D8-A1JA-01A TCGA-D8-A1JB-01A TCGA-D8-A1JC-01A TCGA-D8-A1JD-01A TCGA-D8-A1JE-01A TCGA-D8-A1JF-01A TCGA-D8-A1JG-01B TCGA-D8-A1JH-01A TCGA-D8-A1JI-01A TCGA-D8-A1JJ-01A TCGA-D8-A1JK-01A TCGA-D8-A1JL-01A TCGA-D8-A1JM-01A TCGA-D8-A1JN-01A TCGA-D8-A1JP-01A TCGA-D8-A1JS-01A TCGA-D8-A1JT-01A TCGA-D8-A1JU-01A TCGA-D8-A1X5-01A TCGA-D8-A1X6-01A TCGA-D8-A1X7-01A TCGA-D8-A1X8-01A TCGA-D8-A1X9-01A TCGA-D8-A1XA-01A TCGA-D8-A1XB-01A TCGA-D8-A1XC-01A TCGA-D8-A1XD-01A TCGA-D8-A1XF-01A TCGA-D8-A1XG-01A TCGA-D8-A1XJ-01A TCGA-D8-A1XK-01A TCGA-D8-A1XL-01A TCGA-D8-A1XM-01A TCGA-D8-A1XO-01A TCGA-D8-A1XQ-01A TCGA-D8-A1XR-01A TCGA-D8-A1XS-01A TCGA-D8-A1XT-01A TCGA-D8-A1XU-01A TCGA-D8-A1XV-01A TCGA-D8-A1XW-01A TCGA-D8-A1XY-01A TCGA-D8-A1XZ-01A TCGA-D8-A1Y0-01A TCGA-D8-A1Y1-01A TCGA-D8-A1Y2-01A TCGA-D8-A1Y3-01A TCGA-D8-A27E-01A TCGA-D8-A27F-01A TCGA-D8-A27G-01A TCGA-D8-A27H-01A TCGA-D8-A27I-01A TCGA-D8-A27K-01A TCGA-D8-A27L-01A TCGA-D8-A27M-01A TCGA-D8-A27N-01A TCGA-D8-A27P-01A TCGA-D8-A27R-01A TCGA-D8-A27T-01A TCGA-D8-A27V-01A TCGA-D8-A27W-01A TCGA-D8-A3Z5-01A TCGA-D8-A3Z6-01A TCGA-D8-A4Z1-01A TCGA-D8-A73U-01A TCGA-D8-A73W-01A TCGA-D8-A73X-01A TCGA-E2-A105-01A TCGA-E2-A106-01A TCGA-E2-A107-01A TCGA-E2-A108-01A TCGA-E2-A109-01A TCGA-E2-A10B-01A TCGA-E2-A10C-01A TCGA-E2-A10E-01A TCGA-E2-A10F-01A TCGA-E2-A14N-01A TCGA-E2-A14U-01A TCGA-E2-A15I-01A TCGA-E2-A15J-01A TCGA-E2-A15K-01A TCGA-E2-A1AZ-01A TCGA-E2-A1B0-01A TCGA-E2-A1B1-01A TCGA-E2-A1B4-01A TCGA-E2-A1B5-01A TCGA-E2-A1B6-01A TCGA-E2-A1BC-01A TCGA-E2-A1BD-01A TCGA-E2-A1IE-01A TCGA-E2-A1IF-01A TCGA-E2-A1IG-01A TCGA-E2-A1IH-01A TCGA-E2-A1II-01A TCGA-E2-A1IJ-01A TCGA-E2-A1IK-01A TCGA-E2-A1IL-01A TCGA-E2-A1IN-01A TCGA-E2-A1IO-01A TCGA-E2-A1IU-01A TCGA-E2-A1L6-01A TCGA-E2-A1L7-01A TCGA-E2-A1L8-01A TCGA-E2-A1L9-01A TCGA-E2-A1LA-01A TCGA-E2-A1LB-01A TCGA-E2-A1LE-01A TCGA-E2-A1LG-01A TCGA-E2-A1LH-01A TCGA-E2-A1LI-01A TCGA-E2-A1LK-01A TCGA-E2-A1LL-01A TCGA-E2-A1LS-01A TCGA-E2-A2P5-01A TCGA-E2-A2P6-01A TCGA-E2-A3DX-01A TCGA-E2-A56Z-01A TCGA-E2-A570-01A TCGA-E2-A572-01A TCGA-E2-A573-01A TCGA-E2-A574-01A TCGA-E2-A576-01A TCGA-E9-A1N3-01A TCGA-E9-A1N4-01A TCGA-E9-A1N5-01A TCGA-E9-A1N6-01A TCGA-E9-A1N8-01A TCGA-E9-A1N9-01A TCGA-E9-A1NA-01A TCGA-E9-A1NC-01A TCGA-E9-A1ND-01A TCGA-E9-A1NE-01A TCGA-E9-A1NF-01A TCGA-E9-A1NG-01A TCGA-E9-A1NH-01A TCGA-E9-A1NI-01A TCGA-E9-A1QZ-01A TCGA-E9-A1R0-01A TCGA-E9-A1R2-01A TCGA-E9-A1R3-01A TCGA-E9-

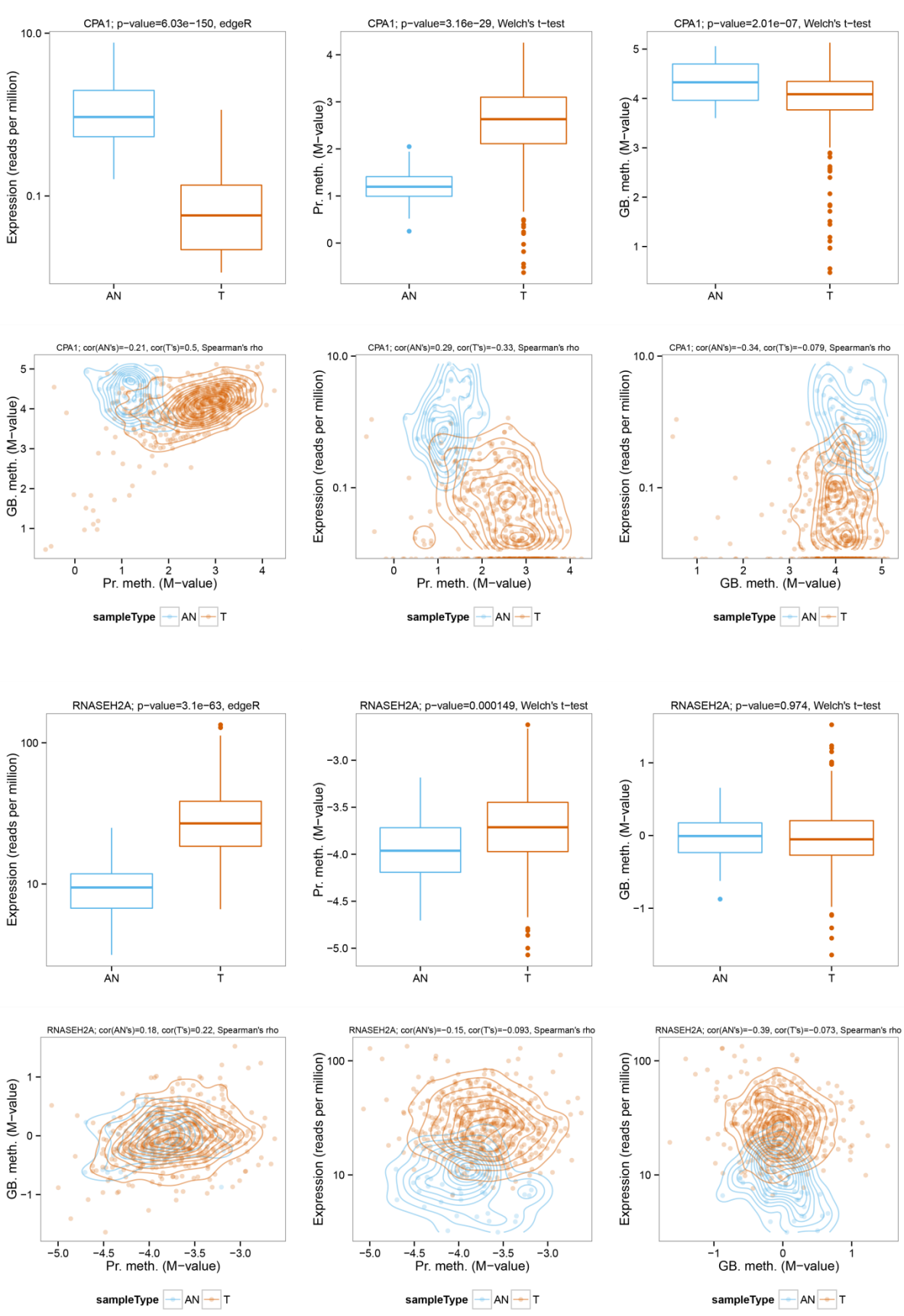
	A1R4-01A TCGA-E9-A1R5-01A TCGA-E9-A1R6-01A TCGA-E9-A1R7-01A TCGA-E9-A1RA-01A TCGA-E9-A1RB-01A TCGA-E9-A1RC-01A TCGA-E9-A1RD-01A TCGA-E9-A1RE-01A TCGA-E9-A1RF-01A TCGA-E9-A1RG-01A TCGA-E9-A1RH-01A TCGA-E9-A1RI-01A TCGA-E9-A226-01A TCGA-E9-A227-01A TCGA-E9-A228-01A TCGA-E9-A229-01A TCGA-E9-A22A-01A TCGA-E9-A22B-01A TCGA-E9-A22D-01A TCGA-E9-A22E-01A TCGA-E9-A22G-01A TCGA-E9-A22H-01A TCGA-E9-A243-01A TCGA-E9-A244-01A TCGA-E9-A245-01A TCGA-E9-A247-01A TCGA-E9-A248-01A TCGA-E9-A249-01A TCGA-E9-A24A-01A TCGA-E9-A295-01A TCGA-E9-A2JS-01A TCGA-E9-A2JT-01A TCGA-E9-A3HO-01A TCGA-E9-A3Q9-01A TCGA-E9-A3QA-01A TCGA-E9-A3X8-01A TCGA-E9-A54X-01A TCGA-E9-A5UO-01A TCGA-E9-A5UP-01A TCGA-E9-A6HE-01A TCGA-EW-A1IW-01A TCGA-EW-A1IX-01A TCGA-EW-A1IY-01A TCGA-EW-A1IZ-01A TCGA-EW-A1J1-01A TCGA-EW-A1J2-01A TCGA-EW-A1J3-01A TCGA-EW-A1J5-01A TCGA-EW-A1J6-01A TCGA-EW-A1OV-01A TCGA-EW-A1OW-01A TCGA-EW-A1OX-01A TCGA-EW-A1OY-01A TCGA-EW-A1OZ-01A TCGA-EW-A1P0-01A TCGA-EW-A1P1-01A TCGA-EW-A1P3-01A TCGA-EW-A1P4-01A TCGA-EW-A1P5-01A TCGA-EW-A1P6-01A TCGA-EW-A1P7-01A TCGA-EW-A1P8-01A TCGA-EW-A1PA-01A TCGA-EW-A1PB-01A TCGA-EW-A1PC-01B TCGA-EW-A1PD-01A TCGA-EW-A1PE-01A TCGA-EW-A1PF-01A TCGA-EW-A1PG-01A TCGA-EW-A1PH-01A TCGA-EW-A2FR-01A TCGA-EW-A2FS-01A TCGA-EW-A2FV-01A TCGA-EW-A2FW-01A TCGA-EW-A3E8-01B TCGA-EW-A3U0-01A TCGA-EW-A423-01A TCGA-EW-A424-01A TCGA-EW-A6S9-01A TCGA-EW-A6SA-01A TCGA-EW-A6SB-01A TCGA-EW-A6SC-01A TCGA-EW-A6SD-01A TCGA-GI-A2C8-01A TCGA-GI-A2C9-01A TCGA-GM-A2D9-01A TCGA-GM-A2DA-01A TCGA-GM-A2DB-01A TCGA-GM-A2DC-01A TCGA-GM-A2DD-01A TCGA-GM-A2DF-01A TCGA-GM-A2DH-01A TCGA-GM-A2DI-01A TCGA-GM-A2DK-01A TCGA-GM-A2DL-01A TCGA-GM-A2DM-01A TCGA-GM-A2DN-01A TCGA-GM-A2DO-01A TCGA-GM-A3NW-01A TCGA-GM-A3NY-01A TCGA-GM-A3XG-01A TCGA-GM-A3XL-01A TCGA-GM-A3XN-01A TCGA-GM-A4E0-01A TCGA-GM-A5PV-01A TCGA-GM-A5PX-01A TCGA-HN-A2NL-01A TCGA-JL-A3YW-01A TCGA-JL-A3YX-01A TCGA-LD-A66U-01A TCGA-LD-A74U-01A TCGA-LD-A7W5-01A TCGA-LD-A7W6-01A TCGA-LL-A440-01A TCGA-LL-A441-01A TCGA-LL-A442-01A TCGA-LL-A50Y-01A TCGA-LL-A5YL-01A TCGA-LL-A5YM-01A TCGA-LL-A5YN-01A TCGA-LL-A5YO-01A TCGA-LL-A5YP-01A TCGA-LL-A6FP-01A TCGA-LL-A6FQ-01A TCGA-LL-A6FR-01A TCGA-LL-A73Y-01A TCGA-LL-A73Z-01A TCGA-LL-A740-01A TCGA-LL-A7SZ-01A TCGA-LL-A7T0-01A TCGA-LL-A8F5-01A TCGA-LQ-A4E4-01A TCGA-MS-A51U-01A TCGA-OL-A5RU-01A TCGA-OL-A5RV-01A TCGA-OL-A5RW-01A TCGA-OL-A5RX-01A TCGA-OL-A5RY-01A TCGA-OL-A5RZ-01A TCGA-OL-A5S0-01A TCGA-OL-A66H-01A TCGA-OL-A66I-01A TCGA-OL-A66J-01A TCGA-OL-A66K-01A TCGA-OL-A66L-01A TCGA-OL-A66N-01A TCGA-OL-A66O-01A TCGA-OL-A66P-01A TCGA-OL-A6VO-01A TCGA-OL-A6VR-01A TCGA-PL-A8LZ-01A TCGA-S3-A6ZF-01A TCGA-S3-A6ZG-01A TCGA-S3-A6ZH-01A TCGA-V7-A7HQ-01A TCGA-W8-A86G-01A TCGA-XX-A899-01A TCGA-XX-A89A-01A

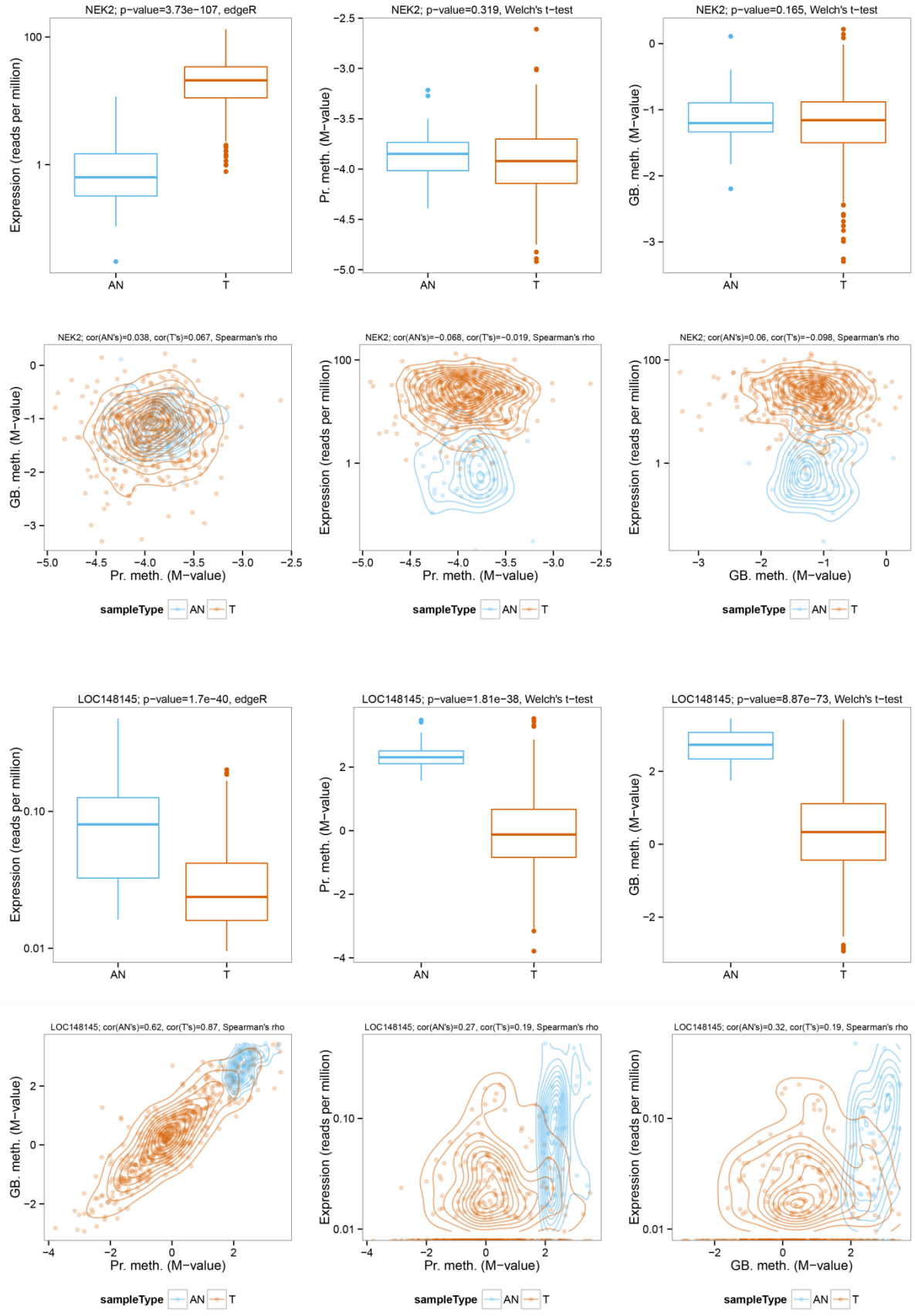
Table S8

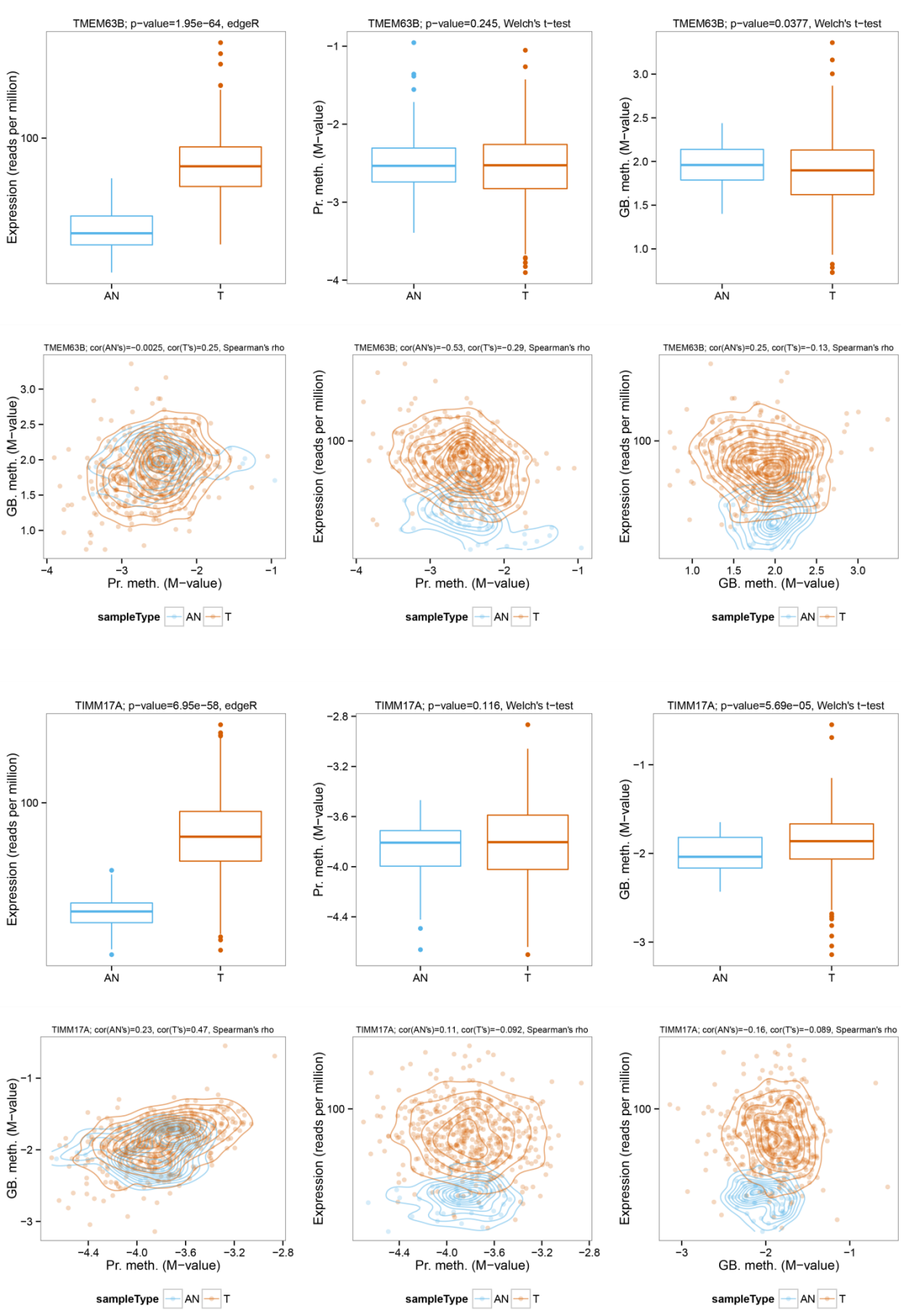
Sample type	**List with TCGA patient IDs**
Progressed disease n=14	TCGA-A7-A3RF TCGA-A7-A425 TCGA-LL-A5YM TCGA-E9-A243 TCGA-A7-A13G TCGA-A7-A26H TCGA-LQ-A4E4 TCGA-A7-A13H TCGA-A8-A08O TCGA-E9-A226 TCGA-A2-A3XY TCGA-E9-A2JS TCGA-A2-A3XU TCGA-AR-A5QQ
Non-progressed disease n=57	TCGA-A7-A0CE TCGA-A7-A0CH TCGA-E9-A1RI TCGA-E9-A1NE TCGA-OL-A5RW TCGA-E9-A1NA TCGA-E9-A1N5 TCGA-A7-A0D9 TCGA-AR-A1AS TCGA-AR-A2LN TCGA-GM-A3NY TCGA-A2-A3Y0 TCGA-E9-A22A TCGA-AR-A2LO TCGA-E9-A1NC TCGA-A2-A3KD TCGA-AR-A2LQ TCGA-AC-A2FB TCGA-GM-A3XG TCGA-A2-A0YL TCGA-A2-A3XW TCGA-BH-A0HY TCGA-EW-A2FS TCGA-EW-A1P3 TCGA-BH-A0HA TCGA-EW-A2FR TCGA-AR-A255 TCGA-AR-A1AV TCGA-AR-A2LJ TCGA-AR-A1AX TCGA-AR-A1AM TCGA-AR-A2LJ TCGA-AR-A1AW TCGA-OL-A66J TCGA-GM-A3XN TCGA-GM-A3XL TCGA-GM-A4E0 TCGA-AR-A254 TCGA-AR-A252 TCGA-AR-A24T TCGA-AR-A1AU TCGA-AR-A251 TCGA-AR-A24N TCGA-AR-A24Z TCGA-A2-A3XT TCGA-AR-A24X TCGA-AR-A24V TCGA-B6-A401 TCGA-AR-A0U4 TCGA-AR-A0TT TCGA-AR-A24R TCGA-AR-A24M TCGA-AR-A0TW TCGA-AR-A24Q TCGA-B6-A40B TCGA-A2-A0EP TCGA-A2-A0CR TCGA-GM-A3NW TCGA-AR-A0TP TCGA-AR-A0U3 TCGA-AQ-A04L

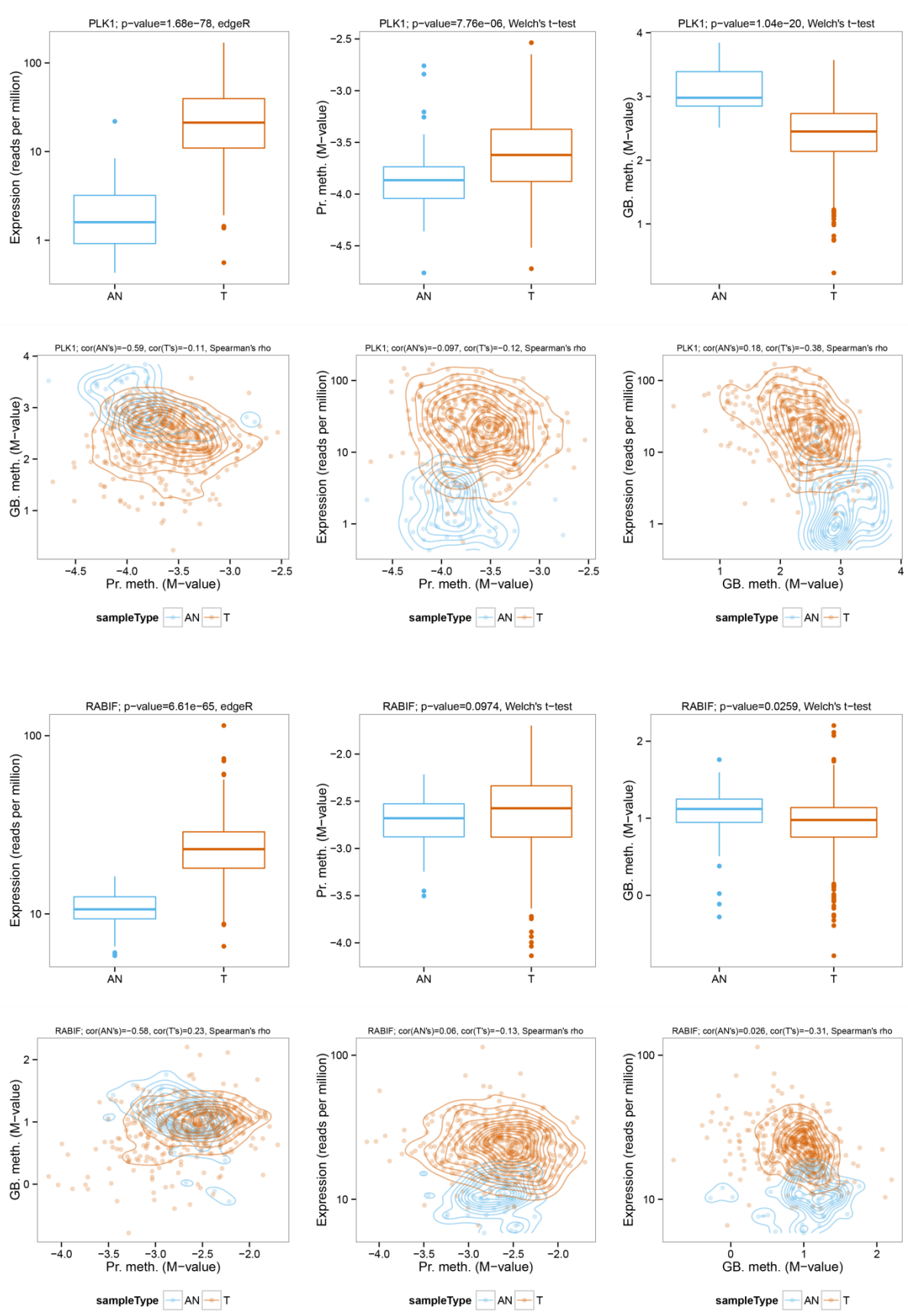
Figure S8 Marginal and pairwise distributions of gene expression, promoter methylation, and gene body methylation for the top-10 genes identified by integrative PINCAGE in the comparison between tumour and adjacent normal samples. For each gene **Top rows:** Marginal distributions of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value across BRCA Tumour (T) and Adjacent Normal (AN) samples. For each gene **Bottom rows:** 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).




CPA1; p−value=6.03e−150, edgeR
Expression (reads per million)
10.0
0.1
AN
T
CPA1; p−value=3.16e−29, Welch's t−test
Pr. meth. (M−value)
CPA1; p−value=2.01e−07, Welch's t−test
GB. meth. (M−value)
CPA1; cor(AN's)=−0.21, cor(T's)=0.5, Spearman's rho
CPA1; cor(AN's)=0.29, cor(T's)=−0.33, Spearman's rho
CPA1; cor(AN's)=−0.34, cor(T's)=−0.079, Spearman's rho
sampleType AN T
RNASEH2A; p−value=3.1e−63, edgeR
RNASEH2A; p−value=0.000149, Welch's t−test
RNASEH2A; p−value=0.974, Welch's t−test
RNASEH2A; cor(AN's)=0.18, cor(T's)=0.22, Spearman's rho
RNASEH2A; cor(AN's)=−0.15, cor(T's)=−0.093, Spearman's rho
RNASEH2A; cor(AN's)=−0.39, cor(T's)=−0.073, Spearman's rho


NEK2; p−value=3.73e−107, edgeR
Expression (reads per million)
AN
T
NEK2; p−value=0.319, Welch's t−test
Pr. meth. (M−value)
NEK2; p−value=0.165, Welch's t−test
GB. meth. (M−value)
NEK2; cor(AN's)=0.038, cor(T's)=0.067, Spearman's rho
NEK2; cor(AN's)=−0.068, cor(T's)=−0.019, Spearman's rho
NEK2; cor(AN's)=0.06, cor(T's)=−0.098, Spearman's rho
sampleType AN T
LOC148145; p−value=1.7e−40, edgeR
LOC148145; p−value=1.81e−38, Welch's t−test
LOC148145; p−value=8.87e−73, Welch's t−test
LOC148145; cor(AN's)=0.62, cor(T's)=0.87, Spearman's rho
LOC148145; cor(AN's)=0.27, cor(T's)=0.19, Spearman's rho
LOC148145; cor(AN's)=0.32, cor(T's)=0.19, Spearman's rho


TMEM63B; p−value=1.95e−64, edgeR
Expression (reads per million)
AN
T
TMEM63B; p−value=0.245, Welch's t−test
Pr. meth. (M−value)
TMEM63B; p−value=0.0377, Welch's t−test
GB. meth. (M−value)
TMEM63B; cor(AN's)=−0.0025, cor(T's)=0.25, Spearman's rho
TMEM63B; cor(AN's)=−0.53, cor(T's)=−0.29, Spearman's rho
TMEM63B; cor(AN's)=0.25, cor(T's)=−0.13, Spearman's rho
sampleType AN T
TIMM17A; p−value=6.95e−58, edgeR
TIMM17A; p−value=0.116, Welch's t−test
TIMM17A; p−value=5.69e−05, Welch's t−test
TIMM17A; cor(AN's)=0.23, cor(T's)=0.47, Spearman's rho
TIMM17A; cor(AN's)=0.11, cor(T's)=−0.092, Spearman's rho
TIMM17A; cor(AN's)=−0.16, cor(T's)=−0.089, Spearman's rho


PLK1; p−value=1.68e−78, edgeR
Expression (reads per million)
AN
T
PLK1; p−value=7.76e−06, Welch's t−test
Pr. meth. (M−value)
PLK1; p−value=1.04e−20, Welch's t−test
GB. meth. (M−value)
PLK1; cor(AN's)=−0.59, cor(T's)=−0.11, Spearman's rho
PLK1; cor(AN's)=−0.097, cor(T's)=−0.12, Spearman's rho
PLK1; cor(AN's)=0.18, cor(T's)=−0.38, Spearman's rho
sampleType AN T
RABIF; p−value=6.61e−65, edgeR
RABIF; p−value=0.0974, Welch's t−test
RABIF; p−value=0.0259, Welch's t−test
RABIF; cor(AN's)=−0.58, cor(T's)=0.23, Spearman's rho
RABIF; cor(AN's)=0.06, cor(T's)=−0.13, Spearman's rho
RABIF; cor(AN's)=0.026, cor(T's)=−0.31, Spearman's rho

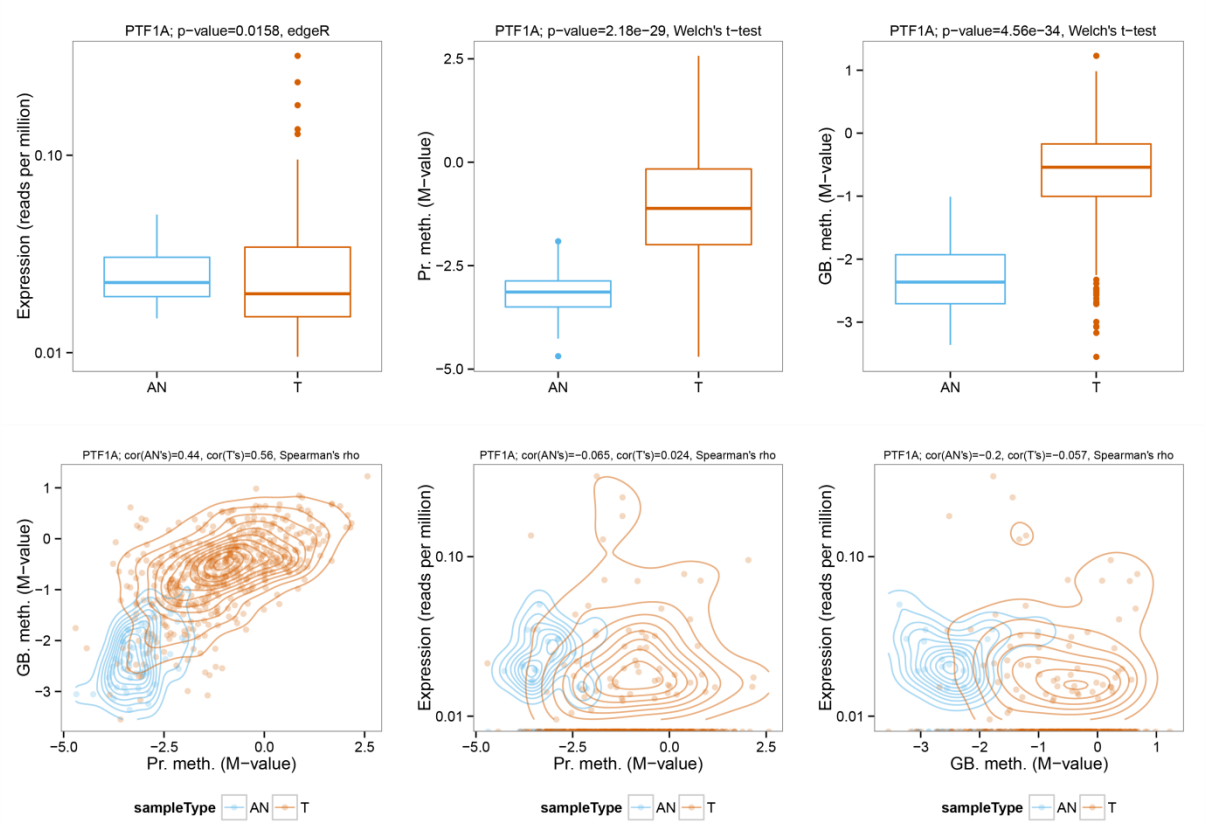
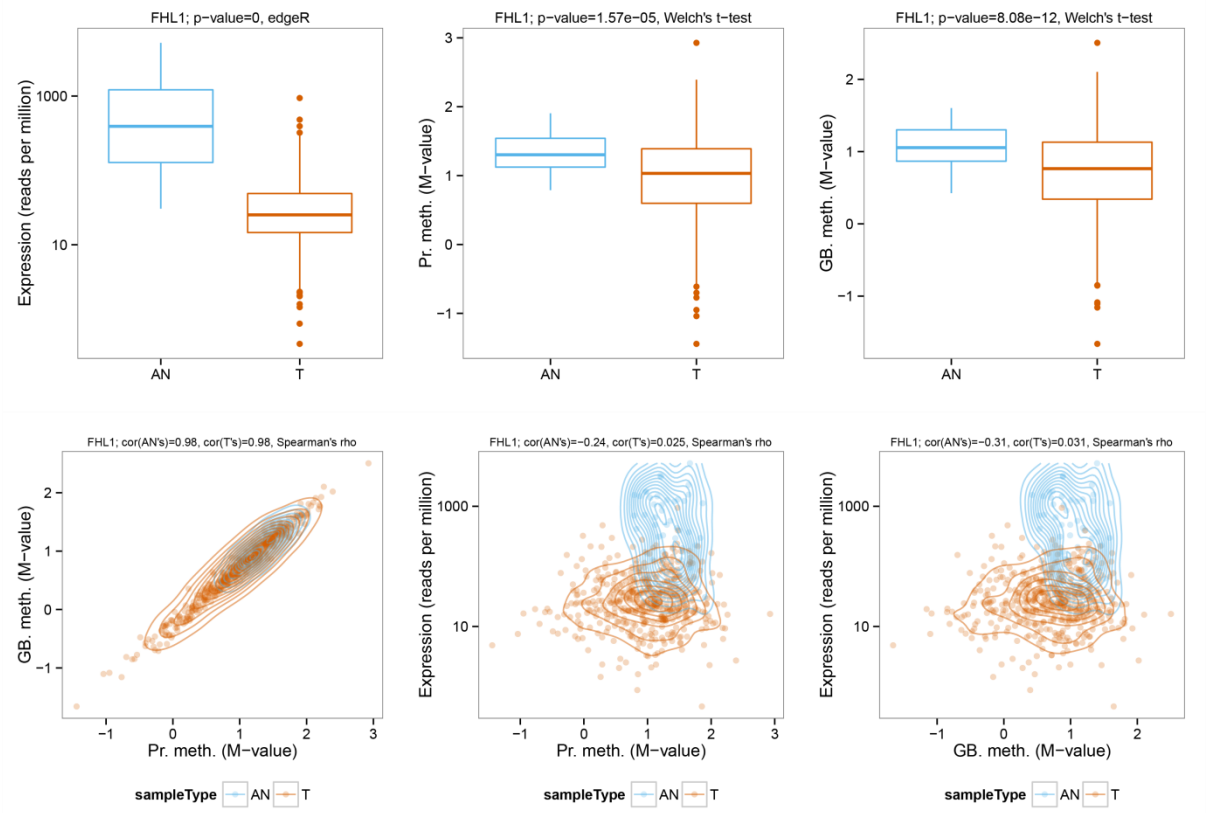
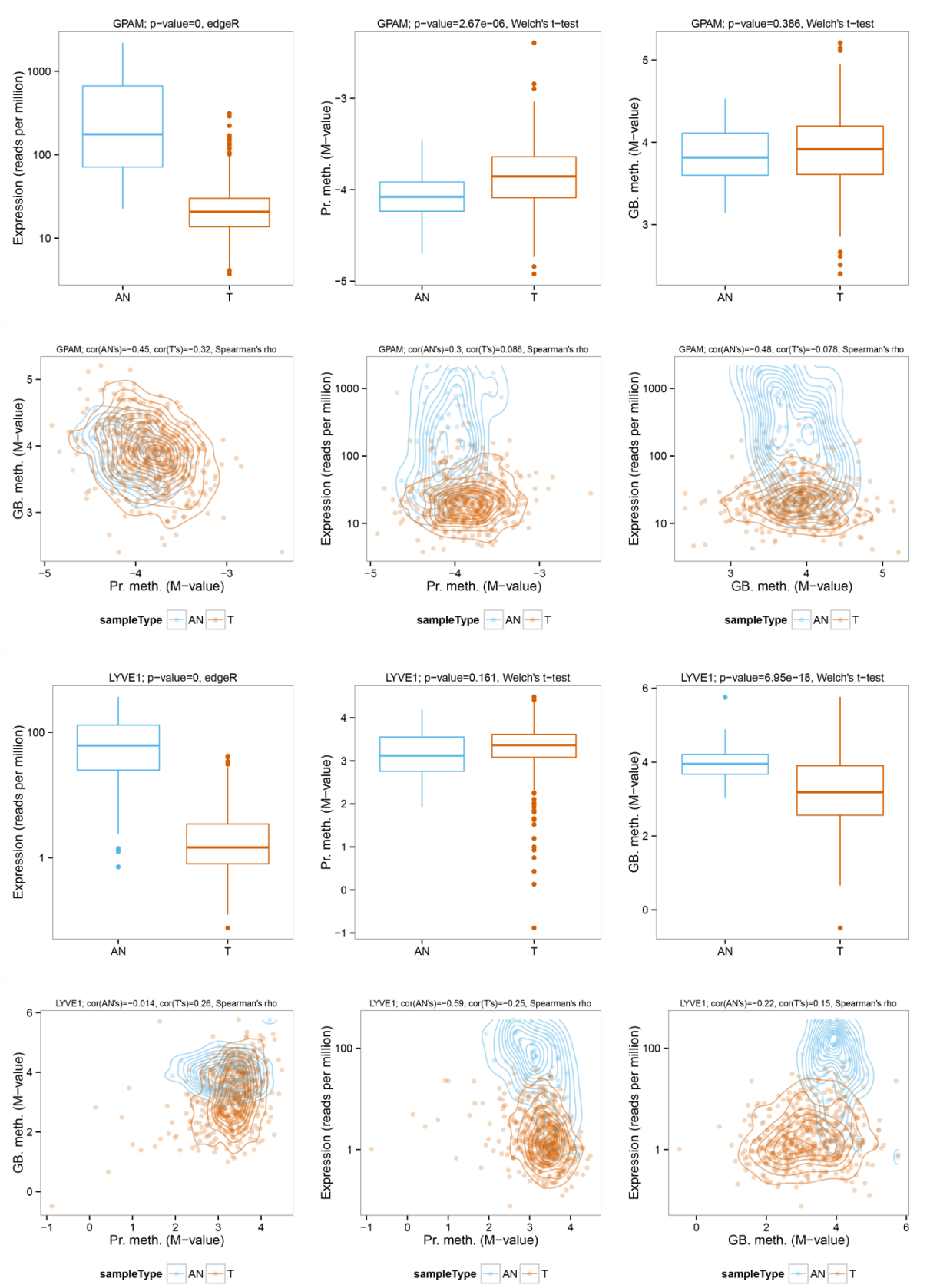
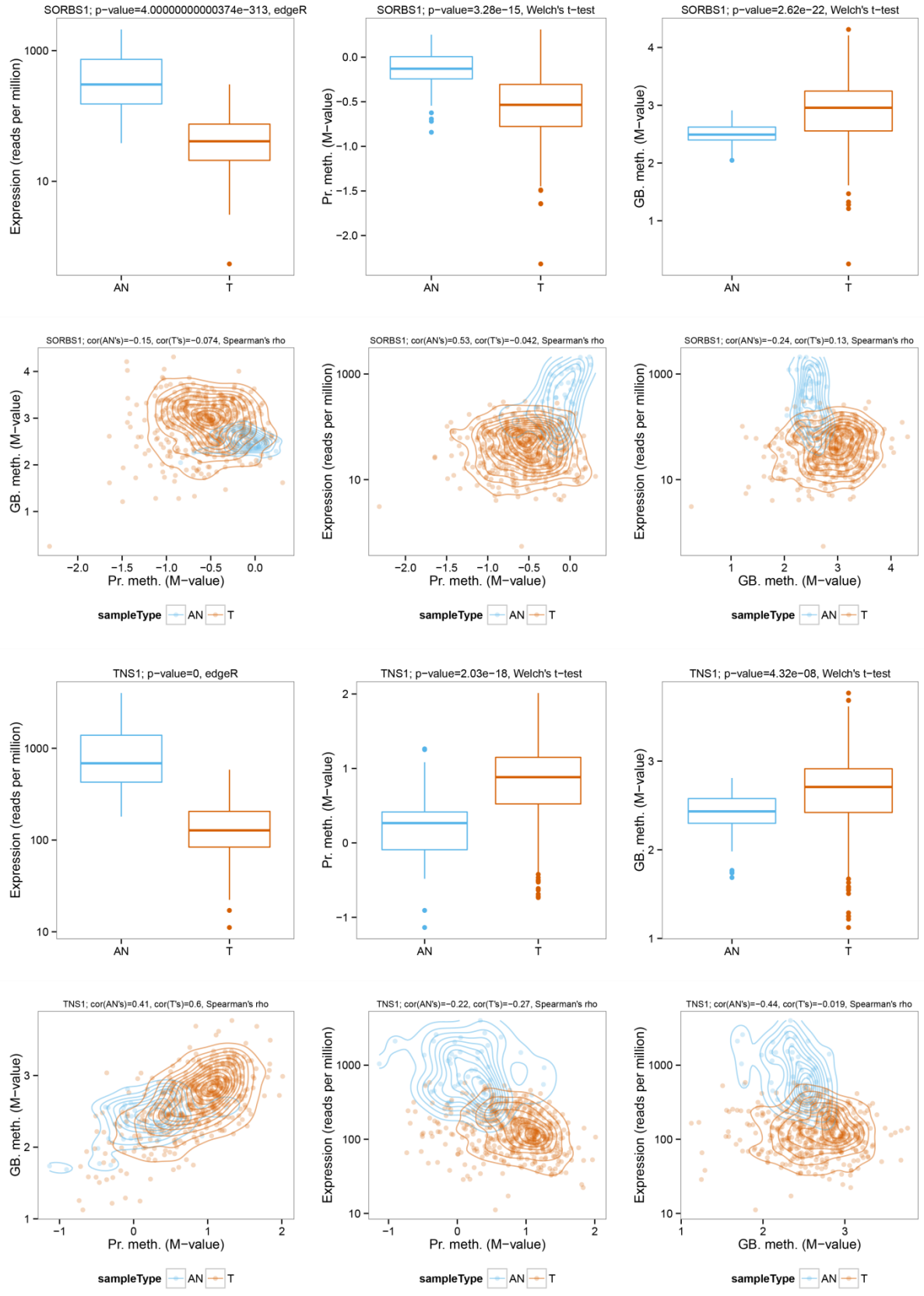
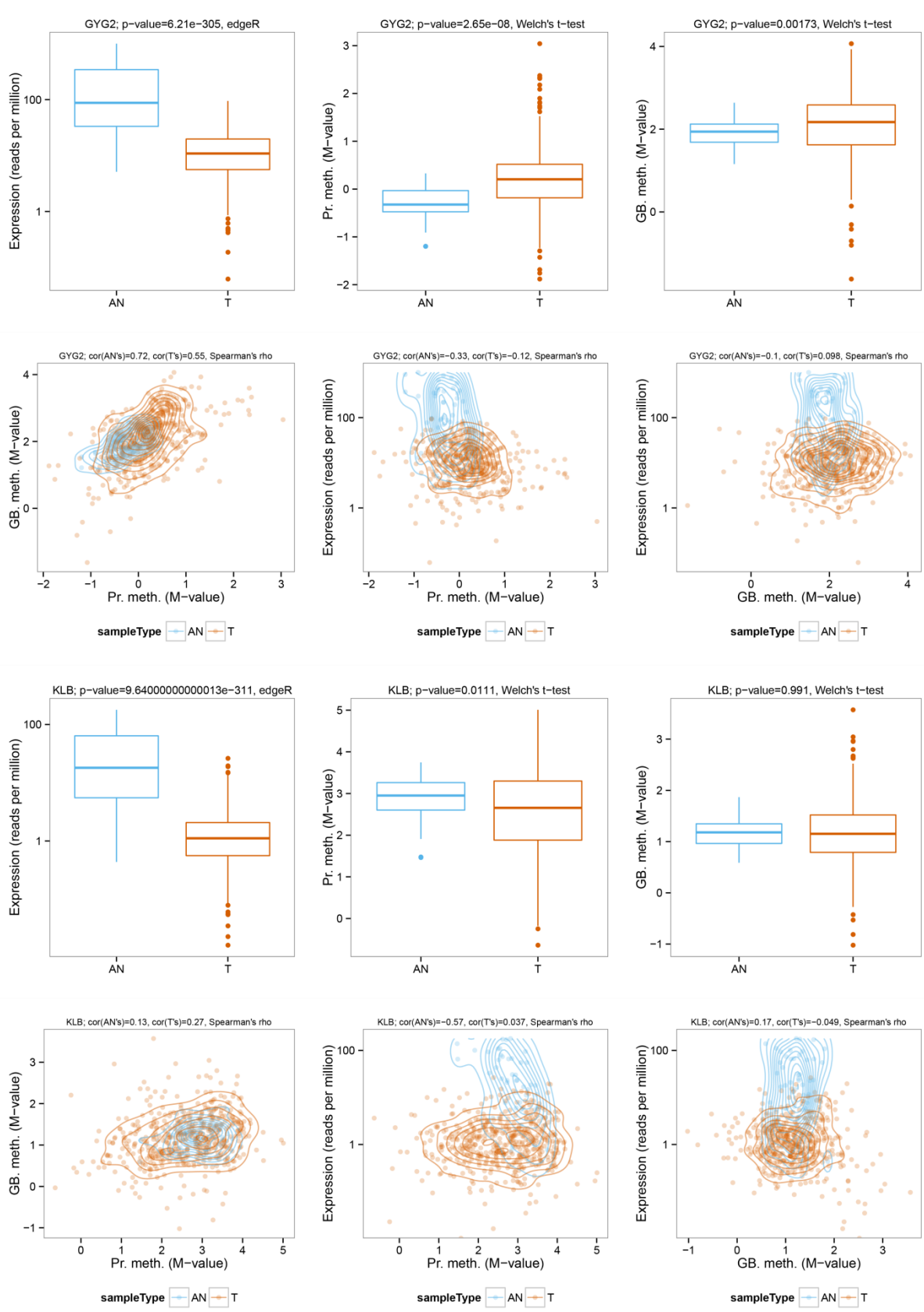

PTF1A; p−value=0.0158, edgeR
Expression (reads per million)
0.10
0.01
AN
T
PTF1A; p−value=2.18e−29, Welch's t−test
Pr. meth. (M−value)
2.5
0.0
−2.5
−5.0
AN
T
PTF1A; p−value=4.56e−34, Welch's t−test
GB. meth. (M−value)
1
0
−1
−2
−3
AN
T
PTF1A; cor(AN's)=0.44, cor(T's)=0.56, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
−5.0
−2.5
0.0
2.5
sampleType AN T
PTF1A; cor(AN's)=−0.065, cor(T's)=0.024, Spearman's rho
Expression (reads per million)
0.10
0.01
Pr. meth. (M−value)
sampleType AN T
PTF1A; cor(AN's)=−0.2, cor(T's)=−0.057, Spearman's rho
Expression (reads per million)
0.10
0.01
−3
−2
−1
0
1
GB. meth. (M−value)
sampleType AN T

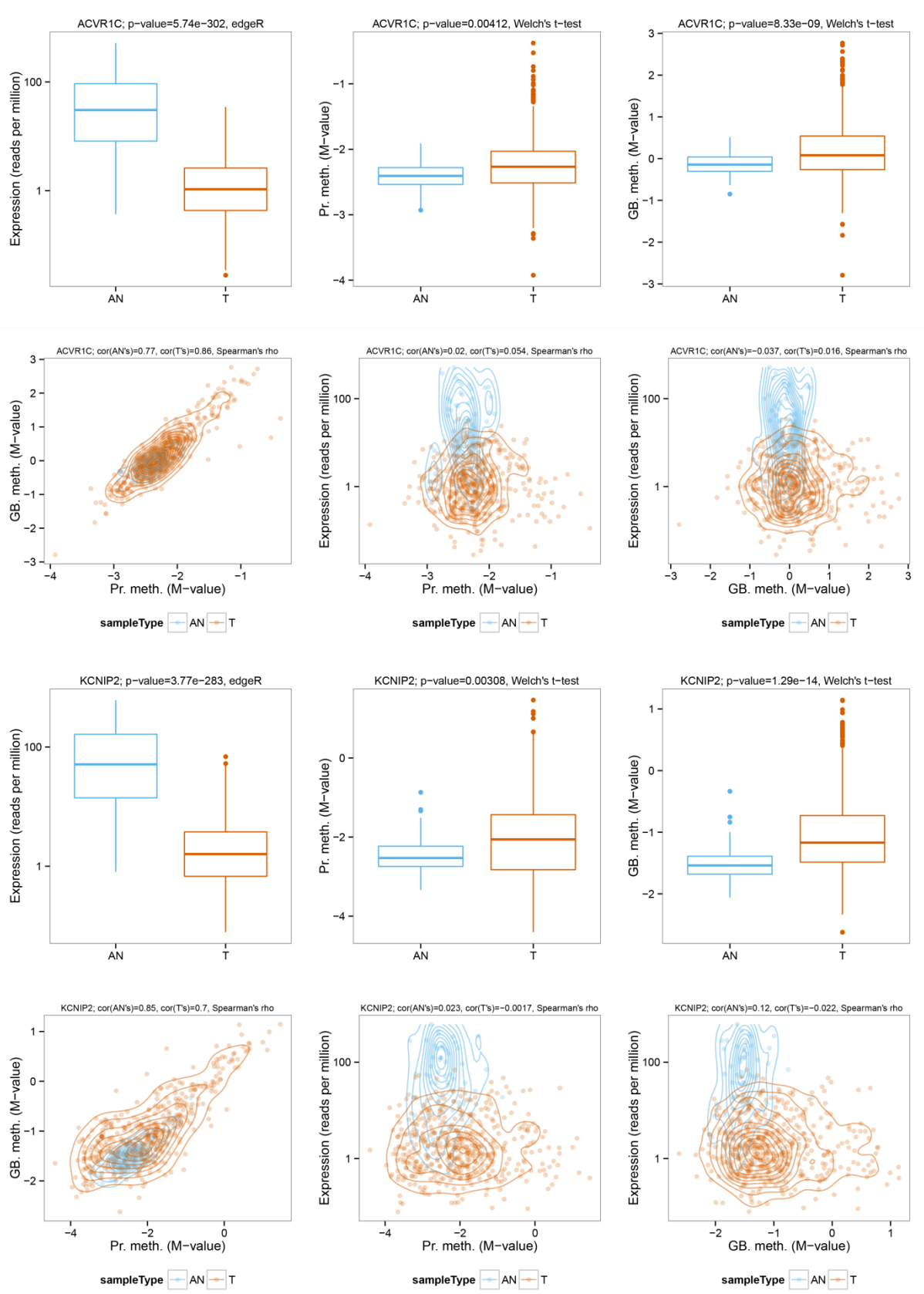
Figure S9 Marginal and pairwise distributions of gene expression, promoter methylation, and gene body methylation for the top-10 genes identified by combination of established methods with Fisher's in the comparison between tumour and adjacent normal samples. For each gene **Top rows:** Marginal distributions of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value across BRCA Tumour (T) and Adjacent Normal (AN) samples. For each gene **Bottom rows:** 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).




GPAM; p−value=0, edgeR
Expression (reads per million)
1000
100
10
AN
T
GPAM; p−value=2.67e−06, Welch's t−test
Pr. meth. (M−value)
−3
−4
−5
GPAM; p−value=0.386, Welch's t−test
GB. meth. (M−value)
5
4
3
GPAM; cor(AN's)=−0.45, cor(T's)=−0.32, Spearman's rho
Pr. meth. (M−value)
GPAM; cor(AN's)=0.3, cor(T's)=0.086, Spearman's rho
GPAM; cor(AN's)=−0.48, cor(T's)=−0.078, Spearman's rho
GB. meth. (M−value)
sampleType AN T
LYVE1; p−value=0, edgeR
LYVE1; p−value=0.161, Welch's t−test
LYVE1; p−value=6.95e−18, Welch's t−test
LYVE1; cor(AN's)=−0.014, cor(T's)=0.26, Spearman's rho
LYVE1; cor(AN's)=−0.59, cor(T's)=−0.25, Spearman's rho
LYVE1; cor(AN's)=−0.22, cor(T's)=0.15, Spearman's rho


SORBS1; p−value=4.00000000000374e−313, edgeR
Expression (reads per million)
1000
10
AN
T
SORBS1; p−value=3.28e−15, Welch's t−test
Pr. meth. (M−value)
0.0
−0.5
−1.0
−1.5
−2.0
AN
T
SORBS1; p−value=2.62e−22, Welch's t−test
GB. meth. (M−value)
4
3
2
1
AN
T
SORBS1; cor(AN's)=−0.15, cor(T's)=−0.074, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
−2.0
−1.5
−1.0
−0.5
0.0
sampleType
AN
T
SORBS1; cor(AN's)=0.53, cor(T's)=−0.042, Spearman's rho
Expression (reads per million)
1000
10
Pr. meth. (M−value)
sampleType
AN
T
SORBS1; cor(AN's)=−0.24, cor(T's)=0.13, Spearman's rho
Expression (reads per million)
1000
10
1
2
3
4
GB. meth. (M−value)
sampleType
AN
T
TNS1; p−value=0, edgeR
Expression (reads per million)
1000
100
10
AN
T
TNS1; p−value=2.03e−18, Welch's t−test
Pr. meth. (M−value)
2
1
0
−1
AN
T
TNS1; p−value=4.32e−08, Welch's t−test
GB. meth. (M−value)
3
2
1
AN
T
TNS1; cor(AN's)=0.41, cor(T's)=0.6, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
−1
0
1
2
sampleType
AN
T
TNS1; cor(AN's)=−0.22, cor(T's)=−0.27, Spearman's rho
Expression (reads per million)
1000
100
10
Pr. meth. (M−value)
sampleType
AN
T
TNS1; cor(AN's)=−0.44, cor(T's)=−0.019, Spearman's rho
Expression (reads per million)
1000
100
10
1
2
3
GB. meth. (M−value)
sampleType
AN
T


GYG2; p−value=6.21e−305, edgeR
Expression (reads per million)
AN
T
GYG2; p−value=2.65e−08, Welch's t−test
Pr. meth. (M−value)
GYG2; p−value=0.00173, Welch's t−test
GB. meth. (M−value)
GYG2; cor(AN's)=0.72, cor(T's)=0.55, Spearman's rho
GYG2; cor(AN's)=−0.33, cor(T's)=−0.12, Spearman's rho
GYG2; cor(AN's)=−0.1, cor(T's)=0.098, Spearman's rho
sampleType AN T
KLB; p−value=9.64000000000013e−311, edgeR
KLB; p−value=0.0111, Welch's t−test
KLB; p−value=0.991, Welch's t−test
KLB; cor(AN's)=0.13, cor(T's)=0.27, Spearman's rho
KLB; cor(AN's)=−0.57, cor(T's)=0.037, Spearman's rho
KLB; cor(AN's)=0.17, cor(T's)=−0.049, Spearman's rho


ACVR1C; p−value=5.74e−302, edgeR
Expression (reads per million)
AN
T
ACVR1C; p−value=0.00412, Welch's t−test
Pr. meth. (M−value)
ACVR1C; p−value=8.33e−09, Welch's t−test
GB. meth. (M−value)
ACVR1C; cor(AN's)=0.77, cor(T's)=0.86, Spearman's rho
ACVR1C; cor(AN's)=0.02, cor(T's)=0.054, Spearman's rho
ACVR1C; cor(AN's)=−0.037, cor(T's)=0.016, Spearman's rho
sampleType AN T
KCNIP2; p−value=3.77e−283, edgeR
KCNIP2; p−value=0.00308, Welch's t−test
KCNIP2; p−value=1.29e−14, Welch's t−test
KCNIP2; cor(AN's)=0.85, cor(T's)=0.7, Spearman's rho
KCNIP2; cor(AN's)=0.023, cor(T's)=−0.0017, Spearman's rho
KCNIP2; cor(AN's)=0.12, cor(T's)=−0.022, Spearman's rho

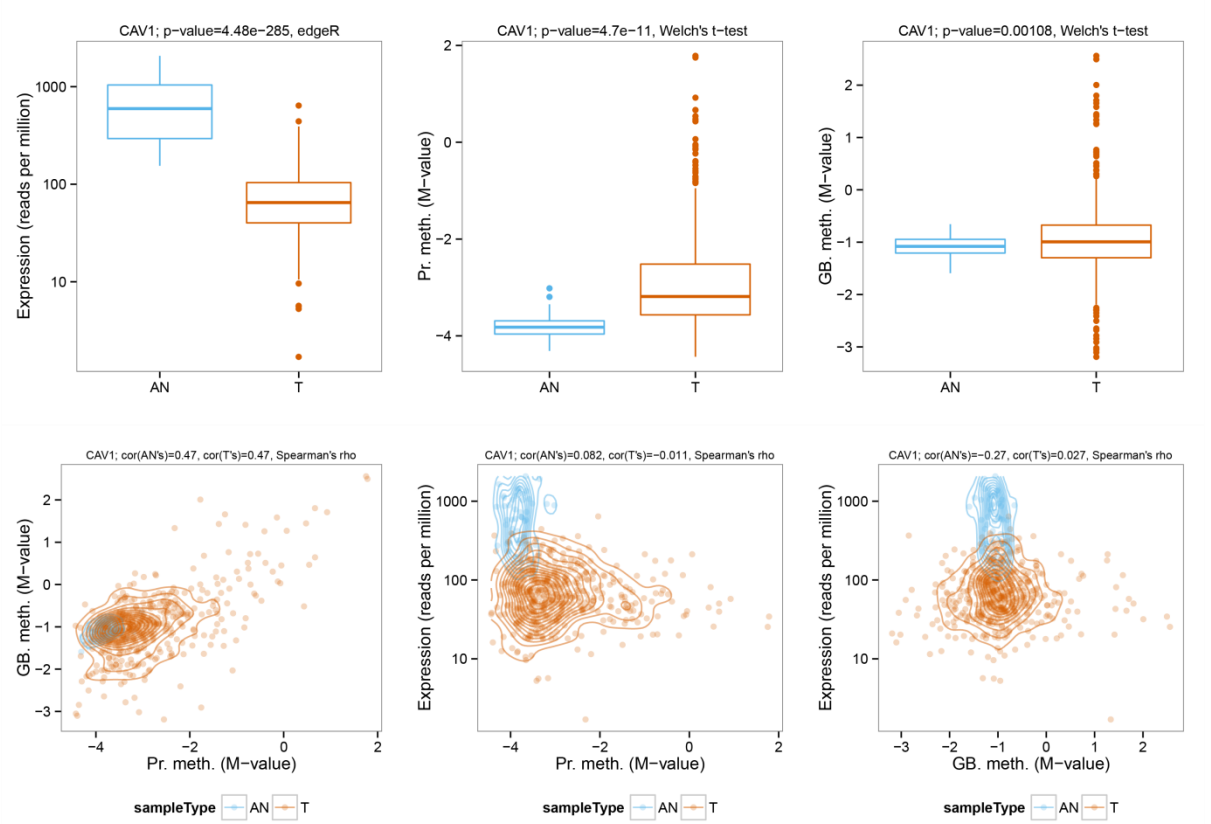
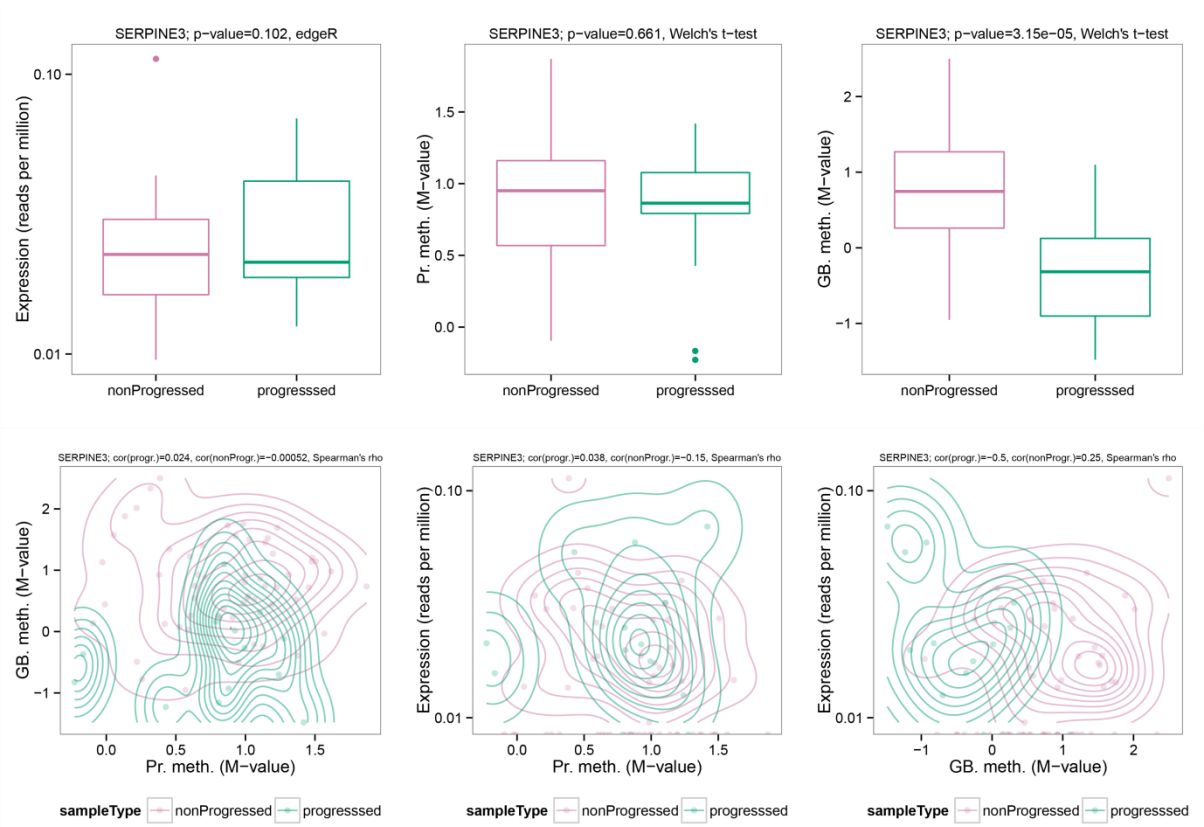
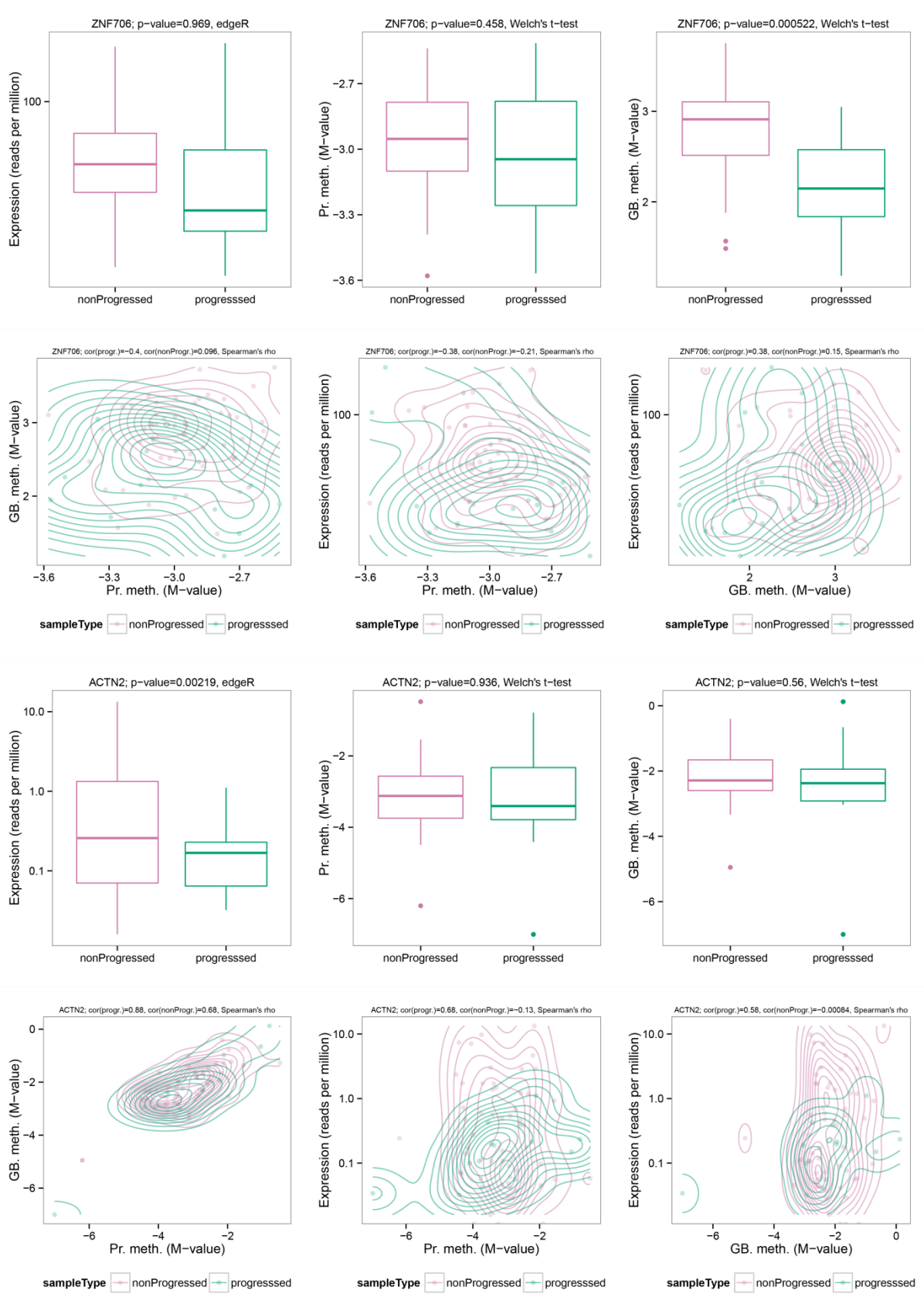
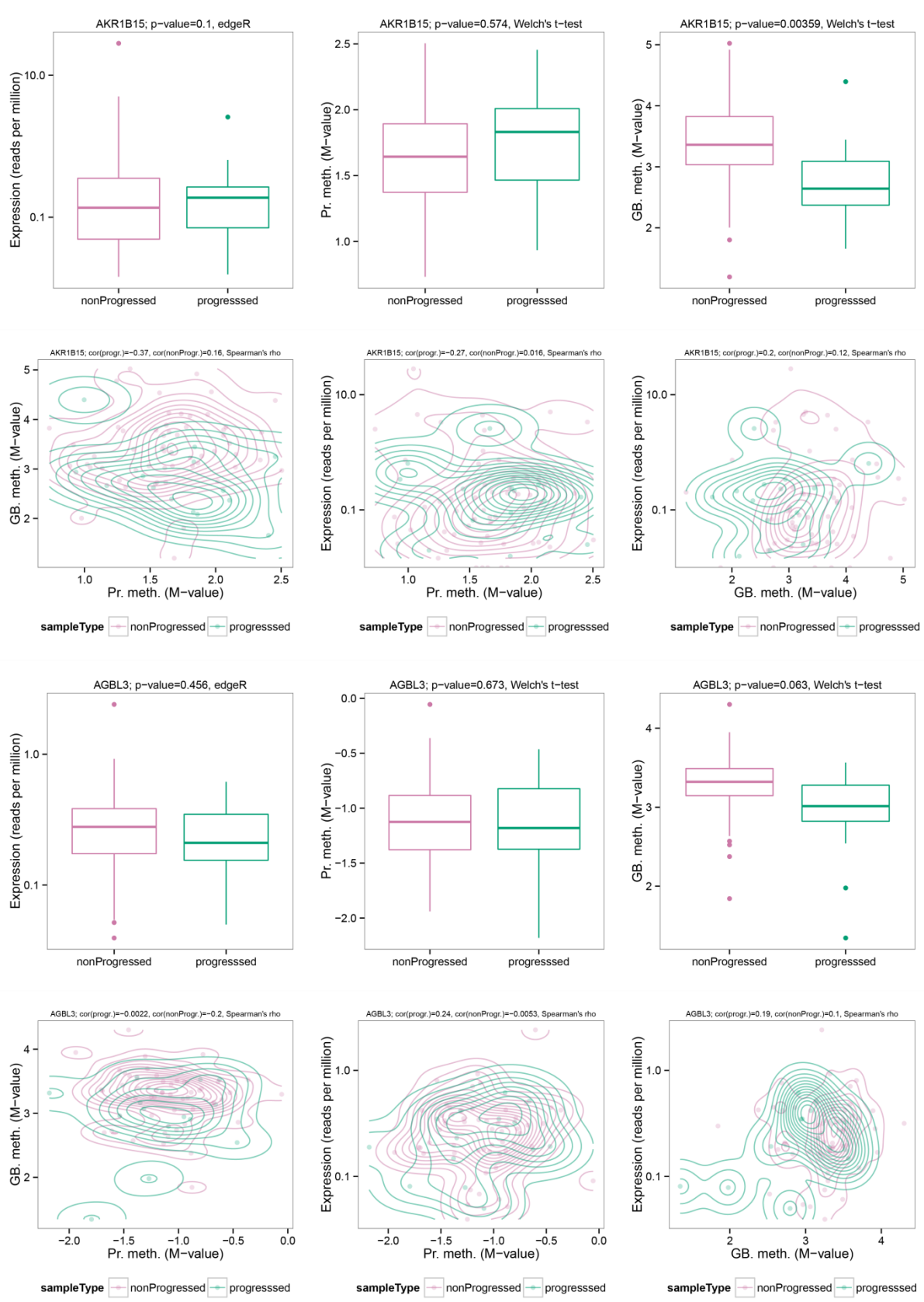
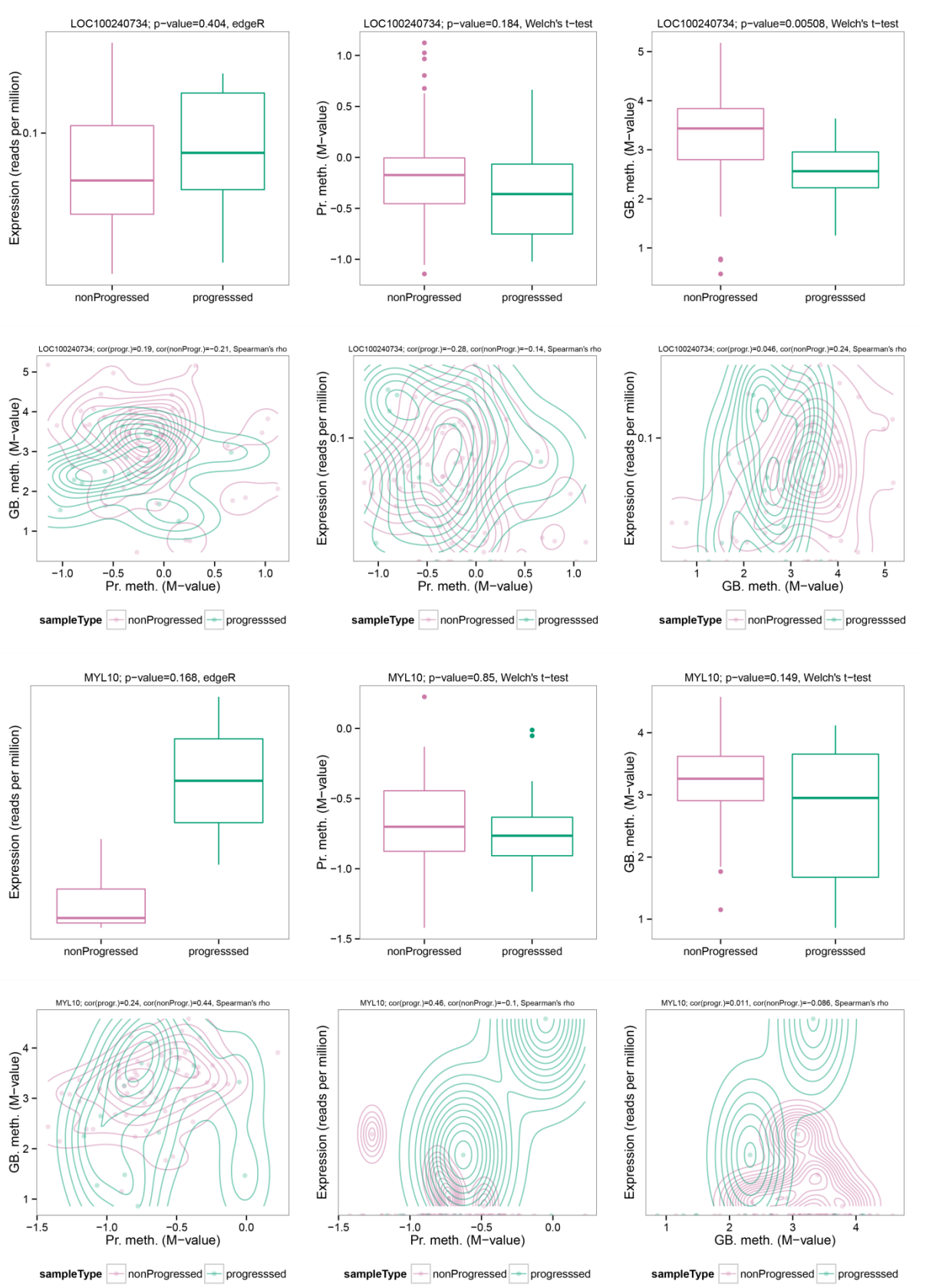

CAV1; p−value=4.48e−285, edgeR
Expression (reads per million)
1000
100
10
AN
T
CAV1; p−value=4.7e−11, Welch's t−test
Pr. meth. (M−value)
2
0
−2
−4
AN
T
CAV1; p−value=0.00108, Welch's t−test
GB. meth. (M−value)
2
1
0
−1
−2
−3
AN
T
CAV1; cor(AN's)=0.47, cor(T's)=0.47, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType AN T
CAV1; cor(AN's)=0.082, cor(T's)=−0.011, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType AN T
CAV1; cor(AN's)=−0.27, cor(T's)=0.027, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType AN T

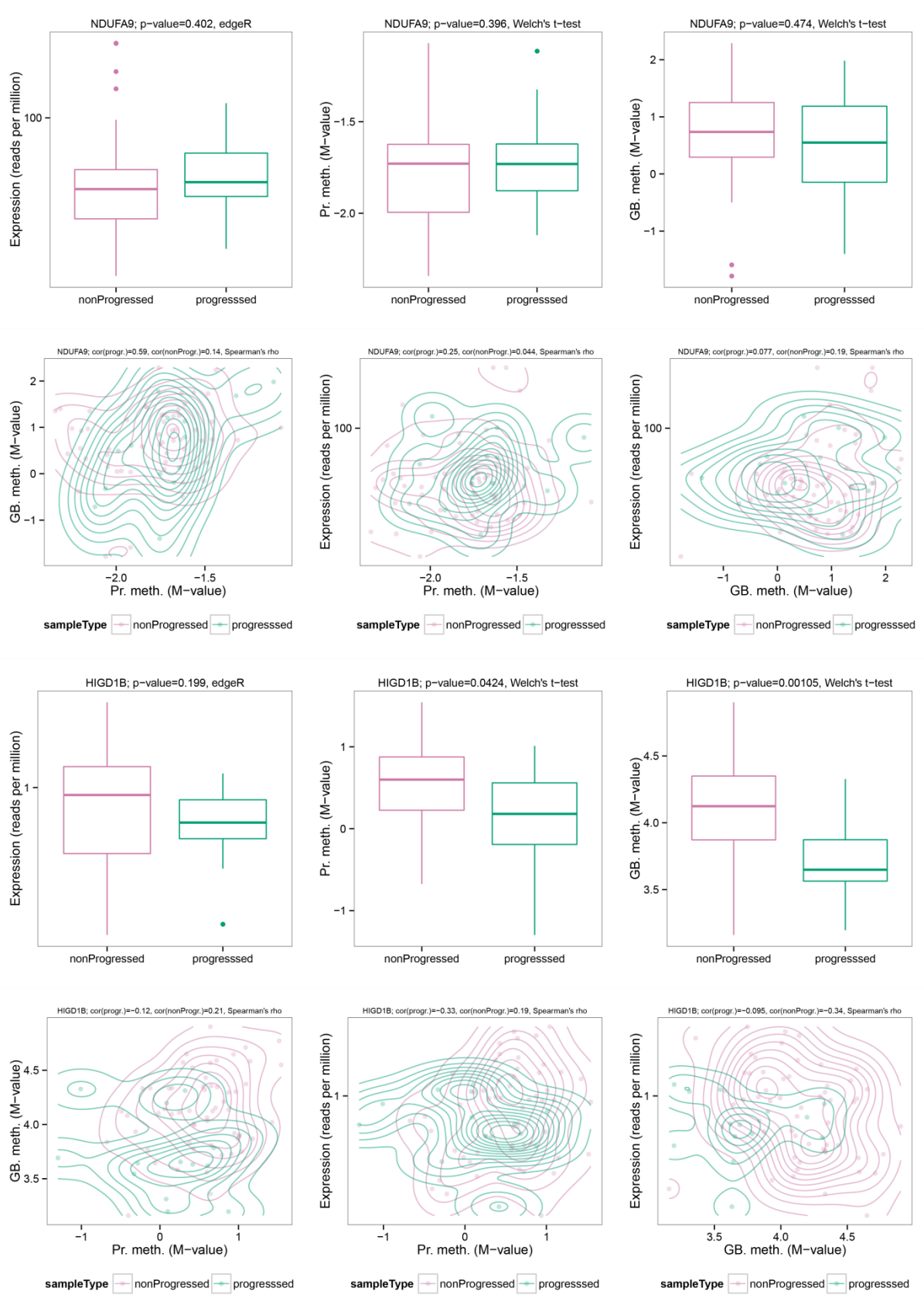
Figure S10 Marginal and pairwise distributions of gene expression, promoter methylation, and gene body methylation for the top-10 genes identified by integrative PINCAGE in comparison between progressing and non-progressing tumour samples. For each gene **Top rows:** Marginal distributions of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value across BRCA progressed and non-progressed samples. For each gene **Bottom rows:** Pairwise distributions of the three data types. Normal-reference-based kernel density contours (Venables, et al., 2002) shown for both progressed (green) and non-progressed samples (violet).

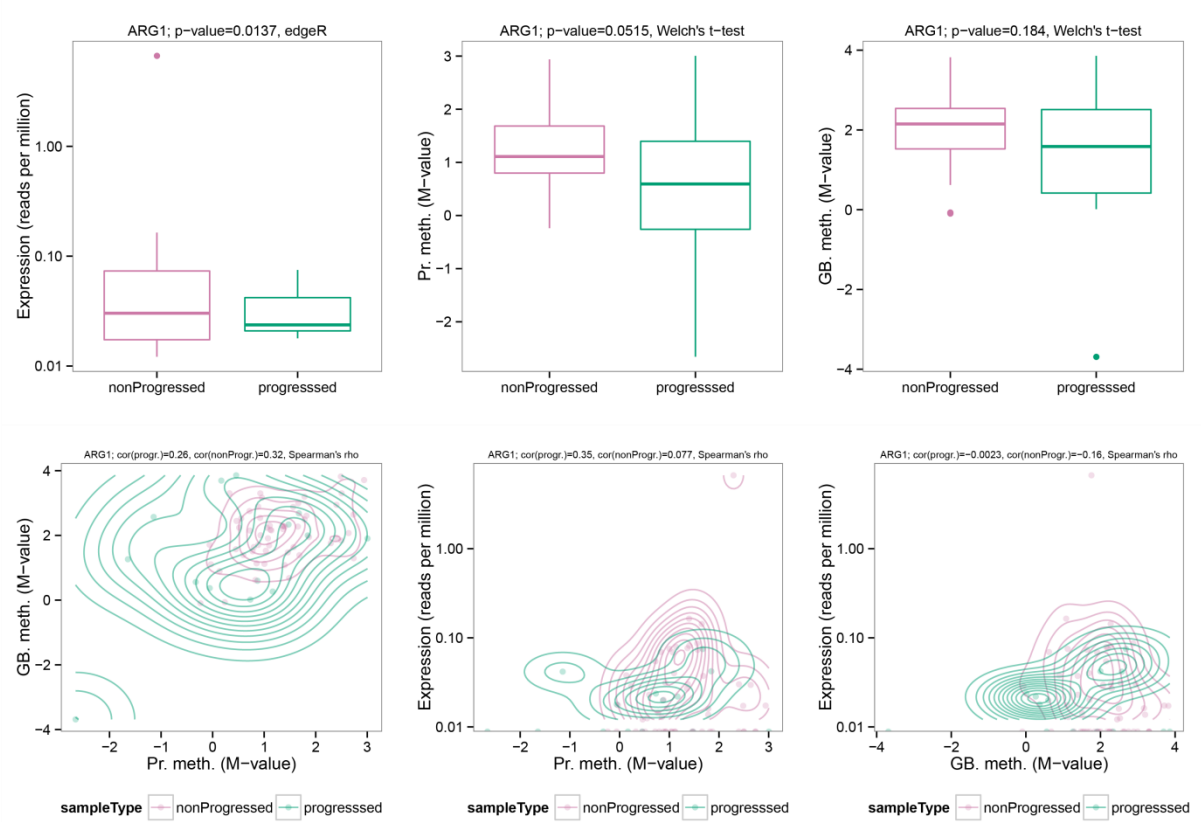



ZNF706; p−value=0.969, edgeR
Expression (reads per million)
nonProgressed
progresssed
ZNF706; p−value=0.458, Welch's t−test
Pr. meth. (M−value)
ZNF706; p−value=0.000522, Welch's t−test
GB. meth. (M−value)
ZNF706; cor(progr.)=−0.4, cor(nonProgr.)=0.096, Spearman's rho
ZNF706; cor(progr.)=−0.38, cor(nonProgr.)=−0.21, Spearman's rho
ZNF706; cor(progr.)=0.38, cor(nonProgr.)=0.15, Spearman's rho
sampleType
ACTN2; p−value=0.00219, edgeR
ACTN2; p−value=0.936, Welch's t−test
ACTN2; p−value=0.56, Welch's t−test
ACTN2; cor(progr.)=0.88, cor(nonProgr.)=0.68, Spearman's rho
ACTN2; cor(progr.)=0.68, cor(nonProgr.)=−0.13, Spearman's rho
ACTN2; cor(progr.)=0.58, cor(nonProgr.)=−0.00084, Spearman's rho


AKR1B15; p−value=0.1, edgeR
Expression (reads per million)
nonProgressed
progresssed
AKR1B15; p−value=0.574, Welch's t−test
Pr. meth. (M−value)
AKR1B15; p−value=0.00359, Welch's t−test
GB. meth. (M−value)
AKR1B15; cor(progr.)=−0.37, cor(nonProgr.)=0.16, Spearman's rho
AKR1B15; cor(progr.)=−0.27, cor(nonProgr.)=0.016, Spearman's rho
AKR1B15; cor(progr.)=0.2, cor(nonProgr.)=0.12, Spearman's rho
sampleType
AGBL3; p−value=0.456, edgeR
AGBL3; p−value=0.673, Welch's t−test
AGBL3; p−value=0.063, Welch's t−test
AGBL3; cor(progr.)=−0.0022, cor(nonProgr.)=−0.2, Spearman's rho
AGBL3; cor(progr.)=0.24, cor(nonProgr.)=−0.0053, Spearman's rho
AGBL3; cor(progr.)=0.19, cor(nonProgr.)=0.1, Spearman's rho


LOC100240734; p−value=0.404, edgeR
Expression (reads per million)
nonProgressed
progresssed
LOC100240734; p−value=0.184, Welch's t−test
Pr. meth. (M−value)
LOC100240734; p−value=0.00508, Welch's t−test
GB. meth. (M−value)
LOC100240734; cor(progr.)=0.19, cor(nonProgr.)=−0.21, Spearman's rho
LOC100240734; cor(progr.)=−0.28, cor(nonProgr.)=−0.14, Spearman's rho
LOC100240734; cor(progr.)=0.046, cor(nonProgr.)=0.24, Spearman's rho
sampleType
MYL10; p−value=0.168, edgeR
MYL10; p−value=0.85, Welch's t−test
MYL10; p−value=0.149, Welch's t−test
MYL10; cor(progr.)=0.24, cor(nonProgr.)=0.44, Spearman's rho
MYL10; cor(progr.)=0.46, cor(nonProgr.)=−0.1, Spearman's rho
MYL10; cor(progr.)=0.011, cor(nonProgr.)=−0.086, Spearman's rho


NDUFA9; p−value=0.402, edgeR
Expression (reads per million)
nonProgressed
progresssed
NDUFA9; p−value=0.396, Welch's t−test
Pr. meth. (M−value)
NDUFA9; p−value=0.474, Welch's t−test
GB. meth. (M−value)
NDUFA9; cor(progr.)=0.59, cor(nonProgr.)=0.14, Spearman's rho
NDUFA9; cor(progr.)=0.25, cor(nonProgr.)=0.044, Spearman's rho
NDUFA9; cor(progr.)=0.077, cor(nonProgr.)=0.19, Spearman's rho
sampleType
HIGD1B; p−value=0.199, edgeR
HIGD1B; p−value=0.0424, Welch's t−test
HIGD1B; p−value=0.00105, Welch's t−test
HIGD1B; cor(progr.)=−0.12, cor(nonProgr.)=0.21, Spearman's rho
HIGD1B; cor(progr.)=−0.33, cor(nonProgr.)=0.19, Spearman's rho
HIGD1B; cor(progr.)=−0.095, cor(nonProgr.)=−0.34, Spearman's rho


ARG1; p−value=0.0137, edgeR
Expression (reads per million)
1.00
0.10
0.01
nonProgressed
progresssed
ARG1; p−value=0.0515, Welch's t−test
Pr. meth. (M−value)
3
2
1
0
−1
−2
nonProgressed
progresssed
ARG1; p−value=0.184, Welch's t−test
GB. meth. (M−value)
4
2
0
−2
−4
nonProgressed
progresssed
ARG1; cor(progr.)=0.26, cor(nonProgr.)=0.32, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType
nonProgressed
progresssed
ARG1; cor(progr.)=0.35, cor(nonProgr.)=0.077, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType
nonProgressed
progresssed
ARG1; cor(progr.)=−0.0023, cor(nonProgr.)=−0.16, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType
nonProgressed
progresssed

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 & differentiation : the molecular biology journal of the American Association for Cancer Research*, **11**, 137-147.
Bibikova, M.*, et al.* (2011) High density DNA methylation array with single CpG site resolution, *Genomics*, **98**, 288-295.
Brendle, A.*, et al.* (2009) Single nucleotide polymorphisms in chromosomal instability genes and risk and clinical outcome of breast cancer: A Swedish prospective case-control study, *Eur J Cancer*, **45**, 435-442.
Brockmoller, S.F.*, et al.* Integration of metabolomics and expression of glycerol-3-phosphate acyltransferase (GPAM) in breast cancer-link to patient survival, hormone receptor status, and metabolic profiling.
Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours, *Nature*, **490**, 61-70.
Eeles, R.A.*, et al.* (2013) Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array, *Nature genetics*, **45**, 385-391.
Forbes, S.A.*, et al.* (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC), *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, **Chapter 10**, Unit 10 11.
Harris, R.A.*, et al.* Cluster analysis of an extensive human breast cancer cell line protein expression map database.
Liu, Z.Y.*, et al.* (2012) Nek2C functions as a tumor promoter in human breast tumorigenesis, *Int J Mol Med*, **30**, 775-782.
Maire, V.*, et al.* (2013) Polo-like Kinase 1: A Potential Therapeutic Option in Combination with Conventional Chemotherapy for the Management of Patients with Triple-Negative Breast Cancer, *Cancer research*, **73**, 813-823.
Martuszewska, D.*, et al.* Tensin3 is a negative regulator of cell migration and all four Tensin family members are downregulated in human kidney cancer.
Matsugi, S.*, et al.* (2007) Serum carboxypeptidase A activity as a biomarker for early-stage pancreatic carcinoma, *Clin Chim Acta*, **378**, 147-153.
Pinilla, S.M.*, et al.* Caveolin-1 expression is associated with a basal-like phenotype in sporadic and hereditary breast cancer.
Pitner, M.K.H. and Saavedra, H.I. (2013) Cdk4 and Nek2 Signal Binucleation and Centrosome Amplification in a Her2+Breast Cancer Model, *Plos One*, **8**.
Poh, W.*, et al.* Klotho-beta overexpression as a novel target for suppressing proliferation and fibroblast growth factor receptor-4 signaling in hepatocellular carcinoma.
Polzehl, J. and Spokoiny, V. (2006) Propagation-separation approach for local likelihood estimation, *Probab Theory Rel*, **135**, 335-362.
Salhab, M.*, et al.* (2012) High TIMM17A expression is associated with adverse pathological and clinical outcomes in human breast cancer, *Breast Cancer-Tokyo*, **19**, 153-160.
Sellick, G.S.*, et al.* (2004) Mutations in PTF1A cause pancreatic and cerebellar agenesis, *Nature genetics*, **36**, 1301-1305.
Shah, S.P.*, et al.* (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution, *Nature*, **461**, 809-U867.
Tang, B.L. and Ng, E.L. (2009) Rabs and Cancer Cell Motility, *Cell Motil Cytoskel*, **66**, 365-370.
Timoshenko, A.V.*, et al.* (2006) COX-2-mediated stimulation of the lymphangiogenic factor VEGF-C in human breast cancer.
Tsunoda, N.*, et al.* (2009) Nek2 as a novel molecular target for the treatment of breast carcinoma, *Cancer Sci*, **100**, 111-116.
Uckun, F.M.*, et al.* (2007) Anti-breast cancer activity of LFM-A13, a potent inhibitor of Polo-like kinase (PLK), *Bioorgan Med Chem*, **15**, 800-814.
V, T.A.*, et al.* COX-2-mediated stimulation of the lymphangiogenic factor VEGF-C in human breast cancer.
Valsasina, B.*, et al.* (2012) NMS-P937, an Orally Available, Specific Small-Molecule Polo-like Kinase 1 Inhibitor with Antitumor Activity in Solid and Hematologic Malignancies, *Mol Cancer Ther*, **11**, 1006-1016.
Van den Eynden, G.G.*, et al.* Overexpression of caveolin-1 and -2 in cell lines and in human samples of inflammatory breast cancer.
Van der Auwera, I.*, et al.* Increased angiogenesis and lymphangiogenesis in inflammatory versus noninflammatory breast cancer by real-time reverse transcriptase-PCR gene expression quantification.

Venables, W.N., Ripley, B.D. and Venables, W.N. (2002) *Modern applied statistics with S*. Statistics and computing. Springer, New York.
Vogelstein, B*., et al.* (2013) Cancer Genome Landscapes, *Science*, **339**, 1546-1558.
Wang, S.L*., et al.* (2012) Nek2A contributes to tumorigenic growth and possibly functions as potential therapeutic target for human breast cancer, *J Cell Biochem*, **113**, 1904-1914.
Xu, X.E*., et al.* (2010) Quantitative proteomics study of breast cancer cell lines isolated from a single patient: Discovery of TIMM17A as a marker for breast cancer, *Proteomics*, **10**, 1374-1390.
Zeng, F*., et al.* Reduced expression of activin receptor-like kinase 7 in breast cancer is associated with tumor progression.
Zhang, F*., et al.* Four-and-a-half-LIM protein 1 down-regulates estrogen receptor alpha activity through repression of AKT phosphorylation in human breast cancer cell.

Chapter 9: Manuscript 2

Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data

Manuscript in preparation

Michał P. Świtnicki[1], Tobias Madsen[1], Jakob S. Pedersen[1,2]

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Brendstrupgårdsvej 21, 8200 Aarhus, Denmark and [2]Bioinformatics Research Centre (BiRC), Aarhus University, C.F.Møllers Allé 8, 8000 Aarhus, Denmark

Running head: *NA*

Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data

Michał P. Świtnicki[1], Tobias Madsen[1], Jakob S. Pedersen[1,2]

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Brendstrupgårdsvej 21, 8200 Aarhus, Denmark and [2]Bioinformatics Research Centre (BiRC), Aarhus University, C.F. Møllers Allé 8, 8000 Aarhus, Denmark

Abstract

Motivation: Cancer development and progression is driven by a complex pattern of genomic and epigenomic perturbations. Effects of these perturbations are then manifested as aberrant gene expression that can affect the studied disease. Hence, different genomic data types are inherently interdependent and their integrative analysis may therefore improve detection of perturbed genes and prediction of disease state. In contrast analysis using general purpose methods based on independence assumptions will make inefficient use of the data and potentially lead to false conclusions. Thus, expert knowledge can be utilized to inform the design of integrative tools that make more optimal use of available data.

Model: Here we present a sparsely parameterized probabilistic model integrating RNA-seq gene expression and 450K array DNA methylation of promoters and gene bodies. It accounts for the dependence between expression and methylation in an attempt to identify integrative biomarkers. It is specified using modular graphical models, enabling future expansion with additional data types. Due to its general parameter sparseness, it permits robust inference even in small cohorts.

Results: We apply our approach to a Breast Invasive Carcinoma data set from The Cancer Genome Atlas consortium, which includes 82 adjacent normal and 730 cancer samples. We identify new biomarker candidates of breast cancer development (TMEM132D, CACNG3, FXYD1, NRSN1, KIR3DX1, LOC388692) and progression (ZFATAS, KAAG1, SERPINE3). The discriminatory performance of the proposed model on individual biomarkers is comparable to established methods assuming independence such as logistic regression, but it better combines evidence across multiple selected genes. Our method can be used for integrative biomarker identification of any genomic disease, especially when the cohort size is small.

Introduction

The overreaching goal of cancer studies is to improve diagnosis, prognosis and treatment of patients. In recent years, high-throughput molecular profiling technologies were widely adopted

by clinical cancer researchers (Sulakhe, et al., 2014), promising to resolve difficulties in hard discrimination problems, for instance between cancer grades. To fulfil the promise, clinicians need biomarkers for the specific disease that are characterized by good discriminatory power. 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). Combined use of biomarkers of different molecular types is expected to improve discriminatory power and clinical performance (Kristensen, et al., 2014). However, the performance gains over using gene expression alone were on average not significant so far, excluding some special cases (Ray, et al., 2014).

We hypothesize that the predictive performance of combined biomarkers can be improved by including existing knowledge on the biological relationships between the different molecular types that these biomarkers are based on. Indeed, we have previously developed a model-based integrative approach that demonstrated such improved predictive performance of novel biomarker candidates for breast cancer progression (Świtnicki, et al., 2015, *in review*). However, our previous approach required relatively high number of training samples to allow robust inference. Hence, in this publication we propose a similar, yet parameter-sparser model-based strategy for identification of integrative biomarkers, which can easily be extended to the increasing array of molecular profiling data types becoming available. The relative sparsity of parameterization should permit analysis and classification in smaller sample sets.

Both gene expression and DNA methylation have long been studied as cancer biomarker candidates (Berse and Lynch, 2015; Parrella, 2010). Individual laboratories typically include only relatively few patients and profile only a single data type when screening for new biomarkers. In recent years, however, more recognition has been given to the necessity of generating multiple molecular profiling data for the same study subjects. Specifically, large patient cohorts profiled for several molecular marks with hundreds of patients are now available from the International Cancer Genome Consortium (ICGC; (Zhang, et al., 2011)) or The Cancer Genome Atlas (TCGA; (Weiss, 2005)). In contrast, smaller research centres cannot generate such massive data sets for their research needs. All these data sets offer new opportunities for exploring and developing integrative predictive approaches. However, new methods should be able to robustly analyse small datasets too as it would greatly expand their applicability domain.

Most classification methods operate on a multivariate principle: first, a selection of relevant features among multiple data types is performed, and second, a multivariate model is trained. A selection of features among normalized data types can be done for example using elastic net (Zou and Hastie, 2005) or Lasso (Tibshirani, 2011). Then, general-purpose machine learning

methods are applied, such as different regressions, random forests, support vector machines, or clustering (Kristensen, et al., 2014; Ray, et al., 2014). These methods, however, typically miss dependencies between data types. Also, interpretation of most of these compound models is difficult and additional feature importance analyses must be performed to elucidate the biomarker candidates.

A class of models characterized by structured integration using prior knowledge is an attractive alternative to the classical multivariate approach. This approach explicitly incorporates prior understanding of the structure of possible interactions between data types. PARADIGM, a well know approach utilizing this strategy (Vaske, et al., 2010) 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. While attractive, PARADIGM simplifies each data type it integrates into three discrete states: nominal, activated and repressed. Also, it identifies affected pathways, rather than individual markers, which is attractive for basic research but not for clinicians working with biomarkers. Hence, a univariate approach is preferred in the clinics.

Here we propose a gene-oriented (essentially univariate) sparse structured integrative model, which includes DNA methylation at individual CpG sites and mRNA expression. The model is modular and may be extended to other data types, as needed. 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 demonstrate its use by analysing DNA methylation and gene expression in the Breast Invasive Carcinoma (BRCA) dataset (Cancer Genome Atlas, 2012).

Materials & Methods

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 (Table 1). The resulting data set consisted of 730 tumour (T) samples and 82 Adjacent Normal (AN) samples. We also defined subsets of progressing (n=14) and non-progressing (n=57) BRCA tumours based on presence or absence of recurrence within close to 3 years of treatment (Table S2).

The more detailed description of the data processing is given in (Świtnicki, et al., 2015, *in review*). In short, the 450k methylation array data was processed using the statistical language R

(R Core Team, 2014) by parsing raw data and inferring peak-corrected (Dedeurwaerder, et al., 2011) M-values (Aryee, et al., 2014). M-values are defined as logit-transformed beta-values, which is a standard metric for the platform, and are preferred for differential analysis due to their homoscedasticity (Du, et al., 2010).

Promoters and gene bodies were defined using native Illumina's probe annotation categories (450k Manifest File v1.2 (Bibikova, et al., 2011)). Gene bodies were defined with Illumina's Gene Body and 3' UTR regions while promoters with TSS1500, TSS200, 5' UTR and 1st Exon (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 was already summarized per gene and no further processing was needed.

The data was summarized and organized by disease groups (T vs 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) (Table 1). The data types, their distribution across samples, and their pairwise correlations are exemplified by the TMEM132C gene (Fig. 1).

Table 1 Definition of data sets: sizes and 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.

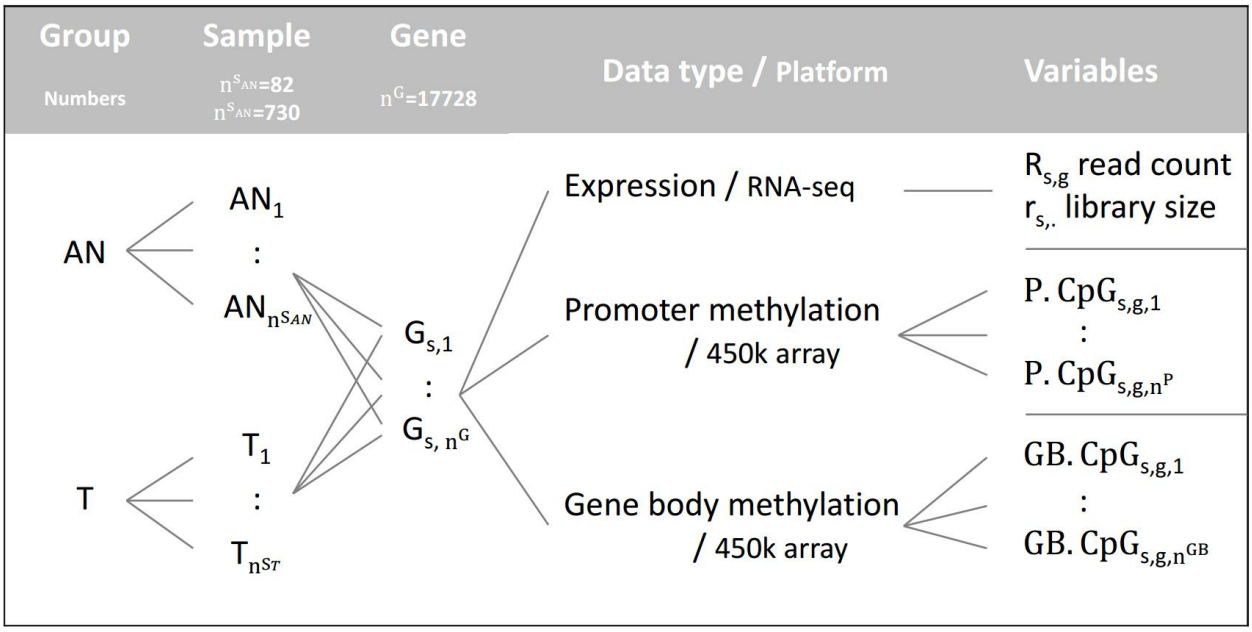
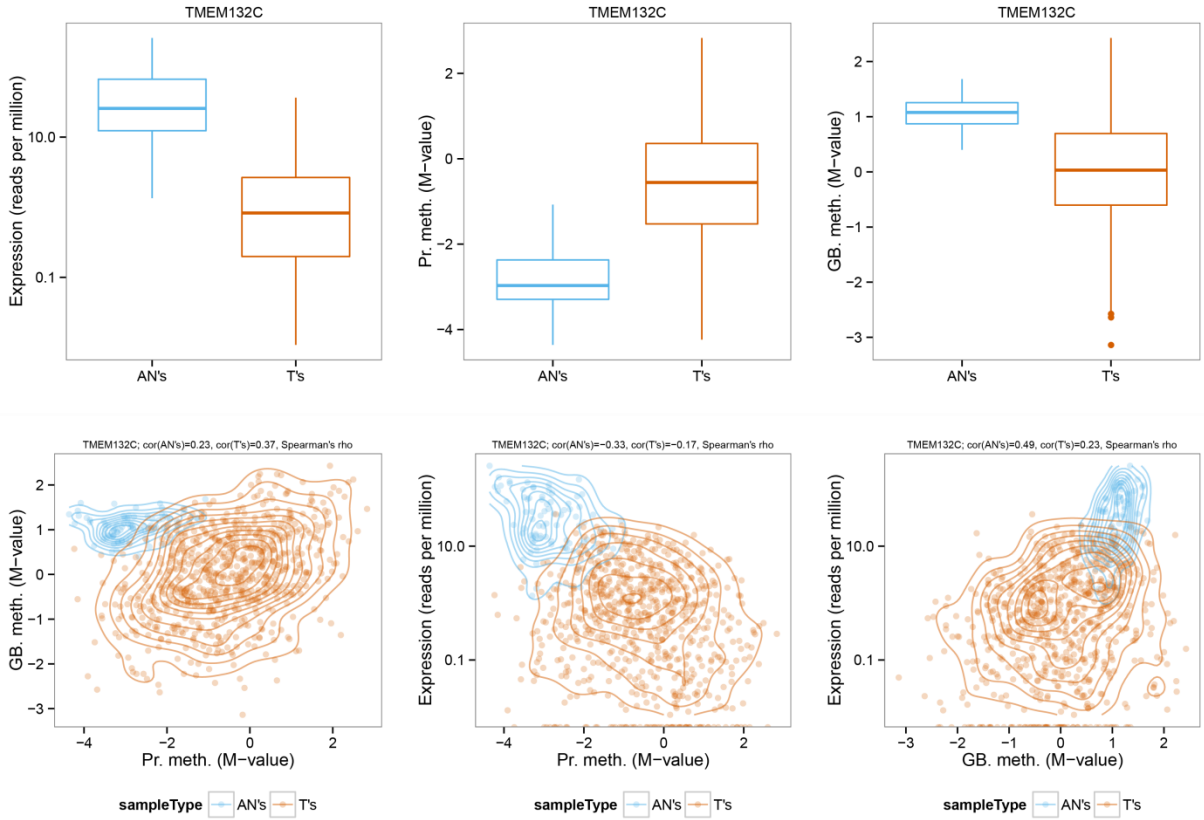


Fig. 1 Marginal and pairwise distribution of gene expression, promoter methylation, and gene body methylation for the TMEM132C gene. A) Marginal distribution of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value across BRCA Tumour (T) and Adjacent 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).



Model specification

We specify a model for each gene separately using a probabilistic graphical model (Fig. 2). Graphical models are a convenient tool for encoding expert knowledge from literature and for conveying relationships in a clear visual form. Our model is able to define a joint distribution of the observed data as well as to capture potential dependencies between data types, as seen for the TMEM132C gene (Fig. 1). We introduce our model by describing the data type specific probability distributions first, and then combining these parts into joint distribution under integrative model.

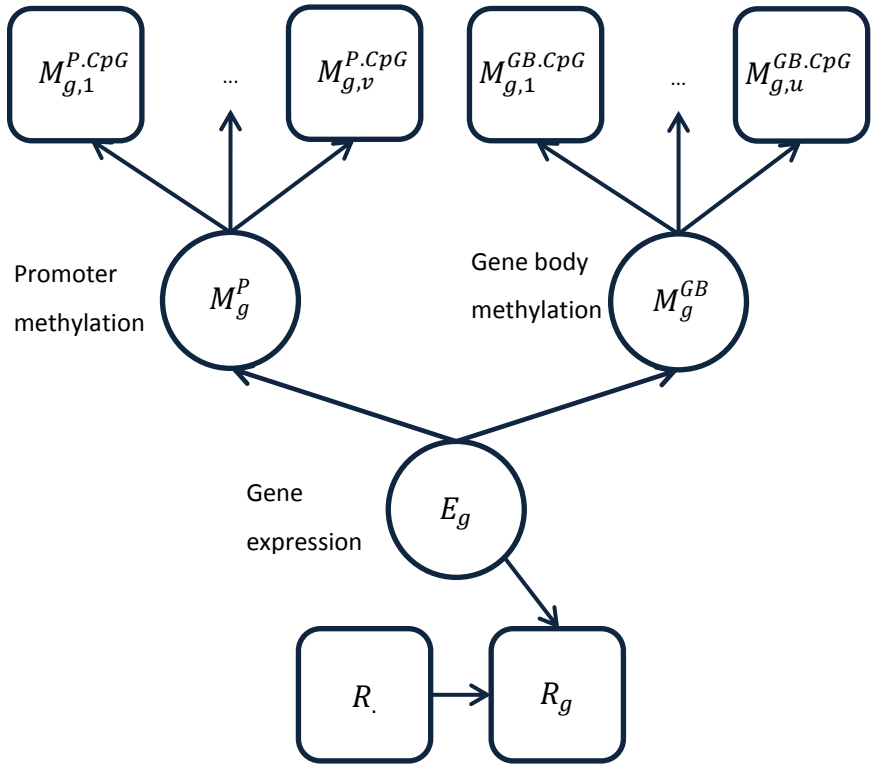


Fig. 2 Directed acyclic graph representation of our probabilistic graphical model. Variables in square boxes are directly observed while variables in circles are inferred.

Gene expression

The expression, E_g of a gene g, is the expected fraction of transcripts in the total pool of transcripts that maps to the gene. RNA-seq can be viewed as sampling of transcripts, thus we do not directly observe E_g, but instead a library size, L, i.e. the total number of sampled transcripts, and a read count, R_g, the number of transcripts mapping to g. Modelling R_g using a beta-binomial model, we have:

$$E_g \sim Beta(\alpha, \beta) \quad (1)$$

$$R_g \sim Binom(L, E_g) \quad (2)$$

DNA methylation

We previously defined the gene body and promoter regions in the *Data sources* section as we model these two categories separately. This choice comes from our understanding of their different roles in regulation of transcription (Jjingo, et al., 2012; Jones, 2012; You and Jones, 2012), but also from our observation of their different empirical distributions (Świtnicki, et al., 2015). We denote the CpG site methylation measurements for given gene's promoter region by $M_{g,1}^p, \ldots M_{g,n_g^p}^p$ and by $M_{g,1}^{gb}, \ldots M_{g,n_g^{gb}}^{gb}$ for the gene body region, respectively. We propose a model

where the methylation of each promoter and gene body's consisting CpG sites is governed by an unobserved methylation state variables M_g^p, M_g^{gb} i.e.

$$M_{g,j}^p \sim N\left(M_g^p, {\sigma_g^p}^2\right) \quad (3)$$

$$M_{g,j}^{gb} \sim N\left(M_g^{gb}, {\sigma_g^{gb}}^2\right) \quad (4)$$

Joint distribution

The link between expression and methylation is modelled by a linear relationship between unobserved expression and methylation states.

$$M_g^p \sim N\left(a^p E_g + b^p, {\sigma_g^{e,p}}^2\right) \quad (5)$$

$$M_g^{gb} \sim N\left(a^{gb} E_g + b^{gb}, {\sigma_g^{e,gb}}^2\right) \quad (6)$$

Given the conditional independence assumptions the graphical model encodes (Fig. 2), the joint distribution of the complete data, $D^{full} = \{E, R, L, M^p, \{M_j^p\}_{j=1}^{n^p}, M^{gb}, \{M_j^{gb}\}_{j=1}^{n^{gb}}\}$ is now fully specified. The likelihood of the parameters, $\theta = (\alpha, \beta, {\sigma^{e,p}}^2, {\sigma^p}^2, a^p, b^p, {\sigma^{e,gb}}^2, {\sigma^{gb}}^2, a^{gb}, b^{gb})$ can be calculated as

$$\begin{aligned} L\left(\theta, D^{full}\right) &= Binom(R; L, E)\, Beta(E; \alpha, \beta)\, \emptyset\left(M^p; a^p E + b^p, {\sigma^{e,p}}^2\right) \emptyset\left(M^{gb}; a^{gb} E \right. \\ &\quad \left. + b^{gb}, {\sigma^{e,gb}}^2\right) \prod_{j=1}^{n^p} \emptyset\left(M_j^p; M_p, {\sigma_p}^2\right) \prod_{j=1}^{n^{gb}} \emptyset\left(M_j^{gb}; M_{gb}, {\sigma_{gb}}^2\right) \\ &= \binom{L}{R} \frac{E^{\alpha-1+R}(1-E)^{\beta-1+L-R}}{B(\alpha,\beta)}\, \emptyset\left(M^p; a^p E \right. \\ &\quad \left. + b^p, {\sigma^{e,p}}^2\right) \emptyset\left(M^{gb}; a^{gb} E \right. \\ &\quad \left. + b^{gb}, {\sigma^{e,gb}}^2\right) \prod_{j=1}^{n^p} \emptyset\left(M_j^p; M_p, {\sigma_p}^2\right) \prod_{j=1}^{n^{gb}} \emptyset\left(M_j^{gb}; M_{gb}, {\sigma^{gb}}^2\right) \end{aligned} \quad (7)$$

Parameter inference

We have made a factor graph R library called dgRaph to implement the above probabilistic graphical model, which handles only discrete random variables. The model implementation therefore relies on discretization of the continuous random variables. In theory, calculating the

likelihood of the observed data $D = \{R, L, \{M_j^p\}_{j=1}^{n^p}, \{M_j^{gb}\}_{j=1}^{n^{gb}}\}$ amounts to integrating out the unobserved variables $\{E, M^p, M^{gb}\}$.

$$
\begin{aligned}
& L\left(\theta, D^{obs}\right) \\
& = \int_E \int_{M^p} \int_{M^{gb}} \binom{L}{R} \frac{E^{\alpha-1+R}(1-E)^{\beta-1+L-R}}{B(\alpha,\beta)} \, \emptyset\left(M^p; a^p E + b^p, {\sigma^{e,p}}^2\right) \emptyset\left(M^{gb}; a^{gb} E \right. \\
& \left. + b^{gb}, {\sigma^{e,gb}}^2\right) \prod_{j=1}^{n^p} \emptyset\left(M_j^p; M_p, {\sigma_p}^2\right) \prod_{j=1}^{n^{gb}} \emptyset\left(M_j^{gb}; M_{gb}, {\sigma^{gb}}^2\right) dE dM^p dM^{gb}
\end{aligned}
\quad \textbf{(8)}
$$

In practice this integration is carried out numerically in the dgRaph framework, discretizing the continuous variables into at least 100 bins. We infer the parameters of the model using the EM-algorithm. As our experiments showed, the parameters of the beta distribution converged slowly in our framework and hence we decided to learn them outside of the framework using gradient descent (Venables, et al., 2002) to find maximum likelihood estimates only using the L and R variables.

Classification and gene selection

Here we show how our model is used to predict which group label is the most probable for a given sample *X* (tumour versus normal, progressing versus non-progressing, etc.). At first, *tumour* and *normal* models are trained for a gene under consideration using data from respective groups. Then, we calculate the likelihood ratio as the discriminant function (Eq. (9)).

$$
L(X) = \frac{P^{tumour}(X)}{P^{normal}(X)} \quad \textbf{(9)}
$$

To screen for integrative biomarker candidates, we evaluate these scores using ROC analysis, selecting best performing genes based on training AUC. Then, evaluation is performed and validation AUC is recorded (Fig. 3, top-1). To combine evidence from several selected genes to improve classification performance, we do so using naïve Bayes classifier, assuming independence between genes (Fig. 3, top-2 and top-3 combined). In practice, the *L(X)* is log-transformed so the combination of evidence is done by summation of scores across combined genes.

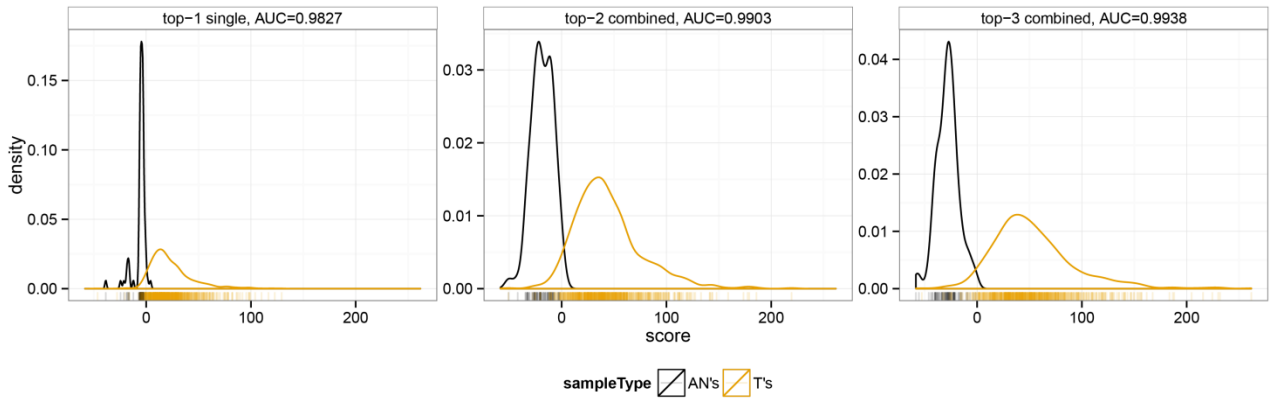


Fig. 3 Combination of evidence across ranks in the comparison between tumours (T's) and adjacent normal (AN's). Top-1 classifier and top-2 and top-3 classifiers combined, respectively. Classification improves as more evidence is absorbed by the naïve Bayes classifier.

Results

BRCA tumours vs normal

We first applied our model for the task of identifying the most discriminating genes between tumour (n=730) and normal (n=82) samples. The analysis was done using an 8-fold cross-validation strategy where 10-11 normal and 91-92 tumour samples were randomly assigned to each fold. In each fold of the procedure, a subset is held out for validation and the remaining training samples are used to (a) train classifiers for each gene, and to (b) rank genes according to the train error (we use training data AUC). This procedure produces a ranked list of genes that perform best at each fold.

Top-ranked candidates

The top-ranked genes were often shared across folds. Specifically, all genes from the top-10 according to the mean rank (Table 2, left-hand side) appeared consistently within the top-14 of each fold in the cross-validation procedure (Table S1). This finding demonstrates a high robustness of identified genes, suggesting them to be good candidate biomarkers of tumour development. The top-10 list includes genes with known associations to cancer such as TMEM132C (Chung, et al., 2013), ULBP1 (Cerwenka and Lanier, 2003), SLC6A2 (Dolled-Filhart, et al., 2006) and A2BP1 (Sengupta, et al., 2013). It also contains poorly characterized genes such as a pseudogene KIR3DX1 or a lncRNA LOC388692. Remaining genes (TMEM132D, CACNG3, FXYD1 and NRSN1) were not previously associated with any cancer type. Interestingly, some of the lowly expressed genes show similar patterns of highly correlated but differential promoter and gene body methylation in BRCA tumours compared to normal (SUPPLEMENTARY: Fig. S2,

KIR3DX1, CACNG3, A2BP1). These patterns are of high diagnostic value, far greater than gene expression alone.

Classification of tumour versus normal

We next evaluated the classification performance of the top-10 individual and combined classifiers using the cross-validation AUC (Table 2, right-hand side) and compared them to corresponding Logistic Regression (LR) classifiers trained using the same genes as found by the integrative model. For individual classifiers, the performance varies and neither our integrative model nor LR models are consistently best in the top10 (6 vs 4, respectively). The average AUC of single rank models remains very high for both methods, with the LR showing lower performance on average (0.9622 vs 0.9777, respectively). When genes are combined across ranks, AUC improves visibly for both methods. Our model combinations consistently achieve higher classification performance than LR combinations (6 vs 2, respectively) with the average difference larger than for single rank models (0.9930 vs 0.9713). LR, however, achieves highest performance of all the classifiers already at the top-2 combined classifier. The performance of LR combined classifiers then erratically degrades suggesting high variance of the combined LR models.

Table 2 Top-10 ranked genes in the evaluation of 82 normal and 730 tumour BRCA samples.

8-fold cross-validation analysis 82 normal and 730 tumour BRCA samples						
Top genes across folds		**Classification performance (AUC)**				
Mean rank	**Gene ID**	**Rank (k)**	**Sparse integrative model**		**Logistic Regression**	
			Single rank	**Combined (1-k)**	**Single rank**	**Combined (1-k)**
1.5	TMEM132D	**1**	0.9827	0.9827	**0.9916**	0.9916
2.1	TMEM132C	**2**	0.9900	0.9903	**0.9934**	**0.9956**
3.5	ULBP1	**3**	**0.9850**	0.9938	0.9786	**0.9944**
4.6	KIR3DX1	**4**	**0.9867**	0.9943	0.9637	0.9943
5.4	CACNG3	**5**	0.9705	**0.9914**	**0.9766**	0.9722
5.9	LOC388692	**6**	**0.9892**	**0.9923**	0.9617	0.9288
6.8	FXYD1	**7**	0.9565	**0.9925**	**0.9833**	0.9616
8.9	SLC6A2	**8**	**0.9815**	**0.9939**	0.9674	0.9524
10.6	NRSN1	**9**	**0.9629**	**0.9942**	0.8676	0.9639
10.8	A2BP1	**10**	**0.9717**	**0.9944**	0.9379	0.9782
		Average	**0.9777**	**0.9930**	0.9622	0.9713

BRCA progressing vs non-progressing tumours

We next applied the integrative model to the challenging problem of discriminating between progressing and non-progressing tumours. We used the recurrence after initial treatment 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 (Table S2). 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.

Classification of progressing versus non-progressing

Following the dichotomization, we applied our model to the task of identifying the most discriminating genes between progressing and non-progressing BRCA tumours. Given the very limited number of progressing tumours, a 14-fold 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. Again, we compared classification performance of our integrative models with the corresponding LR ones (Table 3, right-hand side). For individual classifiers, the performance varies and neither our integrative model nor LR models are consistently best in the top-10 (5 vs 5). The average AUC of single rank models remains low for both methods, with the LR showing marginally lower performance on average (0.6091 vs 0.6103, respectively). When models are combined across ranks, the AUC improves on average considerably for our sparse model only (0.6643) as the LR combined models did not achieve consistent improvement over individual rank models (0.6019 vs 0.6091 for combined and single rank models, respectively). All combined LR models had lower AUC than our combined sparse models. However, the best classification performance of all considered classifiers was achieved by the top-7 single rank LR model (0.7406).

Top-ranked candidates

The top-3 genes in this analysis (Table 3, left-hand side) were at the same time the most robust candidates during the 14-fold cross validation procedure (Table S3), consistently reappearing in the top-20 at each fold. The list includes ZFAT small antisense RNA (ZFATAS), kidney-associated antisense antigen 1 (KAAG1) and Serpin peptidase inhibitor, clade E, member 3 (SERPINE3). The latter candidate was previously identified as the most significant and robust candidate for the progression data set by the parameter-richer implementation of this model (Świtnicki, et al., 2015, *in review*). The KAAG1 was found activated in many tumour types when compared to the host tissue (Van Den Eynde, et al., 1999), including breast. Little is known about the last candidate from the top-3, ZFATAS, other than its classification as a long non-coding RNA. The mostly lowly expressed long non-coding RNAs may therefore show greater potential as biomarkers when methylation data is included. In fact, all top-3 candidates seem to be not differentially expressed (Fig. S3), but are differentially methylated in promoters and/or gene

bodies. Differential methylation of these genes could signify their differential splicing patterns signifying tumour progression (Oltean and Bates, 2014). Further studies would be required to confirm these findings and establish the clinical applicability.

Table 3 Top-10 ranked genes in the evaluation of 14 progressing and 57 non-progressing BRCA tumour samples.

14-fold cross-validation analysis 14 progressing and 57 non-progressing BRCA tumours						
Top genes across folds		**Classification performance (AUC)**				
			Sparse integrative model		**Logistic Regression**	
Mean rank	**Gene ID**	**Rank (k)**	**Single rank**	**Combined (1-k)**	**Single rank**	**Combined (1-k)**
2.4	ZFATAS	**1**	**0.6165**	0.6165	0.5677	0.5677
4.6	SERPINE3	**2**	0.6391	**0.6867**	**0.7055**	0.6654
6.2	KAAG1	**3**	**0.6341**	**0.6591**	0.6165	0.6541
8.1	SFRS8	**4**	0.6278	**0.6842**	**0.6717**	0.6491
14.7	DPY19L3	**5**	**0.6880**	**0.6404**	0.5113	0.5959
14.9	LOC149620	**6**	0.4612	**0.6591**	**0.5564**	0.5426
17.8	ATP9A	**7**	0.6980	**0.6692**	**0.7406**	0.5213
18.4	IQGAP2	**8**	**0.5815**	**0.6692**	0.5689	0.6028
19.8	GPBAR1	**9**	0.5313	**0.6504**	**0.5677**	0.5915
21.4	TMEM198	**10**	**0.6253**	**0.6604**	0.5852	0.5946
		Average	**0.6103**	**0.6643**	0.6091	0.6019

Discussion

Classification using multiple data types is often performed using a simplifying assumption of independence (Hamid, et al., 2009). However, this is often not the case and expert knowledge can help identify the possible interactions. Here we have introduced a sparse probabilistic graphical model for integration of multiple gene-level genomic data types. We applied our model to three types of data, for which we know the expected relationship: gene expression, promoter methylation, and gene body methylation. We integrate these by assuming simplifying linearity between the two methylation types and gene expression. This simplification permits easy interpretation of the gene models and their learned parameters.

Our sparse model permits classification based on sets of data values while considering the expected relationships amongst them. The model also accounts for both the technical and the biological variance of methylation as well as gene expression data, providing predictions robust to technical noise. Benefits of integrated classification using multiple data types are twofold. First, it enables classification of subtle simultaneous deviations of all three variables that would be too weak to detect if classified separately. Also, the inference becomes more robust to noisy

data, especially when the data types are interdependent. The reason is that the model can exploit the partial redundancy among observations.

The general parameter sparseness of the model makes it preferred in several contexts. First, the model is suitable for analysing small cohorts that are typical for smaller research centres and laboratories. Second, the execution becomes faster – our estimates show 10-time faster execution than the previous, parameter-richer implementation of this model (Świtnicki, et al., 2015, *in review*), despite using a demanding cross-validation procedure that involves model retraining at each fold. Last but not least, predictions made with the model should have lower prediction variance and be more robust. 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 to parameter-sparse models.

The weaknesses of the model include its simplifying assumptions. First, as known from the literature (Gelfman, et al., 2013; Raynal, et al., 2012; Sati, et al., 2012), relationship between DNA methylation levels and transcription rate is typically not linear, especially if the entire spectrum of sample methylation is considered. This may lead to significant prediction bias. Second, the model assumes that all tumours are sampled from a single distribution; however, many tumour genes exhibit bimodal gene expression or methylation distributions (Hinoue, et al., 2012; Wyatt, et al., 2014). The latter could be addressed by estimating two gene expression or methylation distributions in each model.

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 stratifying 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.

With the advent of new genome-wide molecular profiling technologies, it has become more important to integrate various data types. If done right, integration should facilitate optimal use of the available complementary and often supplementary information from multiple molecular levels. It can be achieved by utilizing expert knowledge of the integrated data types and ultimately lead towards better diagnosis, prognosis and treatment for patients. The cost of generating enough training data for parameter-rich models is often preventive and therefore the integration should be robust even for smaller data sets. The proposed sparse model meets these criteria; however, further studies are required to test how well the identified biomarker candidates generalize across cohorts.

The freely available software is available as R scripts with instructions on how to prepare the data for analysis at http://.

Bibliography

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.
Berse, B. and Lynch, J.A. (2015) Molecular diagnostic testing in breast cancer, *Seminars in oncology nursing*, **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, N. (2012) Comprehensive molecular portraits of human breast tumours, *Nature*, **490**, 61-70.
Cerwenka, A. and Lanier, L.L. (2003) NKG2D ligands: unconventional MHC class I-like molecules exploited by viruses and cancer, *Tissue antigens*, **61**, 335-343.
Chung, S.*, et al.* (2013) A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients, *Breast cancer research : BCR*, **15**, R81.
Dedeurwaerder, S.*, et al.* (2011) Evaluation of the Infinium Methylation 450K technology, *Epigenomics*, **3**, 771-784.
Dolled-Filhart, M.*, et al.* (2006) Classification of breast cancer using genetic algorithms and tissue microarrays, *Clinical Cancer Research*, **12**, 6459-6468.
Du, P.*, et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, *BMC bioinformatics*, **11**, 587.
Gelfman, S.*, et al.* (2013) DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure, *Genome Res*, **23**, 789-799.
Hamid, J.S.*, et al.* (2009) Data integration in genetics and genomics: methods and challenges, *Human genomics and proteomics : HGP*, **2009**.
Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) The elements of statistical learning : data mining, inference, and prediction. In, *Springer series in statistics,*. Springer, New York, NY, pp. xxii, 745 p.
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, *Nature reviews. Genetics*, **13**, 484-492.
Kristensen, V.N.*, et al.* (2014) Principles and methods of integrative genomic analyses in cancer, *Nature reviews. Cancer*, **14**, 299-313.
Oltean, S. and Bates, D.O. (2014) Hallmarks of alternative splicing in cancer, *Oncogene*, **33**, 5311-5318.
Parrella, P. (2010) Epigenetic Signatures in Breast Cancer: Clinical Perspective, *Breast care*, **5**, 66-73.
R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Ray, B.*, et al.* (2014) Information content and analysis methods for multi-modal high-throughput biomedical data, *Scientific reports*, **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 research*, **72**, 1170-1181.
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.

Sengupta, N.*, et al.* (2013) Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX1 deletion, *Molecular cancer*, **12**, 1.
Shen, R.L., Olshen, A.B. and Ladanyi, M. (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.
Sulakhe, D.*, et al.* (2014) High-throughput translational medicine: challenges and solutions, *Advances in experimental medicine and biology*, **799**, 39-67.
Świtnicki, M.P.*, et al.* (2015) PINCAGE: Probabilistic integration of cancer genomics data for perturbed gene identification and sample classification.
Tibshirani, R. (2011) Regression shrinkage and selection via the lasso: a retrospective, *J R Stat Soc B*, **73**, 273-282.
Van Den Eynde, B.J.*, et al.* (1999) A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription, *The Journal of experimental medicine*, **190**, 1793-1800.
Vaske, C.J.*, et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, *Bioinformatics*, **26**, i237-245.
Venables, W.N., Ripley, B.D. and Venables, W.N. (2002) *Modern applied statistics with S*. Statistics and computing. Springer, New York.
Wang, W.*, et al.* (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data, *Bioinformatics*, **29**, 149-159.
Weiss, R. (2005) NIH Launches Cancer Genome Project. *Washington Post*.
Wyatt, A.W.*, et al.* (2014) Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer, *Genome biology*, **15**, 426.
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 : the journal of biological databases and curation*, **2011**, bar026.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net, *J R Stat Soc B*, **67**, 301-320.

Supplement

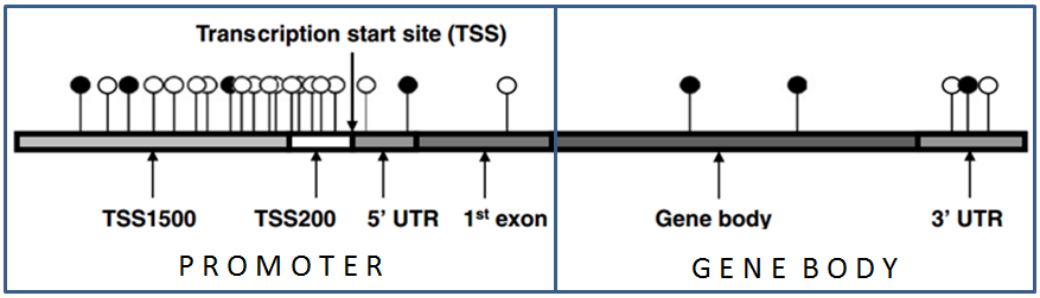


Fig. S1 Division of 450k platform probe annotations between 2 functional groups. Adapted and reprinted from (Bibikova, et al., 2011) with permission from Elsevier.

Table S1 8-fold cross validation table with top-14 ranks at each fold of the 82 adjacent normal vs 730 tumor samples BRCA analysis. All top-10 genes according to the mean rank reappear in every fold within the first 14 ranks (in bold).

Top	Fold							
	1	2	3	4	5	6	7	8
1	**TMEM132D**	**KIR3DX1**	**KIR3DX1**	**TMEM132D**	**TMEM132D**	**TMEM132D**	**TMEM132C**	**TMEM132D**
2	**TMEM132C**	**TMEM132D**	**TMEM132C**	**KIR3DX1**	**TMEM132C**	**TMEM132C**	**TMEM132D**	**TMEM132C**
3	**ULBP1**	**TMEM132C**	**TMEM132D**	**TMEM132C**	**ULBP1**	**FXYD1**	**ULBP1**	**ULBP1**
4	**SLC6A2**	**ULBP1**	**ULBP1**	**ULBP1**	**CACNG3**	**ULBP1**	**CACNG3**	**FXYD1**
5	**KIR3DX1**	**LOC388692**	**CACNG3**	**CACNG3**	**KIR3DX1**	**LOC388692**	A2BP1	**KIR3DX1**
6	**CACNG3**	**CACNG3**	**LOC388692**	**LOC388692**	**LOC388692**	**CACNG3**	**LOC388692**	**LOC388692**
7	LHFPL3	**FXYD1**	**FXYD1**	**FXYD1**	**FXYD1**	MRGPRF	DPP6	**CACNG3**
8	**LOC388692**	LHFPL3	**SLC6A2**	**SLC6A2**	**NRSN1**	**KIR3DX1**	**FXYD1**	TSSK6
9	**A2BP1**	**SLC6A2**	CPVL	**NRSN1**	SEMA4A	CPVL	**NRSN1**	CPVL
10	DPP6	HS3ST2	**NRSN1**	CPVL	CPVL	**SLC6A2**	**KIR3DX1**	**SLC6A2**
11	**FXYD1**	AIM2	**A2BP1**	SEMA4A	**SLC6A2**	**A2BP1**	**SLC6A2**	**NRSN1**
12	CPVL	**NRSN1**	AIM2	**A2BP1**	**A2BP1**	**NRSN1**	AIM2	**A2BP1**
13	**NRSN1**	CPVL	SNCAIP	SNCAIP	LOC285370	LOC134466	LOC134466	SEMA4A
14	NEUROD1	**A2BP1**	NEUROD1	TSSK6	TSSK6	SEMA4A	NEUROD1	AIM2

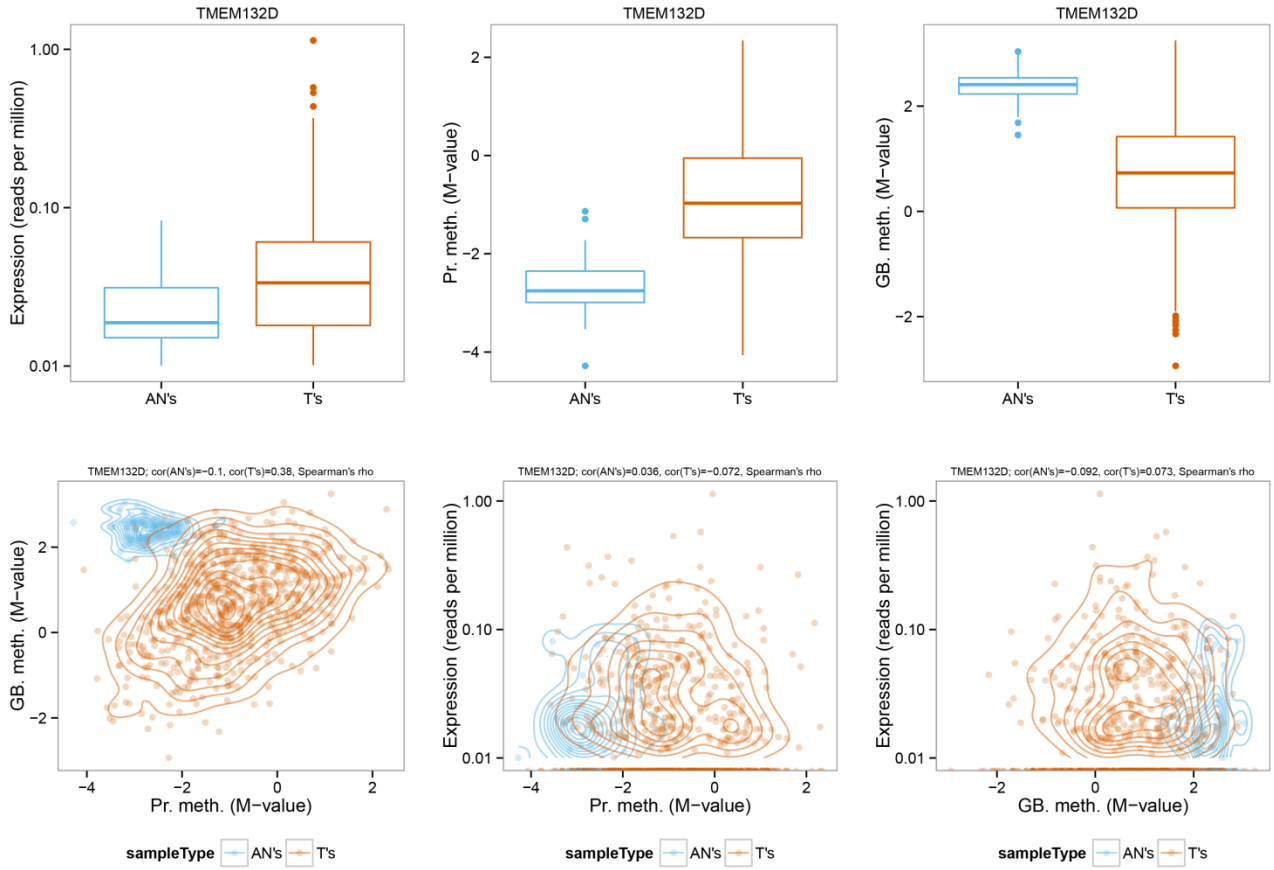
Table S2 Progressed and non-progressed tumours defined based on presence or absence of recurrence within close to 3 years (1065 days) of initial treatment.

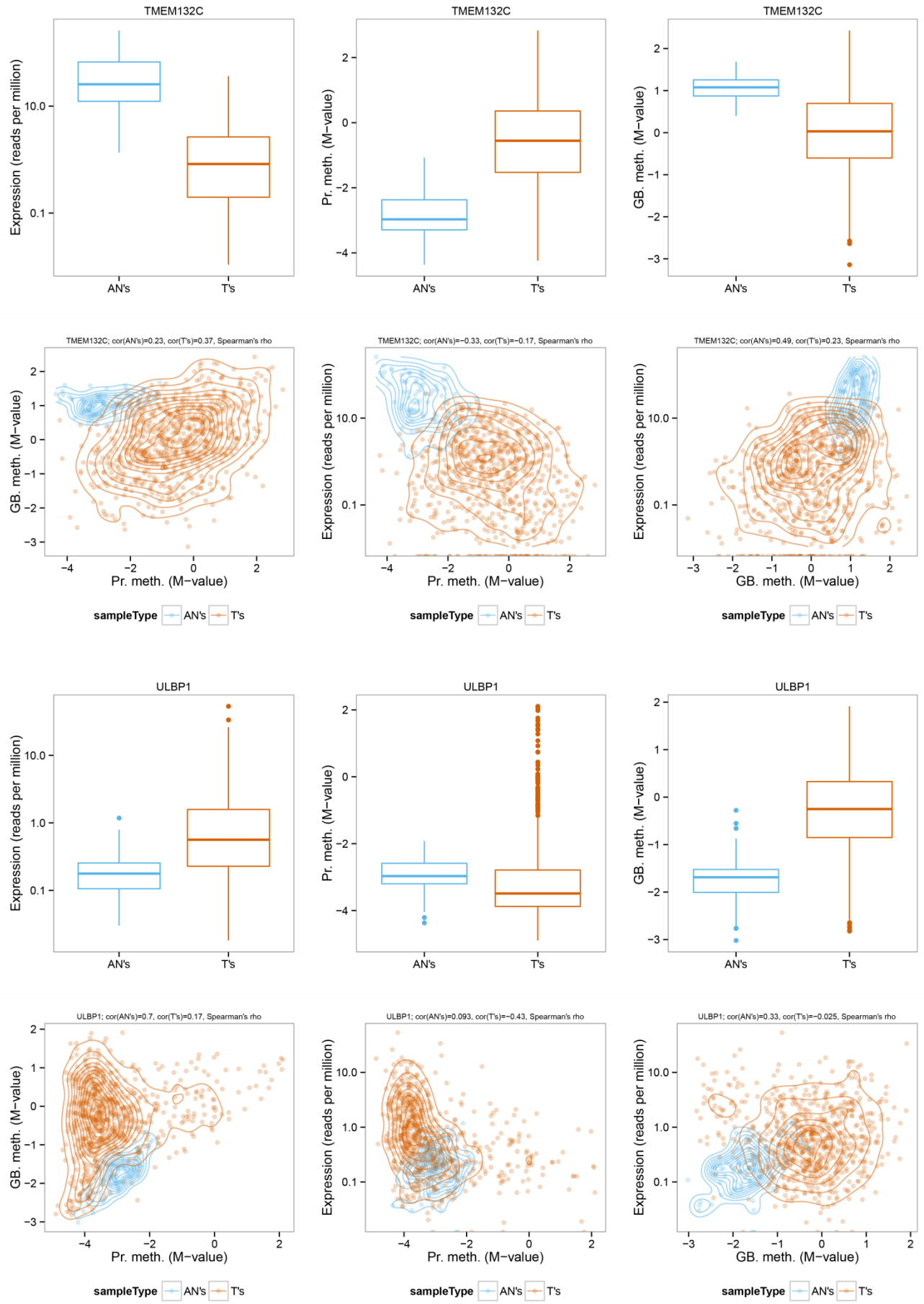
Sample type	**List with TCGA patient IDs**
Progressed disease n=14	TCGA-A7-A3RF TCGA-A7-A425 TCGA-LL-A5YM TCGA-E9-A243 TCGA-A7-A13G TCGA-A7-A26H TCGA-LQ-A4E4 TCGA-A7-A13H TCGA-A8-A08O TCGA-E9-A226 TCGA-A2-A3XY TCGA-E9-A2JS TCGA-A2-A3XU TCGA-AR-A5QQ
Non-progressed disease n=57	TCGA-A7-A0CE TCGA-A7-A0CH TCGA-E9-A1RI TCGA-E9-A1NE TCGA-OL-A5RW TCGA-E9-A1NA TCGA-E9-A1N5 TCGA-A7-A0D9 TCGA-AR-A1AS TCGA-AR-A2LN TCGA-GM-A3NY TCGA-A2-A3Y0 TCGA-E9-A22A TCGA-AR-A2LO TCGA-E9-A1NC TCGA-A2-A3KD TCGA-AR-A2LQ TCGA-AC-A2FB TCGA-GM-A3XG TCGA-A2-A0YL TCGA-A2-A3XW TCGA-BH-A0HY TCGA-EW-A2FS TCGA-EW-A1P3 TCGA-BH-A0HA TCGA-EW-A2FR TCGA-AR-A255 TCGA-AR-A1AV TCGA-AR-A2LJ TCGA-AR-A1AX TCGA-AR-A1AM TCGA-AR-A2LJ TCGA-AR-A1AW TCGA-OL-A66J TCGA-GM-A3XN TCGA-GM-A3XL TCGA-GM-A4E0 TCGA-AR-A254 TCGA-AR-A252 TCGA-AR-A24T TCGA-AR-A1AU TCGA-AR-A251 TCGA-AR-A24N TCGA-AR-A24Z TCGA-A2-A3XT TCGA-AR-A24X TCGA-AR-A24V TCGA-B6-A401 TCGA-AR-A0U4 TCGA-AR-A0TT TCGA-AR-A24R TCGA-AR-A24M TCGA-AR-A0TW TCGA-AR-A24Q TCGA-B6-A40B TCGA-A2-A0EP TCGA-A2-A0CR TCGA-GM-A3NW TCGA-AR-A0TP TCGA-AR-A0U3 TCGA-AQ-A04L

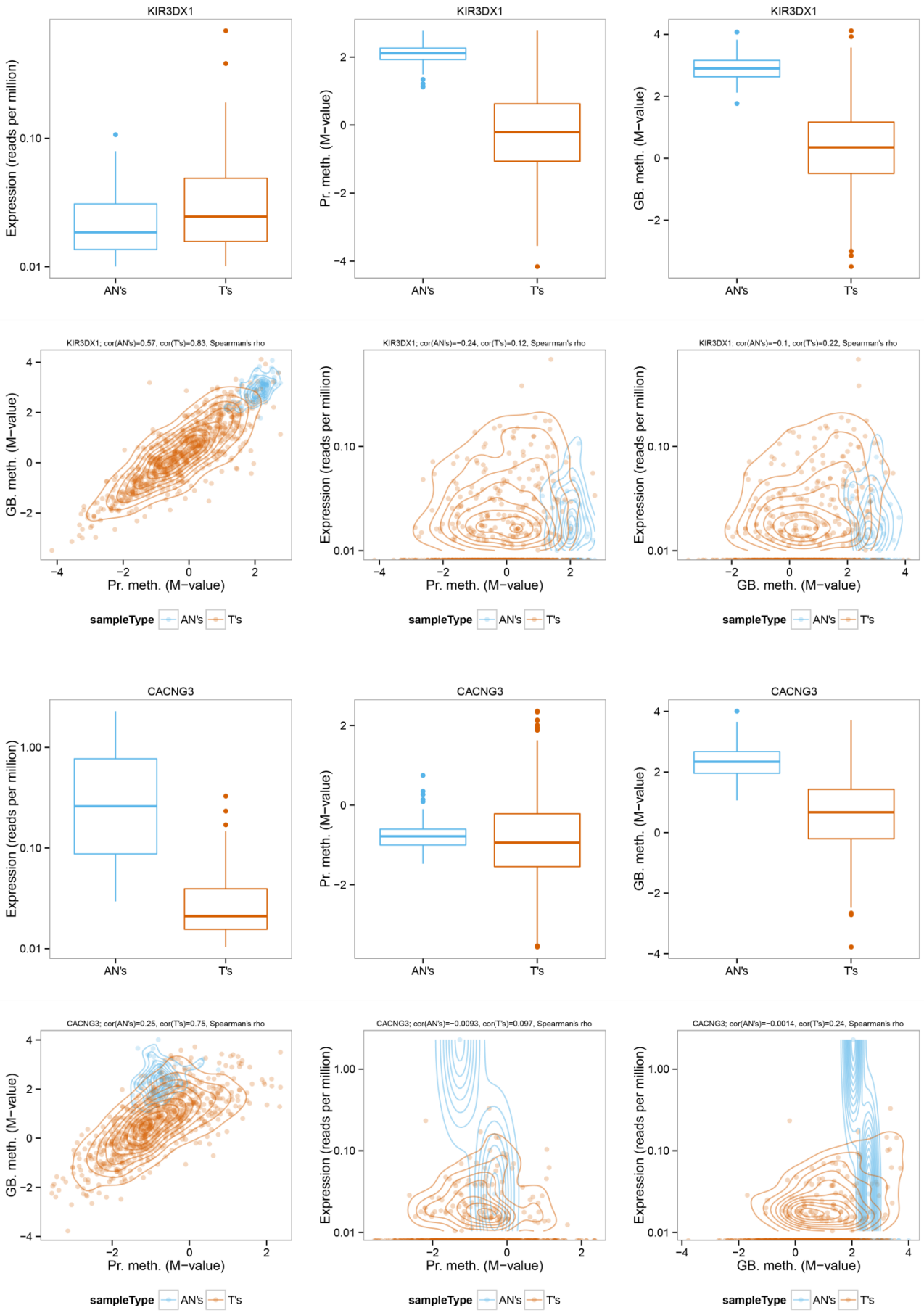
Table S3 14-fold cross validation table with top-20 ranks at each fold. Top-3 genes according to the mean rank reappear in every fold within the first 20 ranks.

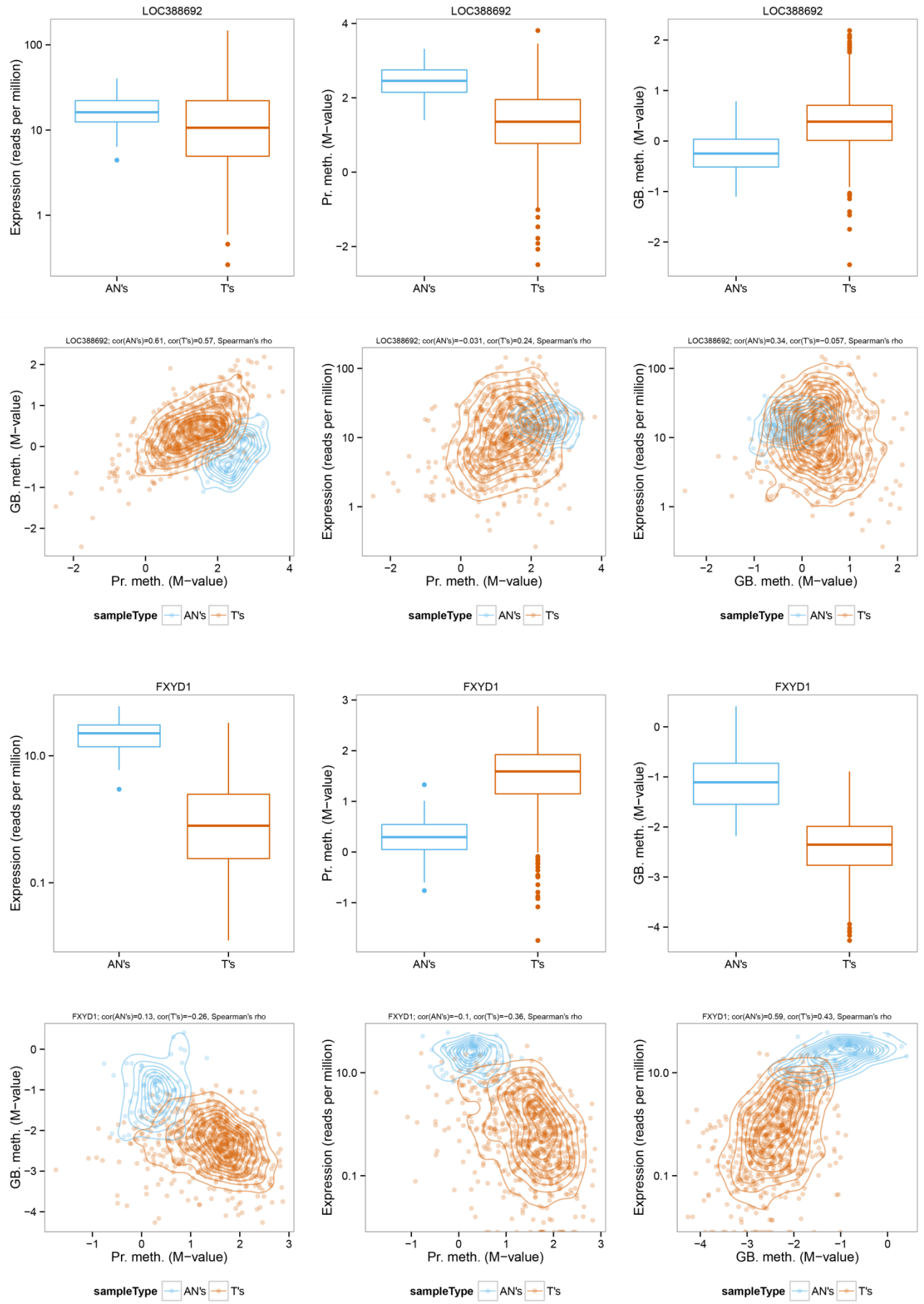
Top	Fold													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	SFRS8	TMEM198	**ZFATAS**	**KAAG1**	**ZFATAS**	**ZFATAS**	**SERPINE3**	**ZFATAS**	**ZFATAS**	**SERPINE3**	EEF1DP3	**SERPINE3**	**SERPINE3**	**ZFATAS**
2	**KAAG1**	ATP9A	DPY19L3	IQGAP2	**KAAG1**	GPRIN3	**ZFATAS**	**SERPINE3**	**SERPINE3**	PDE1A	SFRS8	EEF1DP3	**ZFATAS**	ATP9A
3	FAM27L	**ZFATAS**	SFRS8	SFRS8	WNT7B	SFRS8	**KAAG1**	BTBD2	**KAAG1**	**ZFATAS**	**ZFATAS**	DPY19L3	**KAAG1**	CYCSP52
4	**ZFATAS**	LOC149620	**SERPINE3**	DPY19L3	**SERPINE3**	LOC149620	ATP9A	DPY19L3	CRYBG3	PLCZ1	AIMP1	**ZFATAS**	MYLIP	DKK4
5	IL17C	DPY19L3	WNT7B	ATP9A	CRYBG3	**SERPINE3**	UPP1	EGLN2	IQGAP2	ATP9A	ATAD1	IFIH1	SEMA4C	**KAAG1**
6	GPI	**KAAG1**	ENPP6	NPB	NECAB2	NPB	CCDC73	**KAAG1**	SFRS8	TNIP3	**SERPINE3**	KCNA2	CD2	**SERPINE3**
7	**SERPINE3**	NF1	GPBAR1	**ZFATAS**	GPR171	**KAAG1**	GPBAR1	SFRS8	FAM27L	IQGAP2	TMEM198	**KAAG1**	SLC26A7	SFRS8
8	RHOH	SFRS8	LOC100128023	GPBAR1	QDPR	ATP9A	SLC28A2	PDE1A	EEF1DP3	SFRS8	BTBD2	SFRS8	GPBAR1	GPBAR1
9	BGLAP	LOC100128023	CRYBG3	RNASEN	ATP9A	DCAF10	SFRS8	FAM27L	WNT7B	LPAR4	TCF15	ARID4A	KIAA1683	CD2
10	CRYBG3	NPB	FAM27L	STC1	NPB	IQGAP2	GPRIN3	NF1	DPY19L3	SLC22A9	**KAAG1**	IQGAP2	UACA	CRYBG3
11	TMEM198	PRNT	LOC149620	KIAA1377	SNAR-I	BTBD2	TMEM198	RNASEN	KIAA1377	CRB1	NF1	TCF21	LOC149620	CRB1
12	CHRND	**SERPINE3**	PTPRJ	ZNF706	RCAN2	WNT7B	LOC149620	TMEM198	CTPS2	**KAAG1**	MPHOSPH9	TMEM198	DPY19L3	DPY19L3
13	GPBAR1	DKK4	STC1	ATG2B	RTP2	NECAB2	DPY19L3	LOC149620	CYCSP52	UBE2E2	LOC149620	ODZ3	PDE1A	NF1
14	DPY19L3	QDPR	IQGAP2	**SERPINE3**	LYG2	KIAA1377	FAM155B	CYCSP52	BTBD2	LOC100128023	GPBAR1	ENPP6	WNT7B	LOC149620
..														
20	LOC149620	BTBD2	**KAAG1**	ATAD1	KIAA1683	CCDC73	ENPP6	ANTXR1	DCAF10	LOC149620	CRYBG3	CYCSP52	EEF1DP3	IQGAP2

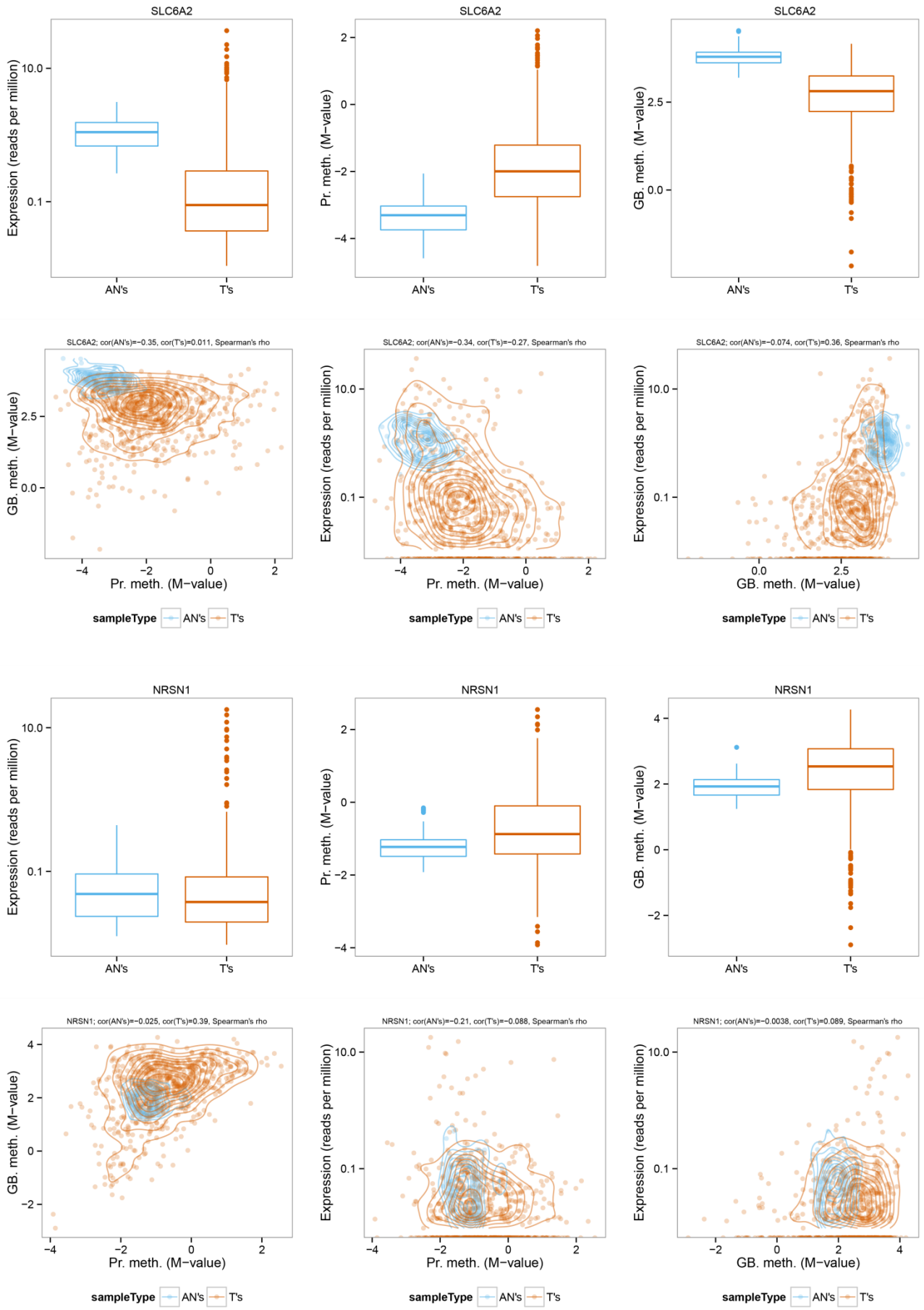
Fig. S2 Marginal and pairwise distributions of gene expression, promoter methylation, and gene body methylation for the top-10 genes identified by integrative PINCAGE in the comparison between tumour and adjacent normal samples. For each gene **Top rows**: Marginal distributions of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value across BRCA Tumour (T) and Adjacent Normal (AN) samples. For each gene **Bottom rows**: 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).




TMEM132C
Expression (reads per million)
10.0
0.1
AN's
T's
TMEM132C
Pr. meth. (M−value)
2
0
−2
−4
AN's
T's
TMEM132C
GB. meth. (M−value)
2
1
0
−1
−2
−3
AN's
T's
TMEM132C; cor(AN's)=0.23, cor(T's)=0.37, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType AN's T's
TMEM132C; cor(AN's)=−0.33, cor(T's)=−0.17, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType AN's T's
TMEM132C; cor(AN's)=0.49, cor(T's)=0.23, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType AN's T's
ULBP1
Expression (reads per million)
10.0
1.0
0.1
AN's
T's
ULBP1
Pr. meth. (M−value)
AN's
T's
ULBP1
GB. meth. (M−value)
AN's
T's
ULBP1; cor(AN's)=0.7, cor(T's)=0.17, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType AN's T's
ULBP1; cor(AN's)=0.093, cor(T's)=−0.43, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType AN's T's
ULBP1; cor(AN's)=0.33, cor(T's)=−0.025, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType AN's T's


KIR3DX1
Expression (reads per million)
0.10
0.01
AN's
T's
KIR3DX1
Pr. meth. (M−value)
2
0
−2
−4
AN's
T's
KIR3DX1
GB. meth. (M−value)
4
2
0
−2
AN's
T's
KIR3DX1; cor(AN's)=0.57, cor(T's)=0.83, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType AN's T's
KIR3DX1; cor(AN's)=−0.24, cor(T's)=0.12, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType AN's T's
KIR3DX1; cor(AN's)=−0.1, cor(T's)=0.22, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType AN's T's
CACNG3
Expression (reads per million)
1.00
0.10
0.01
AN's
T's
CACNG3
Pr. meth. (M−value)
AN's
T's
CACNG3
GB. meth. (M−value)
AN's
T's
CACNG3; cor(AN's)=0.25, cor(T's)=0.75, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType AN's T's
CACNG3; cor(AN's)=−0.0093, cor(T's)=0.097, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType AN's T's
CACNG3; cor(AN's)=−0.0014, cor(T's)=0.24, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType AN's T's


LOC388692
Expression (reads per million)
Pr. meth. (M−value)
GB. meth. (M−value)
AN's
T's
LOC388692; cor(AN's)=0.61, cor(T's)=0.57, Spearman's rho
LOC388692; cor(AN's)=−0.031, cor(T's)=0.24, Spearman's rho
LOC388692; cor(AN's)=0.34, cor(T's)=−0.057, Spearman's rho
sampleType
FXYD1
FXYD1; cor(AN's)=0.13, cor(T's)=−0.26, Spearman's rho
FXYD1; cor(AN's)=−0.1, cor(T's)=−0.36, Spearman's rho
FXYD1; cor(AN's)=0.59, cor(T's)=0.43, Spearman's rho


SLC6A2
Expression (reads per million)
10.0
0.1
AN's
T's
SLC6A2
Pr. meth. (M−value)
2
0
−2
−4
AN's
T's
SLC6A2
GB. meth. (M−value)
2.5
0.0
AN's
T's
SLC6A2; cor(AN's)=−0.35, cor(T's)=0.011, Spearman's rho
GB. meth. (M−value)
2.5
0.0
−4
−2
0
2
Pr. meth. (M−value)
sampleType
AN's
T's
SLC6A2; cor(AN's)=−0.34, cor(T's)=−0.27, Spearman's rho
Expression (reads per million)
10.0
0.1
−4
−2
0
2
Pr. meth. (M−value)
sampleType
AN's
T's
SLC6A2; cor(AN's)=−0.074, cor(T's)=0.36, Spearman's rho
Expression (reads per million)
10.0
0.1
0.0
2.5
GB. meth. (M−value)
sampleType
AN's
T's
NRSN1
Expression (reads per million)
10.0
0.1
AN's
T's
NRSN1
Pr. meth. (M−value)
2
0
−2
−4
AN's
T's
NRSN1
GB. meth. (M−value)
4
2
0
−2
AN's
T's
NRSN1; cor(AN's)=−0.025, cor(T's)=0.39, Spearman's rho
GB. meth. (M−value)
4
2
0
−2
−4
−2
0
2
Pr. meth. (M−value)
sampleType
AN's
T's
NRSN1; cor(AN's)=−0.21, cor(T's)=−0.088, Spearman's rho
Expression (reads per million)
10.0
0.1
−4
−2
0
2
Pr. meth. (M−value)
sampleType
AN's
T's
NRSN1; cor(AN's)=−0.0038, cor(T's)=0.089, Spearman's rho
Expression (reads per million)
10.0
0.1
−2
0
2
4
GB. meth. (M−value)
sampleType
AN's
T's

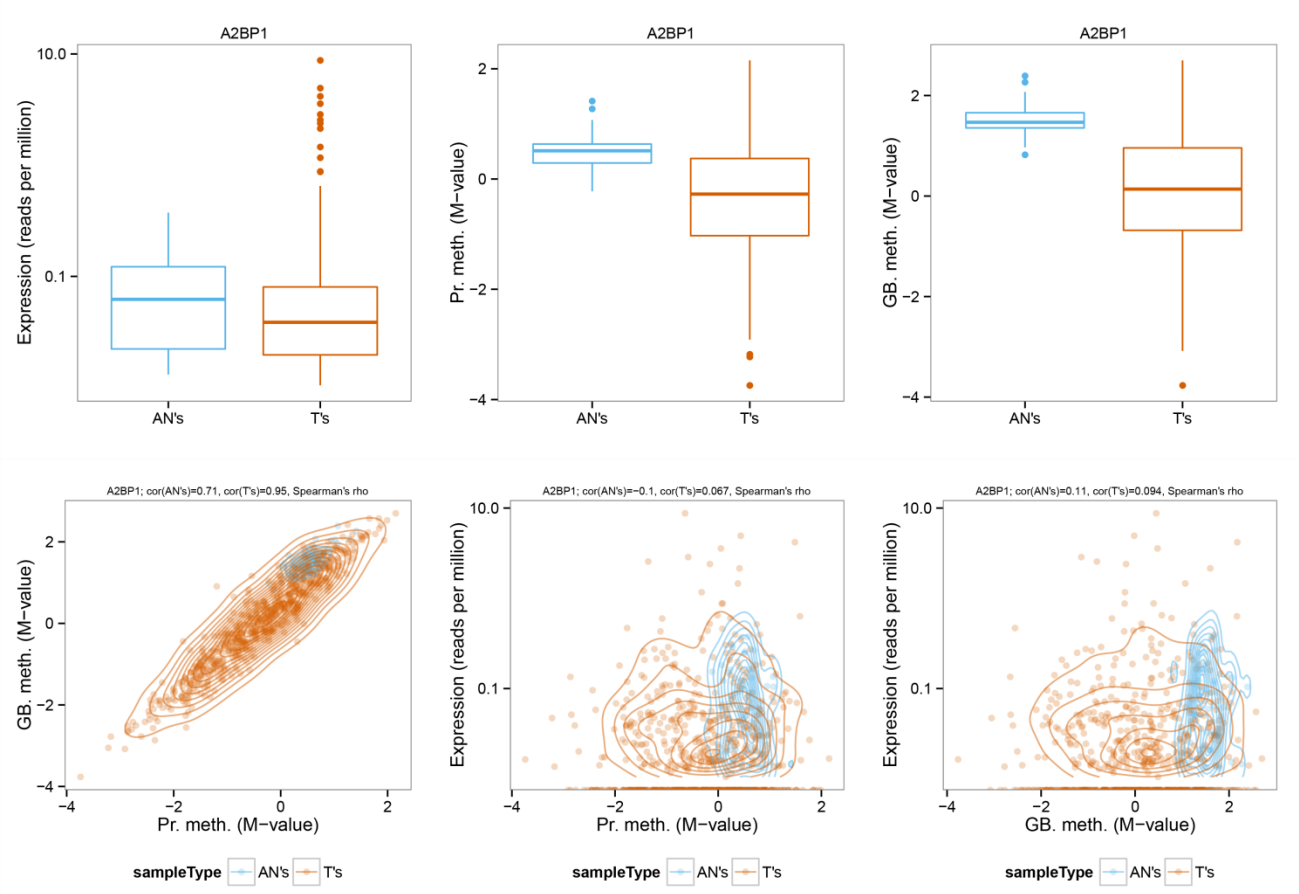
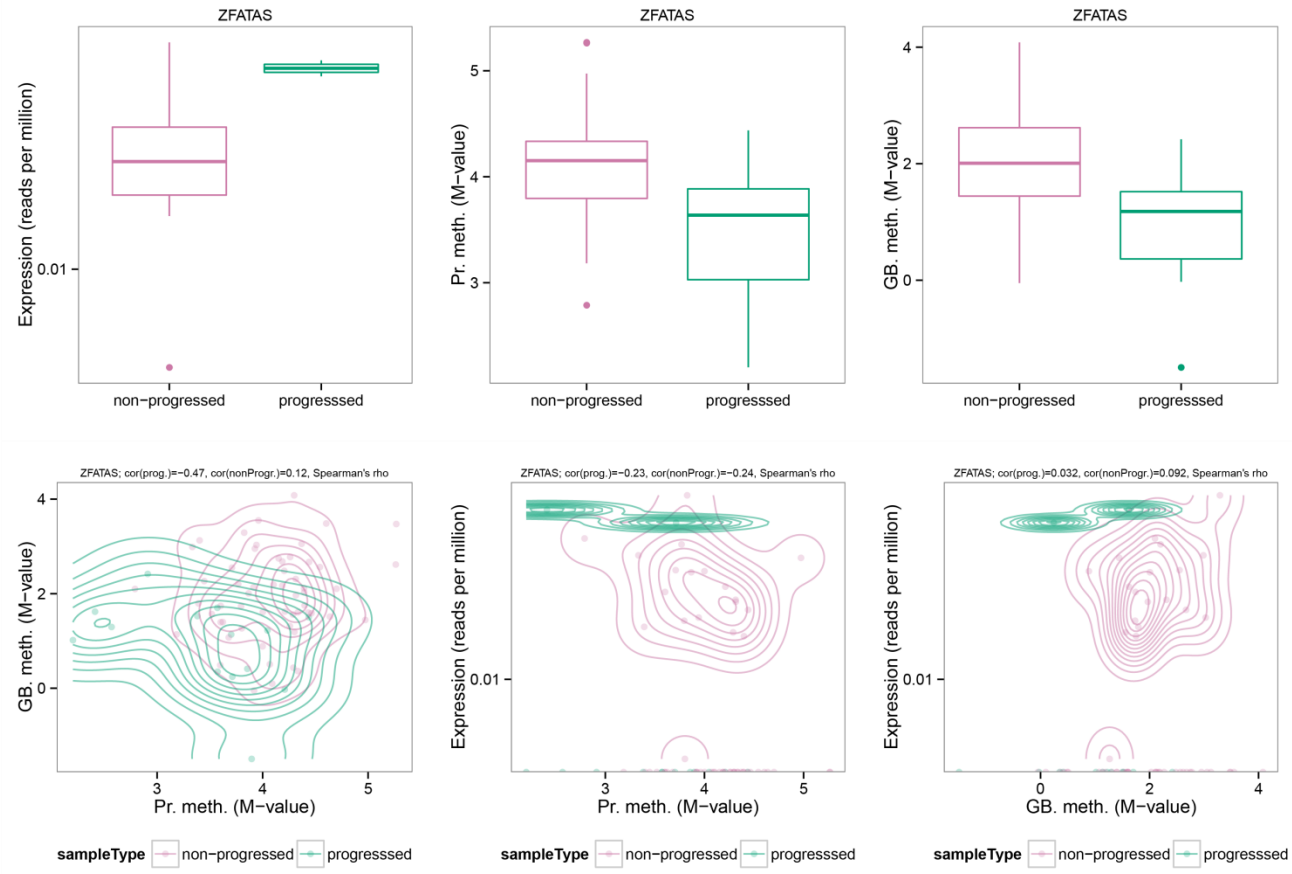
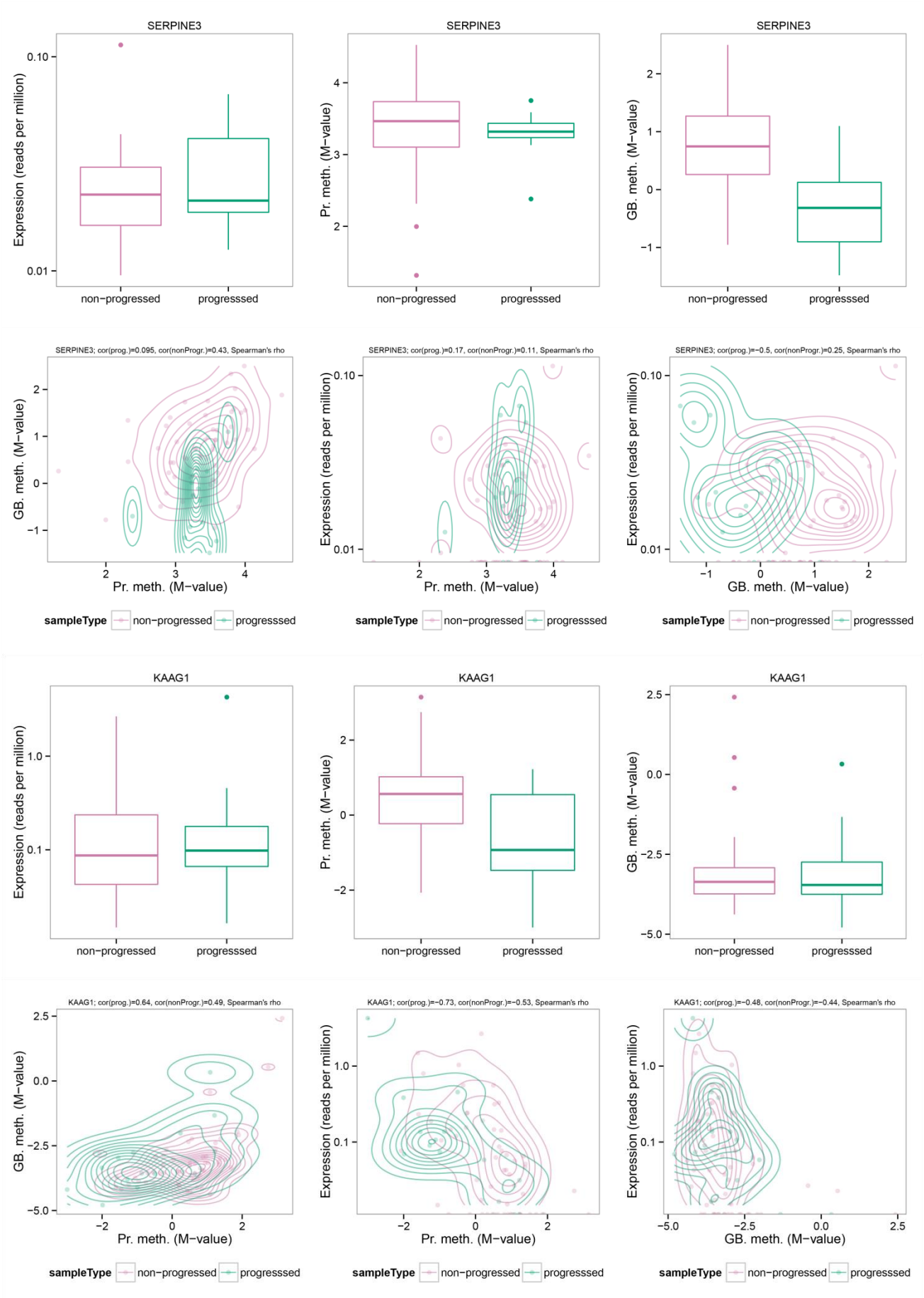
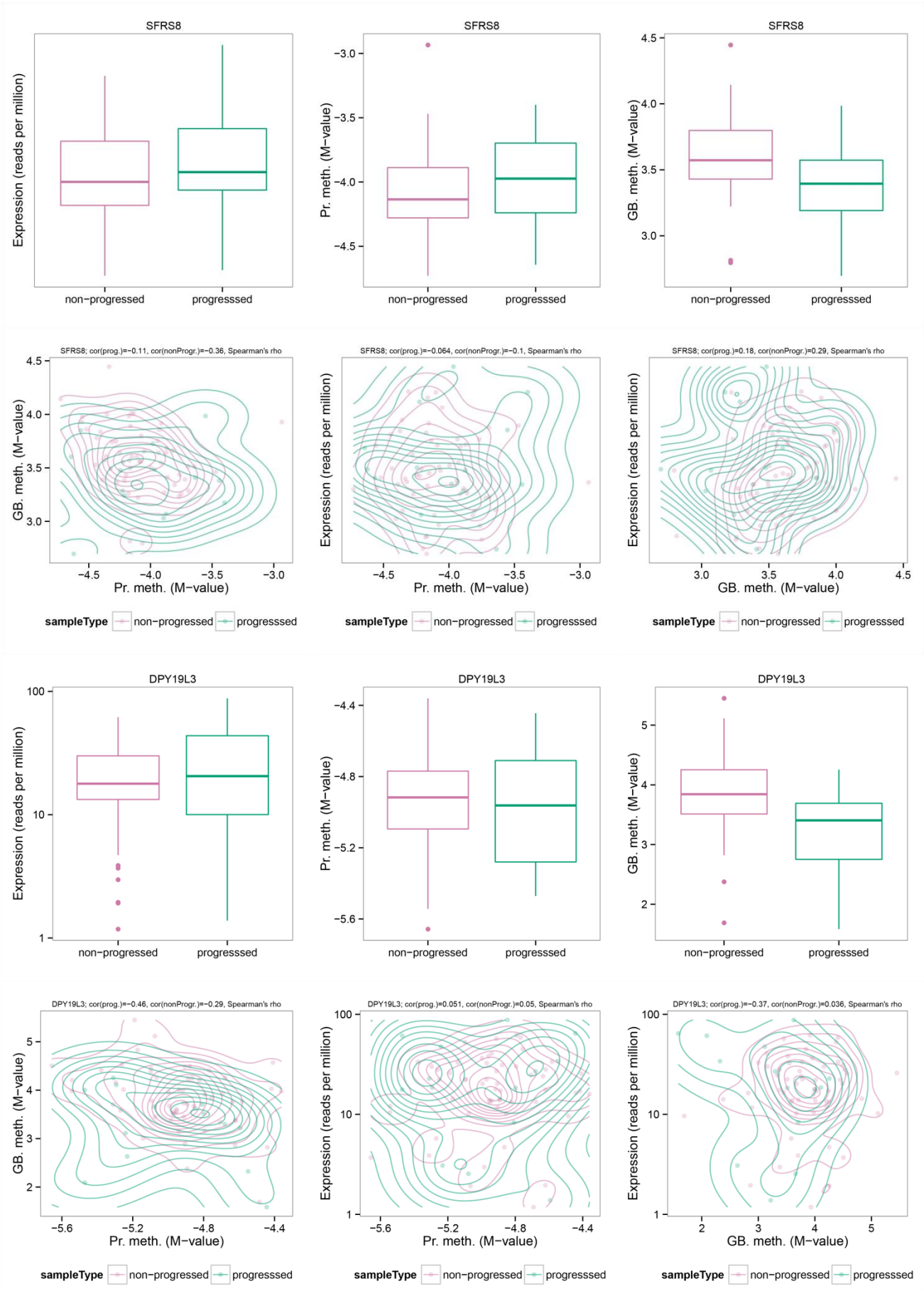
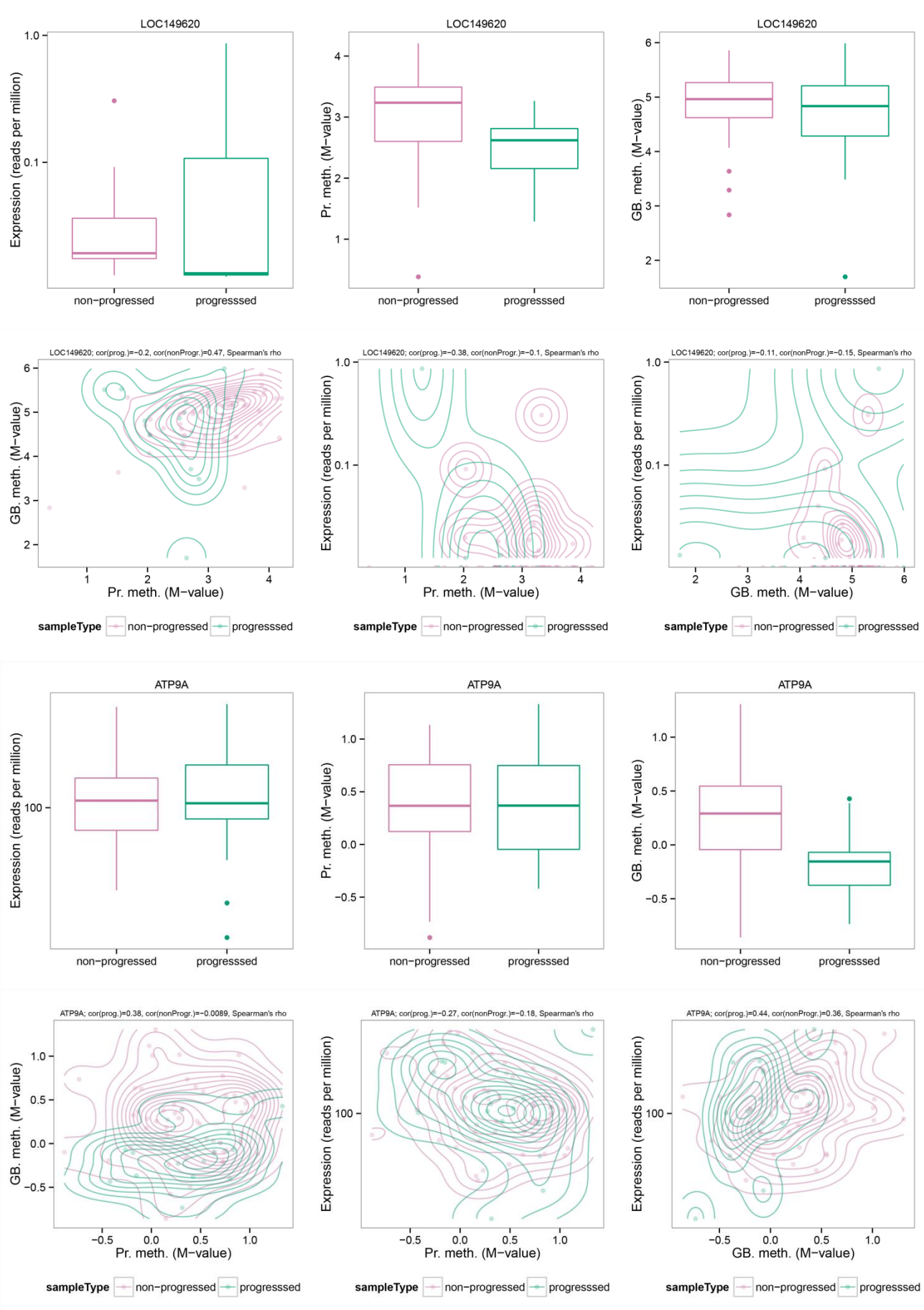

A2BP1
Expression (reads per million)
10.0
0.1
AN's
T's
A2BP1
Pr. meth. (M−value)
2
0
−2
−4
AN's
T's
A2BP1
GB. meth. (M−value)
2
0
−2
−4
AN's
T's
A2BP1; cor(AN's)=0.71, cor(T's)=0.95, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType AN's T's
A2BP1; cor(AN's)=−0.1, cor(T's)=0.067, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType AN's T's
A2BP1; cor(AN's)=0.11, cor(T's)=0.094, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType AN's T's

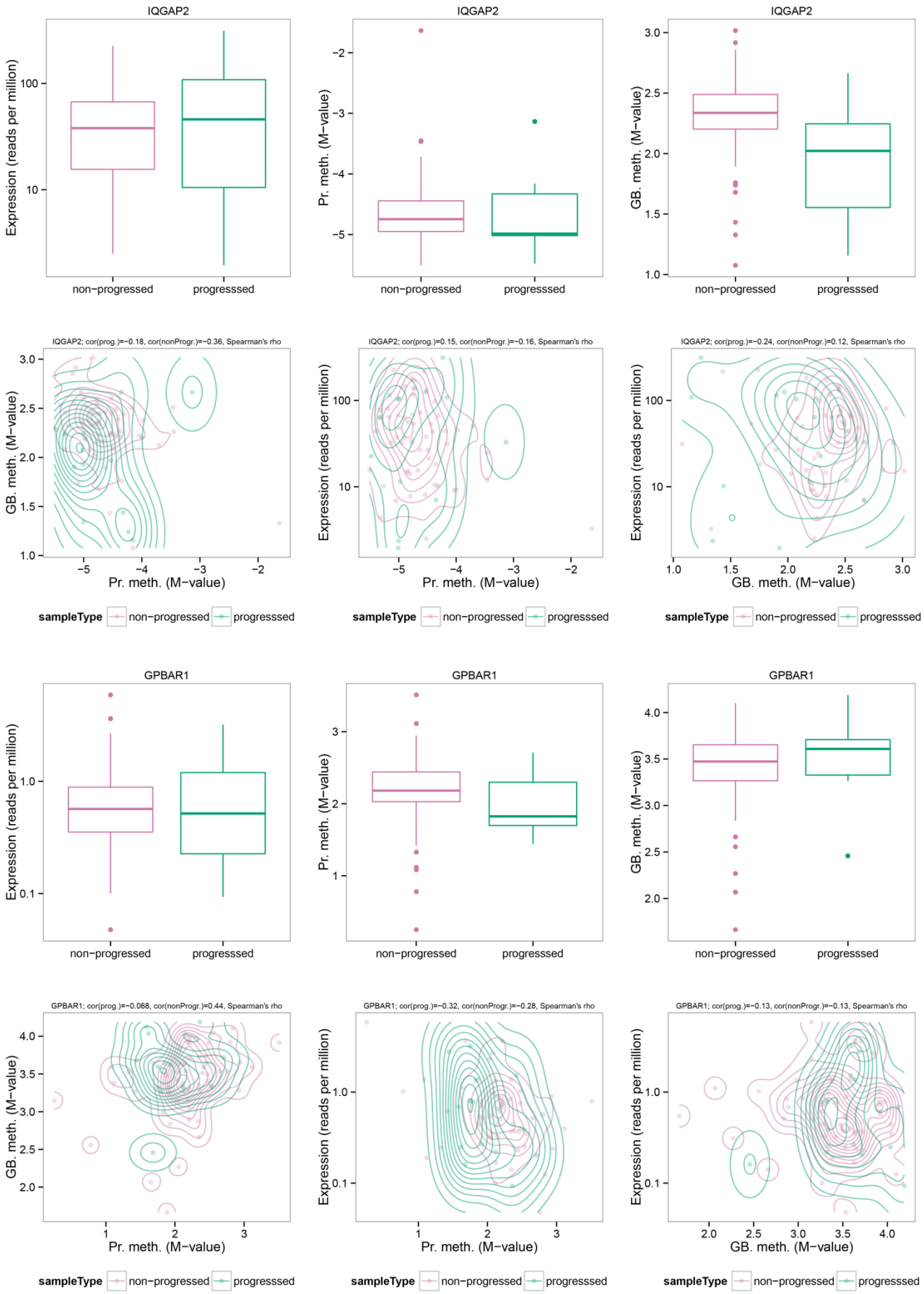
Fig. S3 Marginal and pairwise distributions of gene expression, promoter methylation, and gene body methylation for the top-10 genes identified by integrative sparse model in the comparison between progressing and non-progressing BRCA tumour samples. For each gene **Top rows**: Marginal distributions of gene expression in terms of reads per million (RPM) and promoter and gene body methylation in terms of M-value of BRCA progressed and non-progressed samples. For each gene **Bottom rows**: Pairwise distributions of the three data types. Normal-reference-based kernel density contours (Venables, et al., 2002) shown for both progressed (green) and non-progressed samples (violet).

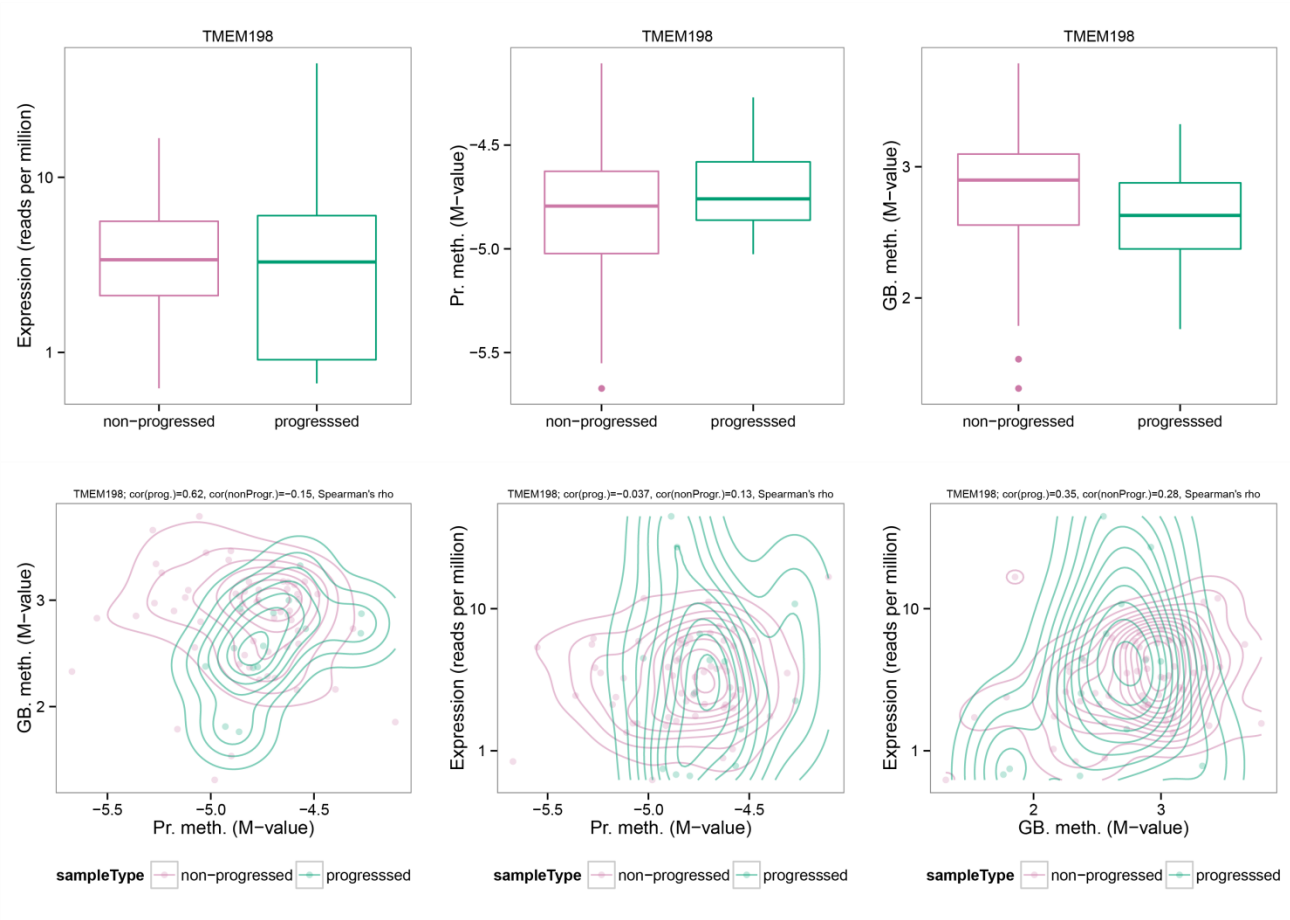



SERPINE3
Expression (reads per million)
0.10
0.01
non−progressed
progresssed
SERPINE3
Pr. meth. (M−value)
4
3
2
non−progressed
progresssed
SERPINE3
GB. meth. (M−value)
2
1
0
−1
non−progressed
progresssed
SERPINE3; cor(prog.)=0.095, cor(nonProgr.)=0.43, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
SERPINE3; cor(prog.)=0.17, cor(nonProgr.)=0.11, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
SERPINE3; cor(prog.)=−0.5, cor(nonProgr.)=0.25, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType
non−progressed
progresssed
KAAG1
Expression (reads per million)
1.0
0.1
non−progressed
progresssed
KAAG1
Pr. meth. (M−value)
non−progressed
progresssed
KAAG1
GB. meth. (M−value)
2.5
0.0
−2.5
−5.0
non−progressed
progresssed
KAAG1; cor(prog.)=0.64, cor(nonProgr.)=0.49, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
KAAG1; cor(prog.)=−0.73, cor(nonProgr.)=−0.53, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
KAAG1; cor(prog.)=−0.48, cor(nonProgr.)=−0.44, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType
non−progressed
progresssed


SFRS8
Expression (reads per million)
non−progressed
progresssed
SFRS8
Pr. meth. (M−value)
non−progressed
progresssed
SFRS8
GB. meth. (M−value)
non−progressed
progresssed
SFRS8; cor(prog.)=−0.11, cor(nonProgr.)=−0.36, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType non−progressed progresssed
SFRS8; cor(prog.)=−0.064, cor(nonProgr.)=−0.1, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType non−progressed progresssed
SFRS8; cor(prog.)=0.18, cor(nonProgr.)=0.29, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType non−progressed progresssed
DPY19L3
Expression (reads per million)
non−progressed
progresssed
DPY19L3
Pr. meth. (M−value)
non−progressed
progresssed
DPY19L3
GB. meth. (M−value)
non−progressed
progresssed
DPY19L3; cor(prog.)=−0.46, cor(nonProgr.)=−0.29, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
sampleType non−progressed progresssed
DPY19L3; cor(prog.)=0.051, cor(nonProgr.)=0.05, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
sampleType non−progressed progresssed
DPY19L3; cor(prog.)=−0.37, cor(nonProgr.)=0.036, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType non−progressed progresssed


LOC149620
Expression (reads per million)
Pr. meth. (M−value)
GB. meth. (M−value)
non−progressed
progresssed
LOC149620; cor(prog.)=−0.2, cor(nonProgr.)=0.47, Spearman's rho
LOC149620; cor(prog.)=−0.38, cor(nonProgr.)=−0.1, Spearman's rho
LOC149620; cor(prog.)=−0.11, cor(nonProgr.)=−0.15, Spearman's rho
sampleType
ATP9A
ATP9A; cor(prog.)=0.38, cor(nonProgr.)=−0.0089, Spearman's rho
ATP9A; cor(prog.)=−0.27, cor(nonProgr.)=−0.18, Spearman's rho
ATP9A; cor(prog.)=0.44, cor(nonProgr.)=0.36, Spearman's rho


IQGAP2
Expression (reads per million)
100
10
non−progressed
progresssed
IQGAP2
Pr. meth. (M−value)
−2
−3
−4
−5
non−progressed
progresssed
IQGAP2
GB. meth. (M−value)
3.0
2.5
2.0
1.5
1.0
non−progressed
progresssed
IQGAP2; cor(prog.)=−0.18, cor(nonProgr.)=−0.36, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
IQGAP2; cor(prog.)=0.15, cor(nonProgr.)=−0.16, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
IQGAP2; cor(prog.)=−0.24, cor(nonProgr.)=0.12, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType
non−progressed
progresssed
GPBAR1
Expression (reads per million)
1.0
0.1
non−progressed
progresssed
GPBAR1
Pr. meth. (M−value)
3
2
1
non−progressed
progresssed
GPBAR1
GB. meth. (M−value)
4.0
3.5
3.0
2.5
2.0
non−progressed
progresssed
GPBAR1; cor(prog.)=−0.068, cor(nonProgr.)=0.44, Spearman's rho
GB. meth. (M−value)
Pr. meth. (M−value)
GPBAR1; cor(prog.)=−0.32, cor(nonProgr.)=−0.28, Spearman's rho
Expression (reads per million)
Pr. meth. (M−value)
GPBAR1; cor(prog.)=−0.13, cor(nonProgr.)=−0.13, Spearman's rho
Expression (reads per million)
GB. meth. (M−value)
sampleType
non−progressed
progresssed


TMEM198
Expression (reads per million)
10
1
non−progressed
progresssed
TMEM198
Pr. meth. (M−value)
−4.5
−5.0
−5.5
non−progressed
progresssed
TMEM198
GB. meth. (M−value)
3
2
non−progressed
progresssed
TMEM198; cor(prog.)=0.62, cor(nonProgr.)=−0.15, Spearman's rho
GB. meth. (M−value)
3
2
−5.5
−5.0
−4.5
Pr. meth. (M−value)
sampleType
non−progressed
progresssed
TMEM198; cor(prog.)=−0.037, cor(nonProgr.)=0.13, Spearman's rho
Expression (reads per million)
10
1
−5.5
−5.0
−4.5
Pr. meth. (M−value)
sampleType
non−progressed
progresssed
TMEM198; cor(prog.)=0.35, cor(nonProgr.)=0.28, Spearman's rho
Expression (reads per million)
10
1
2
3
GB. meth. (M−value)
sampleType
non−progressed
progresssed

Chapter 10: Manuscript 3

ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction

Manuscript in review at *Bioinformatics*

Sudhakar Sahoo[1,] Michał P. Świtnicki[1], Jakob S. Pedersen[1,2]

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Brendstrupgårdsvej 21, 8200 Aarhus, Denmark and [2]Bioinformatics Research Centre (BiRC), Aarhus University, C.F.Møllers Allé 8, 8000 Aarhus, Denmark

Running head: Probabilistic integration of probing data

Doe-Star BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–8

ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction

Sudhakar Sahoo[1], Michał P. Świtnicki[1], and Jakob Skou Pedersen[1,2*]

(1) Department of Molecular Medicine (MOMA), Aarhus University Hospital, Brendstrupgårdsvej 100, 8200 Aarhus N (2) Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, DK-8000 Aarhus C, Denmark

Received ; Revised ; Accepted

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Recently, new RNA secondary structure probing techniques have been developed, including Next Generation Sequencing (NGS) based methods capable of probing transcriptome-wide. These techniques hold great promise for improving structure prediction accuracy. However, each new data type comes with its own signal properties and biases, which may even be experiment specific. There is therefore a growing need for RNA structure prediction methods that can be automatically trained on new data types and readily extended to integrate and fully exploit multiple types of data.

Results: Here we develop and explore a modular probabilistic approach for integrating probing data in RNA structure prediction. It can be automatically trained given a set of known structures with probing data. The approach is demonstrated on SHAPE data sets, where we evaluate and selectively model specific correlations. The approach often makes superior use of the probing data signal compared to other methods. We illustrate the use of ProbFold on multiple data types using both simulations and a small set of structures with both SHAPE, DMS, and CMCT data. Technically, the approach combines stochastic context-free grammars (SCFGs) with probabilistic graphical models. This approach allows rapid adaptation and integration of new probing data types.

Availability: ProbFold is implemented in C++. Models are specified using simple textual formats. Data reformatting is done using separate C++ programs. Source code, statically compiled binaries for x86 Linux machines, C++ programs, example data sets and a tutorial is available from http://moma.ki.au.dk/prj/probfold/.

1 INTRODUCTION

Obtaining accurate secondary structure predictions is a crucial step toward understanding the physical properties of RNA molecules and the biological roles of structural RNA elements. However, computational predictions based only on the primary sequence are often inaccurate and therefore insufficient for in-depth interpretation. Similarly, structure probing data typically only provides partial structural information (Weeks, 2010; Wan *et al*., 2011). Directly modeling both types of data in the folding process generally improves structure prediction accuracy (Mathews *et al*., 2004; Deigan *et al.*, 2009; Swenson *et al*., 2012; Washietl *et al.*, 2012; Sükösd *et al*., 2012; Cordero *et al.*, 2012). While most studies have used methods based on energy-minimization, the inclusion of probing data in probabilistic folding models have not been fully explored (Sükösd *et al*., 2012).

Here we develop and explore a modular probabilistic approach for integrating probing data in RNA secondary structure prediction, which we call ProbFold. The focus is on how to best exploit the structure signal of the probing data, while keeping the models general and easy to adapt to new probing data types or additional layers of probing data. We have not aimed to develop a competitive single-sequence structure prediction method. The main focus is therefore on the use of the probing data signal rather than the overall performance.

Probabilistic modeling offers a coherent framework for combining different types of evidence as they are all naturally measured on the same scale. Structure models based on energy minimization do not extend naturally to additional types of data in the same way. For instance, probing data measurements must be translated to pseudo-energy perturbations before they can be included in the models, though they do not have inherent thermodynamic interpretations.

Probabilistic approaches have previously been used to incorporate comparative evidence in structure folding (Sakakibara *et al*., 1994; Eddy and Durbin, 1994; Knudsen *et al*., 1999; Pedersen *et al*., 2004, 2006; Nawrocki and Eddy, 2013 ; Rivas and Eddy, 2001). This is another example of supplementing the primary sequence with partial structure evidence and as such closely related to the probing data modeling problem studied here. Several of these methods exploit that generative probabilistic methods can be combined and their parameters optimized in a unified approach. For instance, pfold combines previously established models of molecular evolution (phylogenetic models) with probabilistic models of RNA secondary structure (stochastic context-free grammars - SCFGs).

We aimed to develop a method that could be readily extended to disparate data types and could encompass different probabilistic models for these. In particular they should be able to capture correlations both within and between data types. This is achieved by combining SCFGs with probabilistic graphical models (PGMs). The SCFG defines a prior over secondary structures, as it does in most other probabilistic methods (Rivas *et al.*, 2012). The PGMs model the sequence and any layers of probing data given the structure. PGMs are flexible and modular models useful for capturing select

*To whom correspondence should be addressed. Tel: +45 784 55360; Email: jakob.skou@clin.au.dk

dependencies in high dimensional data (Koller and Fridman, 2009). In our case, they model dependencies between the sequence and the probing data as well as dependencies along the sequence.

Standard algorithms allow efficient training of both SCFG and PGMs as well as prediction of the optimal secondary structure. Importantly, this allows ProbFold to be automatically trained without hand-setting any parameters given a sufficiently sized training set of known structures with probing data. The size of the needed training set depends on the number of free parameters in the model.

RNA structure probing has a long history and many different methods exist (Ehresmann *et al.*, 1987), including use of chemical agents (Ramazan *et al.*, 2006; Tijerina *et al.*, 2007; Merino *et al.*, 2005), RNases (Kertesz *et al.*, 2010), and spontaneous cleavage (Regulski and Breaker, 2008). Generally these modify bases or the backbone preferentially at either single or paired positions, allowing positional information on base-pair status through gel electrophoresis or sequencing. In the case of the SHAPE reagent (selective 2'-hydroxyl acylation analyzed by primer extension), it is the flexibility of the backbone that determines reactivity, which is generally higher for unpaired than paired regions (Merino *et al.*, 2005; Weeks, 2010; McGinnis *et al.*, 2008).

The interpretation of structure probing data is challenged by incomplete specificity of the methods, noisy or missing data, nucleotide biases, etc., which results in incomplete labeling of the primary sequence into paired and unpaired positions. For instance, in the case of SHAPE, the distributions of reactivities for paired and unpaired bases are largely overlapping (see Results). There is therefore a great need for computational methods that can integrate and make optimal use of probing data, beyond interpreting the data as a definite labeling of the primary sequence.

Recently, progress have been made on this problem with both physics-based methods (Mathews *et al*., 2004; Merino *et al.*, 2005; Deigan *et al.*, 2009; Washietl *et al.*, 2012; Swenson *et al*., 2012), sampling based methods (Quarrier *et al.*, 2010; Ouyang *et al.*, 2013), and a probabilistic method (Sükösd *et al*., 2012). These are briefly presented below. See (Eddy, 2014) for a detailed review and discussion of their statistical foundations.

In RNAstructure, SHAPE reactivities are converted to pseudo-energy change terms using a linear model optimized by prediction performance on a known structure (23S rRNA from *Escherichia coli* (*E.coli*)) (Deigan *et al.*, 2009; Mathews *et al*., 2004). GTfold provides a fast parallelized multi-core implementation of the energy minimization algorithm and similarly includes SHAPE data (Swenson *et al*., 2012). In a recent extension of RNAfold, pseudo-energy change terms are optimized for each structure given a loss function to maximize the agreement between the energy-based prediction and the experimental observations. In particular, no change is made when the sequence prediction is in complete agreement with the SHAPE data (Washietl *et al.*, 2012).

Another class of approaches samples structures from the Boltzmann-weighted ensemble and selects a representative structure with minimal Manhattan distance to a probing data profile (Quarrier *et al.*, 2010). This has been extended to several layers of probing data by a method that reduces them to a single binary pairing status profile (Ouyang *et al.*, 2013).

The recently proposed probabilistic method, PPfold 3.0 (Sükösd *et al*., 2012), extends pfold (Knudsen *et al*., 1999) by modeling both comparative sequence alignment data and experimental probing data. It uses stochastic context-free grammars (SCFGs) to model secondary structures, phylogenetic models to model alignment columns, and fine-grained discrete probability distributions to model SHAPE probing data. The study provides a proof of concept for including probing data in probabilistic methods.

Compared to PPfold 3.0, ProbFold offers a more general, extendible and parameter-sparse modeling approach that is evaluated using cross-validation. We develop ProbFold using existing SHAPE data (Deigan *et al.*, 2009) combined with an extensive set of known RNA structures (Rivas *et al.*, 2012) and evaluate a hierarchy of increasingly parameter-rich models. We find that including both base-pair stacking interactions and neighbor correlations for the SHAPE data improve performance. We also show how multiple types of probing data can be included in the models and may improve prediction performance. We find that the ProbFold approach exploits the probing data well, generally yielding higher performance gains than other methods, and present automatic procedures for optimizing the models on new data types.

2 MATERIALS & METHODS

2.1 SCFGs

SCFGs are the probabilistic variants of Context-Free-Grammars (CFGs). A CFG defines a formal language used for the generation of strings and is particularly suitable for RNA modeling, as it has the ability to capture nested long-range correlations (Dowel and Eddy, 2004). This approach has been widely applied in the context of RNA modeling and structure analysis (Sakakibara *et al*., 1994; Eddy and Durbin, 1994; Knudsen *et al*., 1999; Pedersen *et al*., 2004, 2006; Rivas *et al.*, 2012; Durbin *et al.*, 1998). An introduction to the use of SCFGs for RNA structure modelling and how they can be extended to handle the multivariate data of our setting is given in the supplementary material and methods section (section S1.1).

For the current study, we use (and extend) the pfold grammar (Knudsen *et al*., 1999), which models RNA secondary structures in terms of individual base pairs and unpaired nucleotides through the set of grammar rules: $S \rightarrow LS|L$; $F \rightarrow bF\hat{b}|LS$; $L \rightarrow a|cF\hat{c}$, where S, L, and F are the nonterminals, a, b, and c refer to the terminal symbols. However, the grammar does not explicitly model stacking interactions between consecutive base pairs, as done in nearest-neighbor energy models (Mathews *et al*., 2004; Xia *et al.*, 1998; Mathews *et al.*, 1999). Apart from hydrogen bonding between paired nucleotides (base pairing), stacking interactions between adjacent base pairs are the largest contributors to helix stability in nucleic acids (Yakovchuk *et al.*, 2006). We model these interactions for consecutive base-pairs by replacing a pair-emitting rule with a stack-emitting rule in the PFold grammar ($L \rightarrow cF\hat{c}$), which takes the previous base-pair into account.

The resulting grammar has six production rules, three of which emit terminals (Figure 1). The probability of the terminals given the transition between nonterminals is specified by emission models. When modeling only the RNA sequence, the terminals consist of nucleotides and the emission distributions can be defined simply by multinomials (Durbin *et al.*, 1998; Pedersen *et al*., 2004). To also model probing data, the emission models instead specify a joint distribution over both RNA sequence data and probing data. To achieve the flexibility needed for specifying joint distributions over multiple, potentially heterogeneous data types, we use a

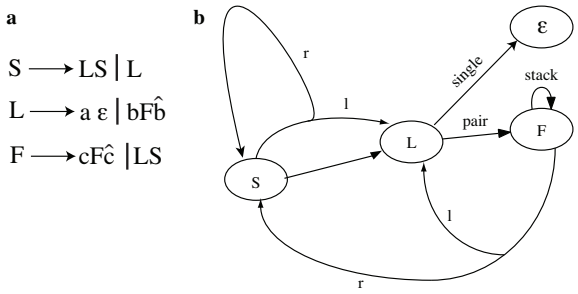


Fig. 1. (a) Grammar rules (see (b) for variable definitions). (b) Pictorial representation of the stacking grammar. The grammar has six production rules involving the four non-terminals S, L, F, and ϵ, which cannot be derived further. Three of the rules emit terminals, named *single*, *pair* and *stack*, and three are non-emitting, including two bifurcation rules. Each bifurcation rule splits into two parts, consisting of a left (l) non-terminal and right (r) non-terminal. The derivation starts in S. S can use either a bifurcation rule, which transits to L (l-part) as well as back to itself (r-part), or a non-emitting rule, which transits to L. L can use either the *single* emitting rule, which transits to ϵ and emits unpaired terminals (a), or use the *pair* rule, which transits to F and emits paired terminals ($a\hat{a}$). Finally, F can use the *stack* emitting rule, which transits back to F and emits (stacked) paired terminals ($b\hat{b}$) dependent on the previous base pair, or a bifurcation rule, which transits to L (l-part) as well as to S (r-part).

probabilistic graphical model framework to define the emission models.

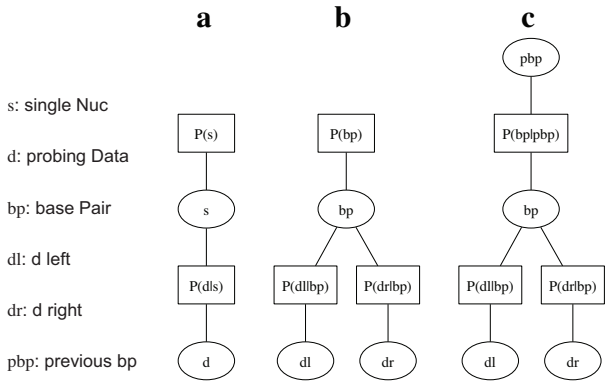
2.2 Emission Distributions and Probabilistic Graphical Models

Probabilistic graphical models (PGMs) offer a coherent and expressive framework for specifying and analyzing joint probability distributions (Koller and Fridman, 2009). PGMs are generally used to capture independence assumptions among a set of random variables and to specify their joint distribution as a factorization of local distributions each defined over subsets of variables. PGMs can be represented by mathematical graphs with nodes denoting random variables and edges denoting potential dependencies. A rich set of algorithms exist for doing inference with PGMs, which have proven a powerful tool for simplifying complex problems (Koller and Fridman, 2009).

We define the emission models as PGMs using the factor graph formalism (Figure 2) (Bishop, 2006). In this formalism, the PGMs are specified as undirected bipartite graphs between random variable nodes (represented by circles) and factor nodes (represented by squares). The factors hold potentially unnormalized probability distributions involving neighboring random variables.

Our current factor graph implementation of PGMs only handles discrete random variables, which simplifies the implementation and speeds up likelihood calculations. Including continuous random variables generally requires potentially slow numeric integration. Probing data are therefore discretized in a preprocessing step (see below).

For a start, ProbFold was developed to take an RNA sequence with a single affiliated sequence of probing data values as input. For each sequence position we thus have observed both a nucleotide and a discretized probing data value. We need to define an emission



d

	single	pair	stack
Joint prob.	P(s,d)	P(bp,dl,dr)	P(bp,pbp,dl,dr)
Factorization	P(s)P(d |s)	P(bp)P(dl|bp) P(dr|bp)	P(dl|bp)P(dr|bp) P(bp|pbp)
RNA Probing data	ACGUCACGU 001232454	ACGUCACGU 001232454	ACGUCACGU 001232454
Mapping	s←A, d←0	bp ← AU dl← 0, dr← 4	bp ←CG , pbp ←AU dl←0 , dr ←5

Fig. 2. Probabilistic graphical models defining the (a) *single*, (b) *pair*, and (c) *stack* emission models. The PGMs are shown as (bipartite) factor graphs, with variable nodes (circles) connected to factors (squares) defining local probability distribution. The variable abbreviations are given to the left. (d) For each emission model, the table gives (i) the joint probability distribution; (ii) its factorization specified by the PGM; (iii) example of short input data sequence with potential input positions highlighted. Note that the probing data has been discretized into six bins (0-5); (iv) mapping of data from highlighted sequence positions to relevant random variables of PGM.

distribution for each of the *single*, *pair*, and *stack* emitting rules (Figure 1). *Single* models only a single sequence position; *pair* models a pair of sequence positions; and *stack* models four sequence positions, consisting of two consecutive pairs (Figure 2d).

For each of the three emission models, we specify a PGM defining a joint distribution over the relevant nucleotides and probing data values (Figure 2). Initially, the *stack* model disregards the probing data of the previous base-pair. The PGM specification should reflect the independence structure of the modeled variables. We let the probing data at a position optionally depend on the observed nucleotides of that position. However, we let the probing data from the two sides of a base pair be independent of each other, given the observed nucleotides of the base pair (Figure 2). These independence assumptions are evaluated separately as part of the model development below.

2.3 Data Sets

The ProbFold models are potentially parameter rich. Optimally we would therefore train and test them on comprehensive sets of known RNA structures encompassing tens of thousands of base-pairs affiliated with consistently generated probing data. As such data sets do not yet exist, we complemented structure sets that include probing data with larger sequence-only sets.

Our primary structure probing set consisted of SHAPE data from *E. coli* 16S and 23S rRNAs (Deigan *et al.*, 2009; Weeks, 2012)

augmented with a set of seven small RNA structures that were downloaded from the RMDB repository (Cordero *et al.*, 2012) (RMDB set) and four taken from (Rice *et al.*, 2014) (see Table 1). Altogether, these include a total of 2,142 unpaired positions and 1,479 base pairs. The SHAPE data was preprocessed by denoting all invalid values (reactivity < 0; n=486) as missing data.

For a subset of the small structures from RMDB (n=6), DMS (dimethyl sulphate) and CMCT (1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene) probing data sets were also available (Cordero *et al.*, 2012) (RMDB set, Table S11). These structures encompass 287 base pairs and 593 unpaired positions, for which 302 positions had missing data for DMS and 415 for CMCT. We used these to illustrate the use of ProbFold on multiple types of probing data.

The probing data data sets were complemented with sets of sequence-only structures called TrainSet A (3,166 structures) and TestSet A (n=697), which were originally compiled by (Rivas *et al.*, 2012). We preprocess these sets by discarding structures with loops with less than three bases, reducing the size of TrainSet A to 2,707 (130,227 base pairs) and TestSet A to 593 (25,596 base pairs). In the supplementary methods and material (section S1.4), we describe training, testing and prediction procedures adopted in this work.

2.4 Probing Data Modeling and Discretization

The value of probing data for secondary structure prediction depends on the difference between its distribution in single stranded and paired regions. In ProbFold, these probing data distributions (P^{single} and P^{pair}) are explicitly modeled as part of the PGMs and may be conditioned on the primary sequence. Given our use of discrete PGMs, we discretize the probing data into k bins and use normalized histogram models (i.e. multinomials), which use $k-1$ free parameters. Initially we visualized these distributions for 16S and 23S SHAPE data using 15 equi-distant bins (Figure S1a). We discuss probing data modelling and discretization including optimal criterion for break points based on Kullback-Leibler (KL) divergence in the supplementary methods and materials (section S1.3).

2.5 Cross-Validation and Overfitting

Unlike some of the previous approaches (Washietl *et al.*, 2012; Sükösd *et al.*, 2012), we use cross-validation in our performance evaluation (section S1.2). We thereby avoid to train and test on the same probing data. For instance, when evaluating the performance on 16S, we train on 23S combined with RMDB and the sequence-only TrainSet A. This further limits the number of probing data related free parameters that can be learned without overfitting during development.

The number of free parameters (fp) in the probing data models is proportional to the number of bins used in the discretization. When the number of free parameters are increased, the models typically learn the properties of the training data well, but generalize poorly to other data sets. To select the optimal number of bins and to illustrate this behavior, we plot the prediction performance on both train and test sets when using probing data from 16S and 23S (Figure S1d). The test performance is optimal between 3 and 15 bins, whereas train performance continues to increase and approaches perfection due to overfitting. Based on this, we use six bins for the discretization in the model evaluation below.

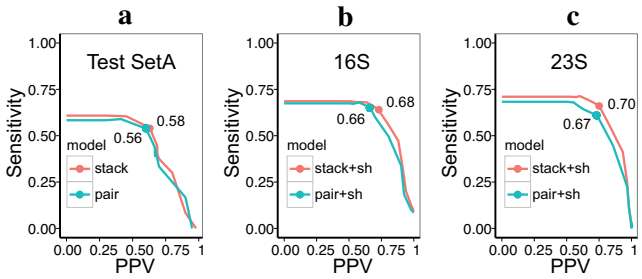


Fig. 3. ROC curves and maximal F-values for (a) sequence-only models on TestSet A, (b) probing data models on *E.* coli 16S rRNA, and (c) probing data models on *E.* coli 23S. The curves are made by varying the value of γ (see text).

3 RESULTS

3.1 Model Selection

We developed and evaluated a hierarchy of models capturing different correlations in the data with increasing number of free parameters (Table S1). The limited amounts of training data enforces a tradeoff as models with too many free parameters will be overfitted and not be robust.

3.2 Sequence-only Models

We started out with two sequence-only models: The *p*air model uses the original PFold grammar Knudsen *et al.* (1999) and is specified by 18 free parameters (fp). The *s*tack model uses the above described grammar extension, which also includes stacking interactions (fp=258). The *s*tack model showed a modest performance gain over the *p*air model in the ROC analysis and by F-measure (Figure 3a).

3.3 Probing Data Models

To extend the sequence-only models to also handle probing data, we developed emission models that generate both sequence data and SHAPE reactivities. To guide the development of these, we started by analyzing correlations in the 16S and 23S SHAPE data sets.

We first evaluated if the SHAPE reactivities were correlated with the primary sequence nucleotides. To control for compositional biases and the different level of reactivities, we did this separately for single (unpaired) and paired regions (Figure 4a,b). In both cases, there were significant differences in the distribution of the SHAPE values for the different nucleotides ($p<4.4e-03$ for *single* and $p<8.6e-06$ for *pair*; Kruskal-Wallis rank sum test).

We then evaluated if the SHAPE reactivities were correlated between the left and the right side of a base-pair (Figure 4c). Surprisingly no correlation was observed (Pearson correlation coefficient, pcc=-0.042; $p-value=0.075$). This may be explained by the overall low SHAPE reactivities of paired bases, causing experimental noise to dominate any underlying signal.

Based on these observations, we defined emission models where the SHAPE reactivities of the left and the right side of a base pair are modeled independently, but with each their own distribution. Altogether, the 25 free parameters are used to model the SHAPE reactivities given discretization in six bins (*single* model: fp=5; *pair* models: fp=2x5=10; *stack* models: fp=2x5=10). Preferably,

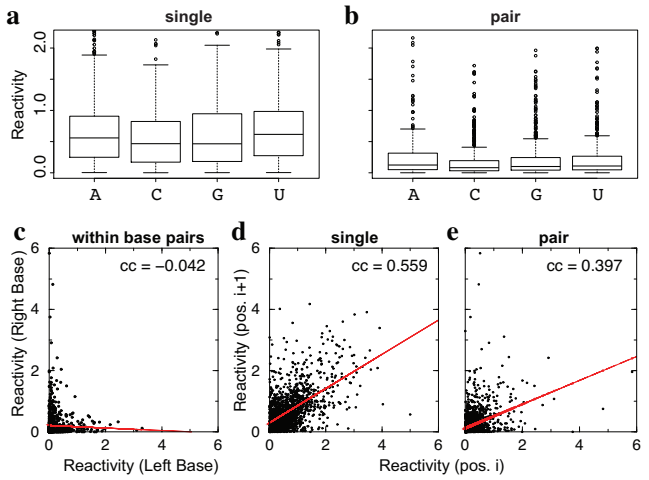


Fig. 4. Correlations in SHAPE data. Box-plots showing distribution of SHAPE reactivities for individual nucleotides for (a) single (unpaired) and (b) paired regions. Scatterplots showing (c) lack of correlation between left and right side of base pairs; (d) positive correlation along the sequence for both unpaired bases; and (e) positive correlation along the sequence for paired bases in stems. The regression line (red dashed line) summarizes the trend in the data.

ProbFold should capture differences in SHAPE reactivities among nucleotides. However, capturing these dependencies requires many additional free parameters (fp=4x25-25=75). Given the limited training data, we therefore model the primary sequence and the SHAPE reactivities independently (e.g., the *single* model becomes $P(s,d) = P(s) \times P(d)$, following the notation of figure 2).

These emission models were combined with the grammar of the sequence-only *s*tack model selected above to give the *s*tack+sh (fp=283) model. The inclusion of SHAPE data dramatically improved sensitivity and overall performance (Figure 3b,c and Table S2)

We finally evaluated if SHAPE reactivities correlate with neighboring positions, again analyzing single and paired regions separately (Figure 4d,e). Significant positive correlations were observed in both cases ($p<0.0001$), with an overall higher correlation in single regions (pcc=0.559) than in paired regions (pcc=0.397). The correlations may reflect overall steric constraints, which are likely to be correlated along the primary sequence. For instance, backbone flexibility of loop positions may decrease toward stems.

We extended the emission models of the *s*tack+sh model to capture these sequential correlations in the SHAPE data (Figure S1), which requires many additional free parameters (fp=125). The resulting model, *s*tack+sh+cor (fp=408), improves performance over simpler models when trained on the 23S data set, but not when trained on the smaller 16S data set. We attribute the decrease in performance on 23S (842 base pairs) to overfitting of the parameter rich correlation model when trained on the 16S data set (468 base pairs). Overall, we recommend the *s*tack+sh+cor model for SHAPE prediction applications. We make a version trained on the combined 16S and 23S data sets available for download together with the other ProbFold models (`http://moma.ki.au.dk/prj/probfold/`).

Table 1. Prediction performance of ProbFold and other methods on set of small RNA structures. Both the F-value and the change in F-value (ΔF) relative to the sequence-only (Seq-only) predictions are shown. See Table S3-10 for the full set of performance statistics.

	ProbFold		PPFold		RNAstructure		GTFold	
Structures	*F-Value*	Δ F	*F-Value*	Δ F	*F-Value*	Δ F	*F-Value*	Δ F
5S RNACordero *et al.* (2012)	0.54	**0.24**	**0.57**	0.22	0.24	0.00	0.25	0.00
Adenine riboswitchCordero *et al.* (2012)	**1.00**	**1.00**	0.96	0.51	0.96	-0.05	**1.00**	0.00
cidGMP riboswitchCordero *et al.* (2012)	0.63	0.14	0.55	**0.21**	**0.73**	-0.15	0.70	-0.01
GlycinCordero *et al.* (2012)	0.76	0.22	0.65	**0.47**	**0.88**	0.23	0.85	0.21
P4-P6 domain (Tetrahymena ribozyme)Cordero *et al.* (2012)	0.87	**0.37**	0.80	0.30	**0.88**	0.01	0.84	0.07
Ribonuclease PCordero *et al.* (2012)	**0.79**	**0.11**	0.20	-0.49	0.57	-0.01	0.39	-0.39
tRNA phenylalanine (yeast)Cordero *et al.* (2012)	**0.98**	**0.79**	0.44	0.12	**0.98**	0.03	0.95	0.71
M-Box riboswitchRice *et al.* (2014)	**0.71**	-0.10	0.47	-0.37	0.52	**0.04**	0.71	-0.18
Lysine riboswitchRice *et al.* (2014)	**0.28**	-0.03	0.22	-0.06	0.28	**0.06**	0.26	-0.00
Group I Intron, T. thermpphilaRice *et al.* (2014)	**0.79**	**0.23**	0.66	0.03	0.78	0.10	0.75	0.16
Group II Intron, O. iheyensisRice *et al.* (2014)	0.51	**0.27**	0.53	0.23	**0.60**	-0.07	0.59	-0.02
Average	**0.71**	**0.29**	0.55	0.11	0.67	0.02	0.66	0.05

We also evaluated the performance of three existing methods, PPfold 3.0 (Sükösd *et al.*, 2012), RNAstructure v5.6 (Mathews *et al.*, 2004) and GTFold-3.0 (Swenson *et al.*, 2012), on the 16S and 23S sequences with and without SHAPE data (Table S2). For the sequence-only predictions, the ProbFold *s*tack model results in a low sensitivity (<0.30) and a high PPV (∼0.80). This pattern is shared by ppfold, though it has somewhat poorer performance, which may be explained by it not modeling stack interactions. RNAstructure and GTFold, which both have richer structure models, have more balanced sensitivity and PPV performance. As a result they both have higher F-values for 23S and, in the case of GTFold, also for 16S. ProbFold *s*tack has the highest accuracy (ACC) on both sequences.

When including SHAPE data, ProbFold continues to have low sensitivity (47-62%) and high PPV (76-90%) compared to the other methods. However, ProbFold's tradeoff between sensitivity and PPV can be adjusted (Figure 3). RNAstructure has the highest overall performance both by accuracy and F-value, closely followed by GTFold on the 16S sequence. However, the 16S and 23S SHAPE data sets used in the development of ProbFold, have also been heavily used for developing the SHAPE models of the other methods. Specifically, PPfold was trained on both 16S and 23S and the RNAstructure parameters were chosen based on analysis of 23S with performance evaluation on 16S (information not available for GTfold; Table S2). This could lead to overfitting and performance statistics that do not generalize. In contrast, the performance results for ProbFold are based on cross-evaluation, to avoid the effect of overfitting to the train data set (Figure S1d). The performance results are therefore not directly comparable between methods.

As an independent test data set, we evaluate the performance on a set of small RNA structures with SHAPE data (Table 1 and S5-10). On this set ProbFold achieves the highest F-value and accuracy on six of the eleven structures as well as the highest overall F-value and accuracy across all structures (Table 1 and S9).

Since our main focus is ProbFold's ability to exploit the probing data, we also evaluated the relative gain in performance when including probing data over sequence-only predictions (Table 1,S6,S8, and S10). ProbFold showed the highest gain for seven of the eleven structures, with an average F-value gain of 0.29. The other methods showed smaller relative gains, with PPfold at 0.11, GTFold at 0.05, and RNAstructure at 0.02.

3.4 ProbFold with other Data Types

To illustrate the use of ProbFold on other types of probing data, we retrained and applied the model on publicly available DMS and

Table 2. Average performance on six small structural RNAs of the Multi-data versions of ProbFold and RNAstructureMathews *et al*. (2004); Cordero *et al.* (2012) with step-wise inclusion of CMCT, DMS and SHAPE structure probing data. Both the F-value and the change in F-value (ΔF) relative to the sequence-only (seq-only) predictions are shown. See Table S13 for the full set of performance statistics.

	ProbFold		RNAstructure	
Data	*F-Value*	Δ F	*F-Value*	Δ F
Seq-only	0.40	0.00	**0.73**	0.00
Seq, CMCT	0.48	0.08	**0.85**	**0.12**
Seq, CMCT, DMS	0.54	**0.14**	**0.85**	0.12
Seq, CMCT, DMS, SHAPE	0.71	**0.31**	**0.82**	0.09

CMCT data sets covering six small RNA structures (see Data Sets). We used leave-one-out cross evaluation to train and evaluate the performance of the *s*tack+sh model from above. Given the limited amount of training data, we use the KL approach to discretize the probing data into three bins only, which reduces the number of free parameters used to model probing data from 15 to 6. For both types of data, the overall performance improved compared to using only the primary sequence (Table S11-12). Overall, using DMS resulted in better prediction performance than CMCT (F-values of 0.54 versus 0.48). The lower power of CMCT compared to DMS is also apparent from the smaller separation between the probing signal intensity distributions for paired and unpaired positions (Figure S3).

As both DMS and CMCT probing have known strong nucleotide dependencies, we also evaluated a version of the *s*tack+sh model where the probing signal distributions depend on the sequence nucleotides (as shown in Figure 2). For DMS the separation of the paired and unpaired signal distributions conditional on sequence nucleotide appear to improve slightly (Figure S4). Whereas for CMCT no improvement is obvious (Figure S5). However, the prediction performance decreased for both models in cross-evaluation (F-values of 0.40 for DMS and 0.37 for CMCT), likely due to the limited training data and many additional free parameters (n=18) introduced in this model even when discretized into three bins only.

3.5 Modeling Multiple Data Types

When available, integration of multiple probing data types should increase prediction accuracy. We here show how ProbFold's emission models can be extended to handle multiple data types, using the *single* model as an example (Figure S6a). Based on the emission models of Figure 2, we suggest to model multiple types of probing data (d_1, d_2, d_3, d_4, and d_5) as independent given the nucleotide of the primary sequence. For the *single* model, the joint distribution thus becomes $P(s, d_1, d_2, d_3, d_4, d_5) = P(s) \times \prod_{i=1}^{5} P(d_i|s)$. The other emission models can be simply defined following the same scheme.

To demonstrate the performance and the advantage of using multiple data sets, we first provide a proof-of-principle experiment based on boot-strapped data, given the limited extent of existing real data sets for well annotated structures. Our results shows that the performance of the model goes up with the use of multiple data sets (section S1.5).

To further demonstrate the benefit of combining multiple probing data sets and to also include different data types, we applied the above described multiple-data-set model to the six previously described RMDB structures for which CMCT, DMS, and SHAPE are all available (Cordero *et al.*, 2012). Given the limited extent of the available data sets, each is modeled using only three bins, with KL optimized break points (section S1.3), to retain the number of model parameters. The experiment is thus still at the proof-of-principle level. As more data become available, more bins can be used to capture the structure signal of the probing data, which is expected to improve performance. The performance was measured using leave-one-out cross evaluation and averaged across all six structures (Table 2, S13).

Both sensitivity and the overall performance as measured by accuracy and F-value increase when incrementally adding each of the three probing data sets, starting with sequence-only (Figure S6c). A slight decrease of PPV is observed when adding DMS. Integration of SHAPE, DMS and CMCT data have previously been carried out using RNAstructure v. 5.3 and pseudo-energy terms (Cordero *et al.*, 2012). For comparison, we evaluate the performance of the multi-data version of RNAstructure on RMDB structures (Table 2, S13). While RNAstructure performs much better on the sequence-only data set and achieves the highest overall F-values, ProbFold shows the largest relative gains from including probing data sets. ProbFold also show consistent gains with each added data set, which is not the case for the RNAstructure model (Table 2). This suggests that the ProbFold emission models are able to make good use of the available structure signal. The emission models can be further extended to account for dependencies both within and among multiple data sets (section S1.6).

4 DISCUSSION

We have presented a probabilistic method for RNA secondary structure prediction that integrates experimental structure probing data. One of the virtues of our approach is its flexibility. The underlying model was initially developed on a capillary electrophoresis SHAPE data set, but it can readily be retrained and applied on other data types given sufficient training data as well as extended to handle multiple data types. We demonstrate these extensions with proof-of-concept examples on existing data sets and on generated (boot-strapped) data. We develop and train versions of ProbFold for SHAPE, DMS, and CMCT probing data - both individually and in combination. We also evaluate different variants of these, for instance including nucleotide dependencies. The flexibility is achieved by the use of a highly modular probabilistic model with accompanying efficient algorithms for training and prediction (Sakakibara *et al*., 1994; Knudsen *et al*., 1999; Eddy and Durbin, 1994; Pedersen *et al*., 2004, 2006; Nawrocki and Eddy, 2013 ; Rivas and Eddy, 2001; Eddy, 2014; Knudsen *et al*., 2003; Rivas *et al.*, 2012; Durbin *et al.*, 1998; Koller and Fridman, 2009; Bishop, 2006).

As with most other probabilistic RNA secondary structure prediction methods (Rivas and Eddy, 2000; Metzler and Neble, 2008), we use stochastic context-free grammars to model the secondary structure. We find limited benefit of modeling stacking interactions, as have others (Rivas *et al.*, 2012). This may be because individual base-pair parameters already account for most of the stacking interaction effects (Yakovchuk *et al.*, 2006). We factorize the grammar rules into transitions and emissions. In

contrast to other methods, we explicitly specify the emission distributions using probabilistic graphical models (PGMs), which may be defined over multi-variate input data. The implementation closely reflects this modularity, with separate textual specifications of the overall grammar, the transitions, the PGMs defining the emission distributions; and the mapping of observed data to PGM variables.

As part of the model development, we evaluated correlations in the probing data. We found modeling SHAPE probing data correlations along the sequence improved performance, given enough training data was present. In general, detecting and modeling the prominent dependencies in the observed data is expected to improve the fit and the discriminatory power of the emission distributions and hence overall model performance. However, including correlations makes the models more complex with many additional free parameters to learn. Given the limited size of probing data sets for training, overfitting and lack of robustness easily becomes a problem, as shown for parameter-rich versions of ProbFold (e.g., Figure S1d and Table S2).

ProbFold bins the continuous probing data values to allow for the use of discrete PGMs. This simplifies the PGM implementation, for instance by avoiding computationally expensive numerical integration, and avoids use of analytical continuous distributions with a potentially poor fit. As the structural signal of the probing data depends on the differences in its distribution in different structural regions, both the number of bins and their boundaries are important parameters of the model. We show that the number of bins should be kept small given the amount of available training data to avoid overfitting. We suggest to select bin boundaries by optimizing the difference between the *single* and *pair* probing data distribution using KL divergence.

As an alternative to discretizing the data, parameter free continuous distributions, such as kernel distributions, could be used. These however easily become computationally heavy, as they in principle are specified by the full training data set. Given knowledge of the uncertainty of the probing data observations, a more satisfying approach would be to explicitly model the uncertainty of the individual probing data values. Such knowledge would be available, e.g., with counts from NGS-based probing data, as some of transcriptome-wide approaches produce (Kertesz *et al.*, 2010; Lucks *et al.*, 2011).

The RNAstructure method converts SHAPE reactivities to pseudo energy change terms, which are incorporated when predicting the minimal free energy structure (Deigan *et al.*, 2009). The conversion is done using a simple linear parametric form, which only requires two free parameters. Using a simple parametric form limits the number of free parameters, but may also introduce bias if the fit is poor in part of the probing data value range. In particular, RNAstructure has been shown to perform extremely well on 16S rRNA when introducing several preprocessing steps and filters, such as (1) selecting parameters performing well on 23S; (2) limiting the allowed distance between base pairs; (3) focusing on sites with useful SHAPE data; and (4) disregarding sites with clear incompatibilities with the comparative reference structure (Deigan *et al.*, 2009). Though individually helpful, manually selecting parameters and introducing many preprocessing steps risk making the approach more liable to overfitting on a concrete data set. Direct calculation of pseudo-energy change terms based on log-likelihood ratios of being paired versus unpaired has also been suggested (Cordero *et al.*, 2012), which is closer to the approach taken by ProbFold.

ProbFold has been designed with the aim of extendibility to multi-variate probing data measurements. This could for instance be the combination of SHAPE with other chemical or enzymatic probing agents, as in the proof-of-concept example using CMCT, DMS, SHAPE data. If the noise in the individual measurements at a site are correlated it becomes important to capture these correlations in the model to retain specificity. Such correlations could for instance be caused by tertiary structure interactions involving single stranded regions, which may affect several types of probing agents, including SHAPE. Specifically, non-canonical base pairs will often give similar signal to canonically paired bases using SHAPE but not DMS. Learning such correlations would therefore be expected to improve prediction performance. As more and more RNA tertiary structures are found, one solution could be to explicitly include tertiary structure aspects into the model (Kopeikin and Chena, 2005; Lorenz *et al.*, 2013).

In the case of SHAPE, we did not observe any correlation between paired bases. However, such correlations may well exist for other types of probing agents, such as double stranded RNases. Even partial evidence for the presence of specific base pairs could significantly improve performance by constraining and simplifying the folding problem. Such evidence would resemble the signal exploited from compensatory base pair substitutions exploited in comparative RNA structure analysis. As suggested by (Sükösd *et al.*, 2012), additional power could be gained by combining experimental probing data with comparative data, though this would be limited to functional and conserved RNA structures.

Through proof-of-principle experiments we have illustrated the applicability of ProbFold to different types of probing data and to multiple complementary data sets. The performance evaluations show that ProbFold can make efficient and competitive use of the probing data, both for SHAPE data and when combining multiple data sets (Table 1, 2). However, model performance is limited by the small size of the available training data sets, which restricts model complexity and hence predictive power.

We hope the advent of NGS-based high-throughput structure probing techniques, as pioneered by (Kertesz *et al.*, 2010; Underwood *et al.*, 2010; Lucks *et al.*, 2011), will result in large uniform probing data sets of known structures assessed by multiple probing agents. This would allow multi-variate versions of ProbFold or similar models to be trained, with an expected boost in performance characteristics. Ultimately such approaches could help improve RNA structure maps transcriptome-wide.

5 FUNDING

This work was supported by a Danish Strategic Research Council grant to Center for Computational and Applied Transcriptomics (COAT) [10-092320/DSF].

6 ACKNOWLEDGMENTS

We thank Kevin Weeks for sharing SHAPE Reactivities for *E. coli* 16S and 23S rRNAs and Zsuzsanna Sükösd, Jeppe Vinther and Lukasz Jan Kielpinski for fruitful discussions and manuscript

comments. This work greatly benefited from discussions at the Benaque RNA workshop in 2012.

REFERENCES

Bishop,M.C. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA.

Cordero,P., Kladwang,W., VanLang,C.C. and Das,R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference, *Biochemistry*, **51**, 7037.

Cordero,P. Lucks,J.B. and Das,R. (2012) An RNA Mapping Database for curating RNA structure mapping experiments. doi: 10.1093/bioinformatics/bts554.

Deigan,K.E., Lia,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination, *Proc. Natl. Acad. Sci.*, **106**, 97-102.

Dowel,R.D. and Eddy,S.R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction (2004), *BMC Bioinformatics*, **5**, 71.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press.

Eddy,S.R. and Durbin,R. (1994) RNA Sequence Analysis Using Covariance Models, *Nucleic Acids Res.*, **22**, 2079-2088.

Eddy, S.R. (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes, *Annu Rev Biophys.*, **43** , 433-56.

Ehresmann, C., Baudin,F., Mougel, M., Romby,P., Ebel. E-P. and Ehresmann, B. (1987) Probing the structure of RNAs in solution, *Nucleic. Acids Res.*, **15**, 9109-9128.

Karadumana,R., Fabrizioa,P., Hartmutha,K., Urlaubb,H. and Luhrmann,H. (2006) RNA Structure and RNA - Protein Interactions in Purified Yeast U6 snRNPs, *J. Mol. Biol.*, **356**, 1248-1262.

Kertesz,M, Wan,Y., Mazor,E., Rinn,J.L., Nutter,R.C., Chang,H.Y. and Segal,E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103-107.

Kertesz,M., Wan,Y., Mazor,E., Rinn,J., Nutter,R., Chang,H. and Segal,E. (2010) Probing RNA structure genome-wide using high throughput sequencing, *Protocol Exchange*, doi:10.1038/nprot.2010.152.

Knudsen,B. and Hein,J.J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars *Nucleic Acids Res.*, **13**, 3423-3428.

Knudsen,B. and Hein,J.J. (1999) Using stochastic context free grammars and molecular evolution to predict RNA secondary structure, *Bioinformatics*, **15**, 446-454.

Koller,D. and Friedman,N. (2009) *Probabilistic Graphical Models*. MIT Press.

Kopeikin,Z. and Chena,Shi-Jie. (2005) Statistical thermodynamics for chain molecules with simple RNA tertiary contacts, *J. Chem. Phy.*, **122**, 094909.

Lorenz,R. and Bernhart,S. and Qin,J. and Höner zu Siederdissen,C. and Tanzer,A., Amman,F., Hofacker,I.L. and Stadler,P. (2013) 2d meet 4g: G-quadruplexes in rna secondary structure prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **99** (PrePrints), 1.

Lucks,J.B., Mortimer,S.A., Trapnell,C., Luo,S., Aviran,S., Schroth,G.P., Pachter,L., Doudna,J.A. and Arkin,A.P. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), *Proc. Natl. Acad. Sci.*, **108**, 11063-11068.

Mathews,D.H., Sabina, J., Zuker. M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, *J. Mol. Biol*, **288**, 911-940.

Mathews,D.H., Dinsey,D.M., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, *PNAS*, **101**, 7287-7292.

McGinnis,J.L., Dunkle,J.A., Cate,J.,H. and Weeks,K.M. (2012) The mechanisms of RNA SHAPE chemistry, *J. Am. Chem. Soc.*, **134**, 12319.

Merino,E.J., Wilkinson,K.A., Coughlan,J.L. and Weeks,K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE), *J. Am. Chem. Soc.*, **127**, 4223-31.

Metzler,D. and and Nebel,M.E. (2008) Predicting RNA secondary structures with pseudoknots by MCMC sampling, *Journal of Math Biol.*, 56, 2008.

Nawrocki,E.,P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches.*Bioinformatics*, **15** , 2933.

Ouyang,Z., Snyder,M.P. and Chang,H.Y. (2013) SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data, *Genome Res.*, **23**, 377 - 387.

Pedersen,J.S., Meyer,I.M., Forsberg,R., Simmonds,P. and Hein,J. (2004) A comparative method for finding and folding RNA secondary structures within protein-coding regions, *Nucleic Acids Res.*, **32**, 4925-4936.

Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad−Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and Classification of Conserved RNA Secondary Structures in the Human Genome, *PLoS Comput. Biol.*, **2**, e33.

Quarrier,S., Martin,J., Davis-Neulander,L., Beauregard,A. and Laederach,A. (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction, *RNA*, **16**, 1108-1117.

Regulski,E.E. and Breaker,R.R. (2008) In-line probing analysis of riboswitches, *Methods Mol. Biol.*, **419**, 53-67.

Rice,G.M., Leonard,C.W, and Weeks,K.M. (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE, RNA , **20**, 846 - 854.

Rivas,E., Lang,R. and Eddy,S.R. (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more, *RNA*, **18**, 193-212.

Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs, *Bioinformatics*, **16**, 583-606.

Rivas, E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis.*BMC Bioinformatics*, **2**, 8.

Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjolander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic Context-Free Grammars for tRNA Modeling, *Nucleic Acids Res.*, **22**, 5112-5120.

Sükösd,Z., Knudsen,B., Kjems,J. and Perdesen,C. (2012) PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, **28**, 2691-2692.

Swenson,M.S.,Anderson, J.,Ash, A.,Gaurav, P.,Sükösd,Z., Bader,D.A., Harvey,S.C. and Heitsch,C.E. (2012) GTfold: enabling parallel RNA secondary structure prediction on multi-core desktops, *BMC Res. Notes*, **5**, 341.

Tijerina,P., Mohr,S. and Russell,R., (2007) DMS footprinting of structured RNAs and RNAprotein complexes. *Nature Protocols*, **2**, 2608 - 2623.

Underwood,J.G., Uzilov,A.V., Katzman,S., Onodera,C.S., Mainzer,J.E., Mathews,D.H., Lowe,T.M., Salama,S.R. and Haussler,D. (2010) FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing, *Nat. Methods*, **7**, 9951001.

Wan,Y.,Kertesz,M., Spitale,C., Segal,E. and Chang, H.Y. (2011) Understanding the transcriptome through RNA structure, *Nat. Rev. Gen*, **12**, 641-655.

Washietl,S., Hofacker,I. L., Stadler,P.F. and Kellis,M. (2012) RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res.*, **40**, 4261-72.

Weeks,K.M. (2012) 16S ans 23S E. coli data, Personal communication.

Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing, *Current Opinion in Structural Biology*, **20**, 295-304.

Xia,T., SantaLucia, J.Jr., Burkard, M.E., Kierzek,R., Schroeder, S.J., Jiao,X.,Cox,C., and Turner,D.H.(1998) Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs, *Biochemistry*, **37**, 14719-14735.

Yakovchuk, P., Protozanova,E. and Frank-Kamenetskii,M.D. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix, *Nucleic Acids Res.*, **34**, 564.

Supplementary information for: ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction

Sudhakar Sahoo, Michał Świtnicki, and Jakob Skou Pedersen

List of Figures

List of Tables

1 Supplementary methods and materials

1.1 Modelling RNA structure using SCFGs

One can define a CFG (G) as composed of a set of nonterminals (or states) (N) including a start symbol $S \in N$, a set of terminals (or observable symbols) (T) (for RNA these symbols are a, c, g and u), and finally a set of production rules P used to re-write a non-terminal into a string with terminals or non-terminals or both. Collectively a CFG can be written as $G = (N, T, P)$. The string generation procedure is an iterative process that starts with S followed by subsequent application of production rules ($V \rightarrow \beta$, where $\beta \in (N \cup T)^*$ and $V \in N$). The application of production rules continues until all non-terminals have been rewritten and the final generated string of terminals is left. There may exist more than one path to generate the same string, and the paths are often represented as parse trees. If there is more than one parse tree for the given string, the grammar is called ambiguous. A CFG takes the form of an SCFG when the production rules are associated with probabilities, such that the sum of probabilities of out-going productions from a given nonterminal is one. The probabilistic variant of CFGs is needed to score and rank the parse trees.

RNA folding problems involve finding a valid and optimal secondary structure for a given input RNA primary sequence. This is generally done either by identifying the structure that minimizes a free energy measure, as in thermodynamics folding methods, or maximizes a probability measure, as in probabilistic methods. When SCFGs are used for structure prediction, the production rules that generate terminals reflect different structural components, such as single-stranded regions or base-pairs, such that the parse tree annotates the observed sequence with a secondary structure.

More concretely, for a given input sequence (x) and a set of probability parameters (Θ), the SCFG G defines a joint distribution $P(x, \sigma|G, \Theta)$ over the sequence and all the parse trees (σ). As x is given, $P(x|G, \Theta)$ is a constant and the joint distribution is proportional to the posterior distribution of σ ($P(\sigma|x, G, \Theta)$). By maximizing $P(x, \sigma|G, \Theta)$ with respect to σ, one can therefore obtain the parse tree ($\hat{\sigma}$) with the highest posterior probability. The cubic-time dynamic programming algorithm, Cocke-Younger-Kasami (CYK) [6], may be used to calculate the probability for $\hat{\sigma} = argmax_\sigma P(x, \sigma|G, \Theta)$. Further a trace-back procedure recovers the parse tree $\hat{\sigma}$ and hence the optimal secondary structure.

The quantity $P(x, \sigma|G, \Theta)$ is the product of the probabilities of all the production rules in σ used to generate the sequence x. It is possible to factorize each production rule into a transition probability and an emission probability [5]. The transition probability refers to the probability of replacing the original nonterminal with one or more new ones and the emission probability similarly refers to the probability of generating the terminals of the rule. In a compact notation, this may be written as, $P(V \rightarrow \beta) = P(V \rightarrow \beta_N) \times P(\beta_T|V \rightarrow \beta_N)$, where β_N and β_T refer to the nonterminal and terminal components of β, respectively. $P(V \rightarrow \beta_N)$ is the transition probability and $P(\beta_T|V \rightarrow \beta_N)$ is the emission probability.

1.2 Performance Measures

We evaluated the relative performance of variants of the ProbFold model capturing different dependencies in the data. For this, we measured the folding accuracy by calculating the two standard statistical measures: (i) Sensitivity (SEN), which is defined as the ratio between the number correctly predicted base pairs (TP) and total number of base pairs in a known reference structure (TP + FN), i.e., SEN = TP/(TP + FN); and (ii) Positive Predictive Value (PPV), which is the ratio between the number of correctly predicted base pairs and total number of predicted base pairs (TP + FP), i.e., PPV=TP/(TP + FP).

We further used SEN and PPV to generate Receiver Operating Characteristic curves (ROC curves). A ROC curve reflects the tradeoff between sensitivity and specificity (here PPV), which can be adjusted via the γ parameter in our case. The performance of a prediction method can be summarized by the area under the ROC curve (AUC), which varies from zero to one, with one equating perfect predictions. ROC curves are normally used for binary classification problems, in which case the AUC is 0.5 for random guessing. Predicting the set of base pairs correctly is a much harder problem than binary classification and the AUC baseline for random guessing will thus be lower. We also calculate AUCs for the simpler binary classification problem of correctly labeling each position as paired or unpaired (AUC^{label}). For most RNA secondary structure prediction methods the tradeoff between sensitivity and specificity cannot be adjusted and the AUC cannot be evaluated.

In addition, we summarize the performance using the F-measure, defined as the harmonic mean of the sensitivity and PPV for a given value of γ (F = 2 × SEN × PPV/(SEN + PPV)). The value of the F-measure lies

between zero and one, with a value of one for perfect predictions. Finally, the performance was also summarized using accuracy (ACC), which is the arithmetic mean of sensitivity and PPV.

1.3 Data discretization and optimality criterion

The Kullback Leibler (KL) divergence (D_{KL}) optimality criterion can be defined as

$$D_{KL}(p^{single}||p^{pair}) = \sum_{i}^{k} p_i^{single} ln \frac{p_i^{single}}{p_i^{pair}}. \quad (1)$$

The KL divergence measures the expected information content, or relative difference in probability from above, of drawing from one distribution (here p^{single}) relative to the other (p^{pair}). Given its relation to likelihood ratio tests, it can be interpreted as the expected information to discriminate between the alternative hypotheses specified by the two distributions [12]. $D_{KL}(p^{single}||p^{pair})$ therefore measures how informative the probing data is for secondary structure prediction. Motivated by this, we define a greedy search procedure for finding the $k-1$ break points that optimize $D_{KL}(p^{single}||p^{pair})$ for k bins (Figure S1b).

The specific values of probing data measurements may differ between experiments due to differences in reagent concentration, reaction time, etc.. As the relative ordering of values is expected to be more robust to these differences, we preprocess the probing data by ranking and normalizing the values to fall between zero and one. The final binning and modeling of the data is therefore done on this scale (Figure S1c).

1.4 Training, Testing and Prediction

We use Expectation-Maximization (EM) algorithms [6, 7] to learn the transition probabilities and the parameters of the emission models given our training data. The EM algorithm is iterative and proceeds through an expectation (E) step and a maximization (M) step. The E step provides expectations for unobserved variables, which are then treated as (weighted) observations in the following maximum likelihood estimation of model parameters (M-step). The iteration continues until convergence. In our case, the Inside-Outside algorithm for SCFGs [6] and the Sum-Product algorithm for PGMs [8] provide the needed expectations.

The structure annotations of the train set are fully specified. Since the ProbFold grammar is unambiguous, meaning that a structure annotation implies a unique parse of the grammar, the use of grammar rules is therefore fully observed. For fully observed data, the E-step reduces to counting and the transitions are therefore learned in a single iteration. As a consequence, the training of the transitions and the emission models become independent. In practice we therefore train them separately (see Software).

We include both structure sets with and without probing data in the training. The structures lacking probing data (TrainSet A and TestSet A) are treated as having missing data values in their place. This corresponds to evaluate these inputs with reduced emission models without the probing data random variables. In effect, these structures will only contribute to the training of the transitions and the nucleotide part of emission model PGMs.

Given a trained model, we use the maximum expected accuracy (MEA) optimality criterion [10, 9] for structure prediction. MEA optimizes the expected prediction accuracy per position given the data and may result in a different structure than the single most probable one given by the CYK algorithm.

To find the MEA structure, we first evaluated the posterior probability of base pairing, $p(i, j)$, for every pair of positions. The posterior probability of being unpaired is then calculated as $q(i) = 1 - \sum_{j<i} p(j, i) - \sum_{j>i} p(i, j)$. A dynamic programming algorithm similar to that of [11] is used to determine the secondary structure with maximum score. To be able to control the tradeoff between sensitivity and specificity, a tradeoff parameter, γ, is introduced in the MEA score (score=$\sum_{paired(i,j)} 2\gamma p(i, j) + \sum_{unpaired(i)} q(i)$) [10]. γ can take any value in the range (0,∞) and has a default value of 1. Increasing the value of γ increases the tendency to predict base pairs.

1.5 Simulation of multiple data sets

To simulate the boot-strapped data, we model the probing data as independent of the primary sequence ($P(s, d_1, d_2, d_3, d_4, d_5) = P(s) \times \prod_{i=1}^{5} P(d_i)$) and use the stacking grammar. Based on our original 16S and 23S SHAPE data, we boot-strap five artificial probing data sets, with the same properties as the original, by permuting the SHAPE reactivities separately within the unpaired and paired regions. The resulting data is thus assumed to lack any correlation between the SHAPE reactivities given the secondary structure,

exactly following the independence assumptions of the constructed multi-data model.

The performance of the model is evaluated in the same type of cross-evaluation experiments as previously with zero to five boot-strapped probing data sets included. Each of these experiments are repeated ten times, with the average performance noted (Figure S6b). The sensitivity increases with the number of probing data sets included for both 16S and 23S. However, the PPV generally declines. The largest gains in F-value are seen with the first two data sets, after which it increases little and even declines slightly. We would have expected the performance to consistently increase with the increasing signal of additional data sets. Once again, we believe overfitting of the model is at cause. We therefore repeat the experiments without cross-evaluation (training on both 16S and 23S), such that the training data and testing data effectively comes from the same source. In this case, the PPV only decreases slightly and the F-values increases consistently as more probing data sets are included, showing the advantage of multiple data sets (Figure S6b, lower panel).

1.6 Modelling dependencies both within and among multiple data sets

Given sufficient training data, dependencies between data sets can for instance be captured by introducing a hidden variable, which the observed probing data depend on (see Figure S7 for the *single* emission model using this approach). The different states of the hidden node can be thought of as different modes of correlation between the data. The joint emission distribution over the sequence data and the probing data then becomes $P(s, d_1, d_2, d_3, d_4, d_5) = \sum_{h \in \{d_{i,i=1..5}\}} P(s, h, d_1, d_2, d_3, d_4, d_5) = \sum_{h \in \{d_{i,i=1..5}\}} P(s) \times P(h|s) \prod_{i=1}^{5} P(d_i|h)$.

2 Supplemental figures and tables

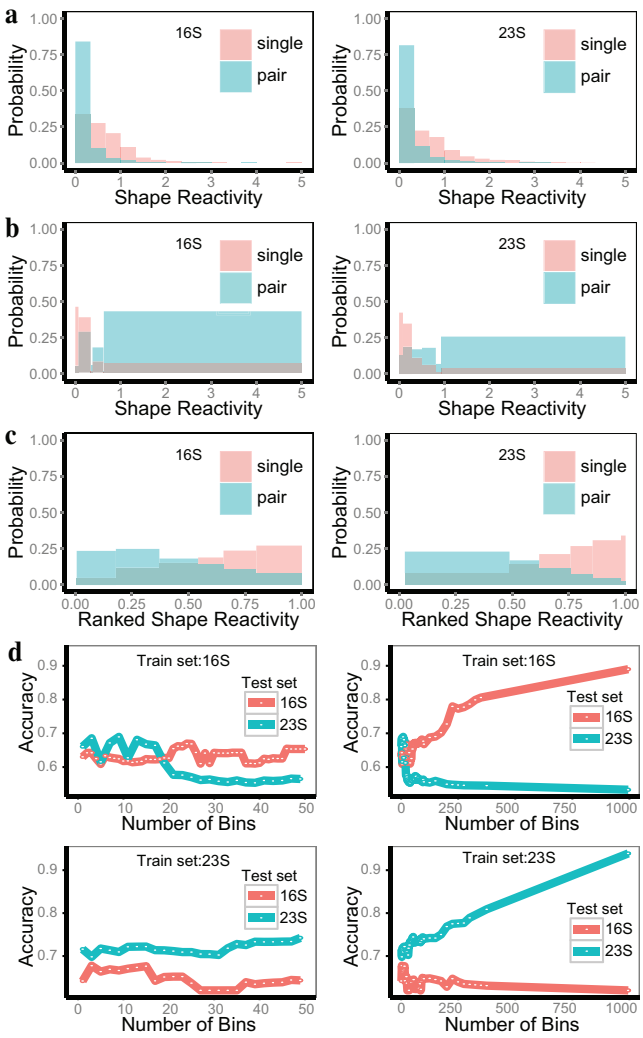
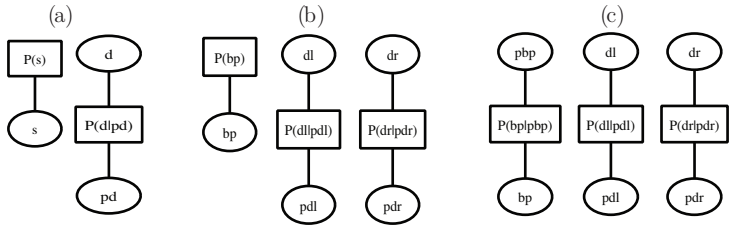


Figure S1 – *E.coli* 16S and 23S SHAPE reactivity distributions and effect of overfitting. (a) Histograms of SHAPE reactivity at paired (blue) and unpaired (red) positions made using 15 equidistant bins. (b) Same as in (a), using six bins with break-points found by optimizing the Kullback Leibler divergence between the *single* and *pair* distributions. (c) Same as in (b), showing rank normalized reactivities. (d) ProbFold (*s*tack+sh model) prediction performance depends on the number of bins used to model the rank normalized SHAPE distributions. The left panels show the performance for up to 50 bins and the rights panels for up to 1,000 bins. ProbFold was separately trained on 16S (top) and 23 (bottom) and evaluted on both 16S (red) and 23S (blue). When trained and tested on different data (cross-evaluation), the performance peaks at a low number (3-15) of bins. When trained and tested on the same sequence the performance increases with the number of bins, which illustrates the effect of overfitting the distributions.

s: Single Nuc
d: Probing data
pd: Previous d
bp: Base pair
dl: d Left
dr: d Right
pdl: Previous d Left
pdr: Previous d Right



(d) Emit Models

	single	pair	stack
Joint prob	P(s,d,pd)	P(bp,dl,dr,pdl,pdr)	P(bp,pbp,dl,dr,pdl,pdr)
Factorization	P(s)P(d)P(d|pd)	P(bp)P(dl)P(dr) P(dl|pdl)P(dr|pdr)	P(bp|pbp)P(dl)P(dr) P(dl|pdl)P(dr|pdr)
Sequence Probing data	ACGUCACGU 001232454	ACGUCACGU 001232454	ACGUCACGU 001232454
Mapping	s ⟵ G, d⟵ 1 pd⟵0	bp⟵AU dl⟵0, dr ⟵ 4 pdl⟵ 0, pdr ⟵ 5	bp ⟵CG , pbp ⟵ AU dl ⟵ 0, dr ⟵ 5 pdl ⟵ 0, pdr ⟵ 4

Figure S 2 – Probabilistic graphical models that include correlations in SHAPE data defining the (a)*single*, (b) *pair*, and (c) *stack* emission models that include correlation along the sequence. The variable abbreviations are given to the left. (d) For each emission model, the table gives (i) the joint probability distribution; (ii) its factorization specified by the PGM; (iii) example of short input data sequence with potential input positions highlighted. The probing data is discretized into six bins and enumerated (0-5); (iv) mapping of data from highlighted sequence positions to relevant random variables of PGM.

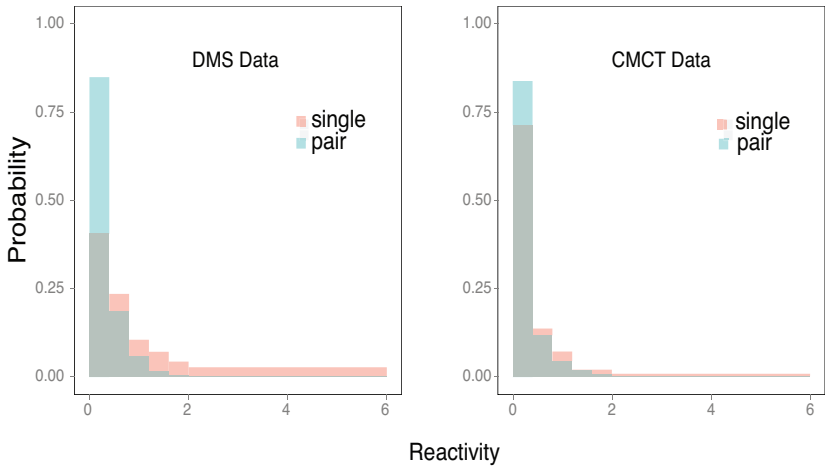


Figure S3 – Distribution of probing signal intensity for base-paired (pair) and unpaired (single) positions for DMS (left) and CMCT data (right)

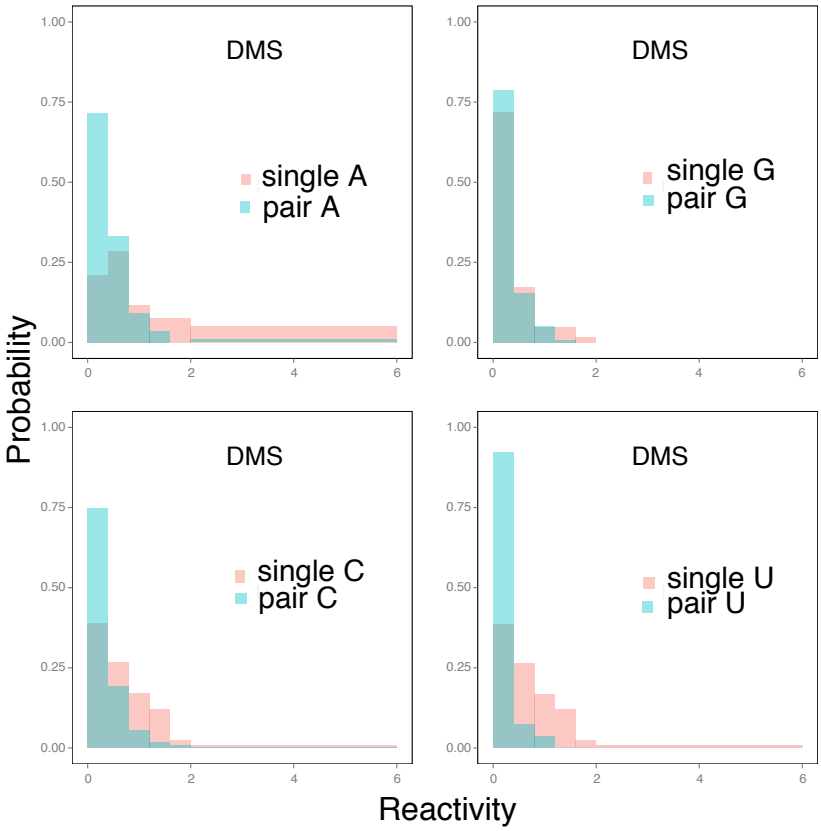


Figure S4 – Distribution of probing signal intensity for base-paired (pair) and unpaired (single) positions conditional on nucleotide of primary sequence for DMS data.

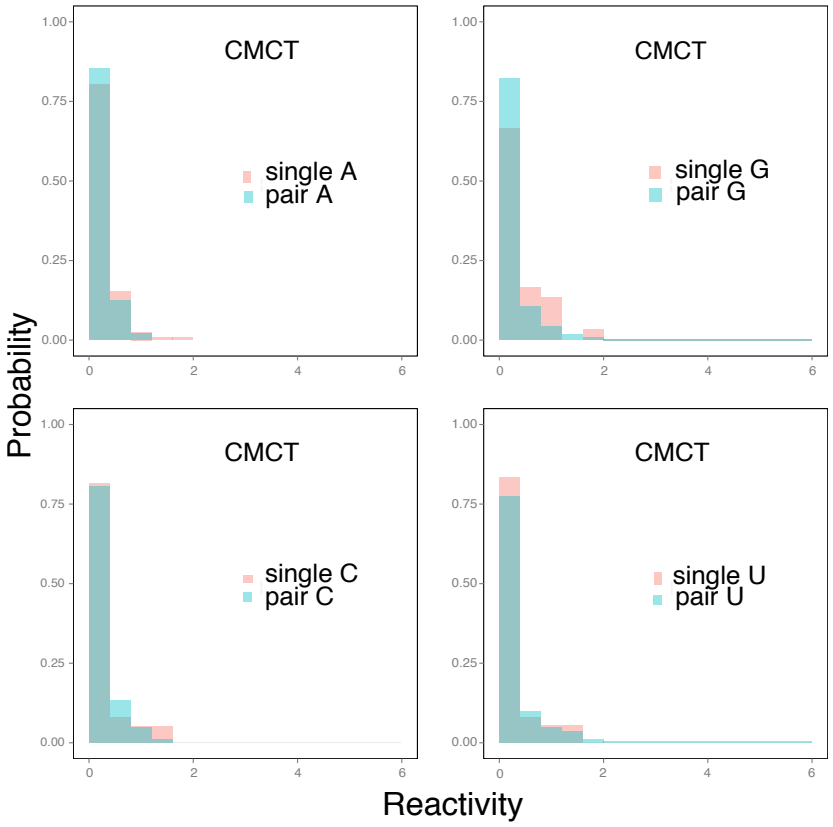


Figure S5 – Distribution of probing signal intensity for base-paired (pair) and unpaired (single) positions conditional on nucleotide of primary sequence for CMCT data.

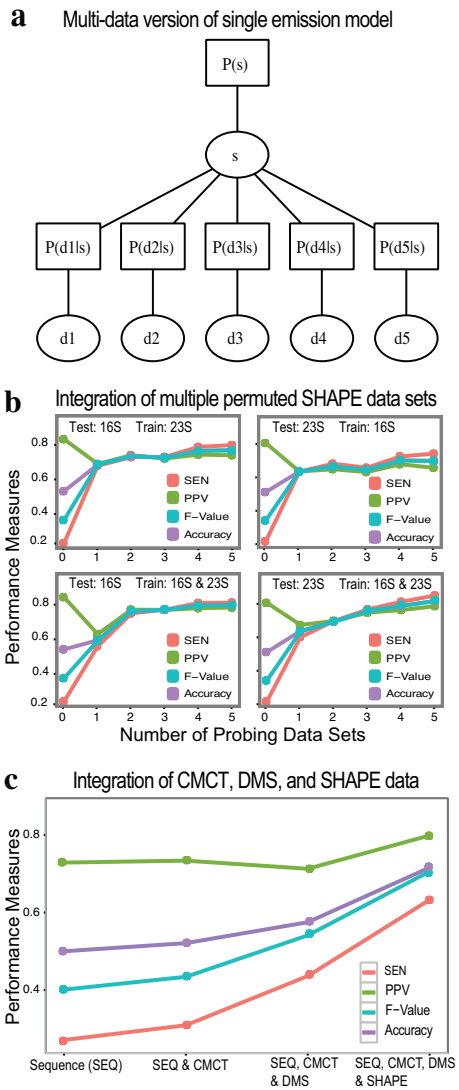


Figure S6 – Multi-data version of ProbFold. (a) PGM for *single* emission distribution that integrates five types of probing data, d_1, d_2, d_3, d_4 and d_5. $P(s)$ is the prior distribution of sequence data, $P(d_i|s), i = 1..5$, are the conditional distributions of the different types of probing data given the sequence data. (b) Prediction performance measured by sensitivity (SEN), positive predictive value (PPV), and F-value, and accuracy for E. coli 16S and 23S rRNAs given varying number of boot-strapped SHAPE probing data sets (0 to 5). 0: nucleotide data only, 1: nucleotide data and a single permuted SHAPE data set, 2: nucleotide data and two permuted SHAPE data sets, etc.. (c) Prediction performance of the multi-data version of ProbFold with incrementally increasing number of probing data types (CMCT, DMS, and SHAPE) on six small RNA structures. Performance evaluated using cross evaluation with same measures as above.

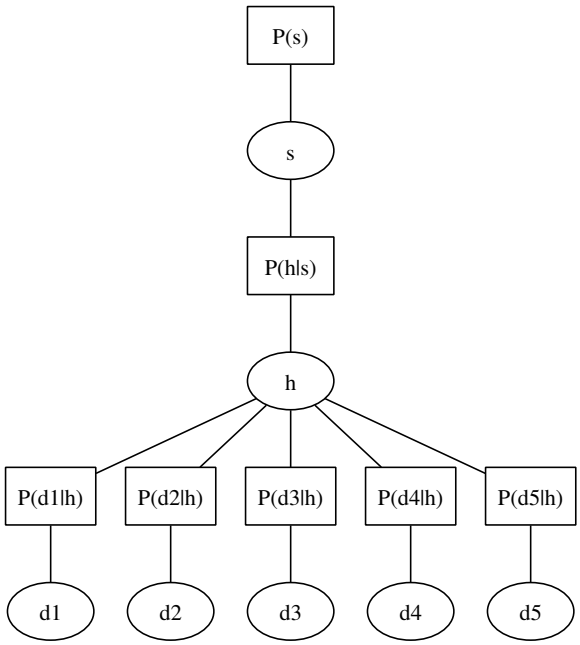


Figure S7 – Multi-data version of the single emission model to account for dependencies both within and among multiple data types (d_1, d_2, d_3, d_4 and d_5). $P(s)$ is the prior distribution of sequence data, $P(h|s)$ is conditional probability of the hidden variable h given sequence data. $P(d_{i,i=1..5}|h)$ are the conditional probabilities of the different types of probing data given the hidden variable. The introduction of a hidden variable in the model allows it to capture the dependencies among the different types of probing data.

Table S1 – Number of free parameters in the different models

Model	Number of free parameters
pair	18
stack	258
pair+sh	33
stack+sh	283
stack+sh+cor	408
stack+sh (3 bins)	268
stack+sh+cor (3 bins)	288

Table S 2 – Prediction performance on *E. coli* 16S and 23S rRNAs secondary structure for different ProbFold models (using $\gamma = 1$) as well as the RNAstructure [2], PPfold [1], and GTFold methods [3]. TA: TrainSet A; sh: SHAPE data; cor: correlation; +sh: SHAPE version of method; n.a.: information not available; '-': information not applicable.

Model	Train set	Test set	PPV	SEN	F-value	ACC	AUC	AUC^{label}
stack	TA	16S	**0.84**	0.24	0.37	**0.54**	0.38	0.55
stack	TA	23S	**0.77**	0.25	0.37	**0.51**	0.38	0.59
stack+sh	TA+23S+RMDB	16S	0.76	0.63	0.69	0.69	0.61	0.79
stack+sh+cor	TA+23S+RMDB	16S	**0.82**	0.59	0.68	0.70	**0.64**	**0.80**
stack+sh	TA+16S+RMDB	23S	**0.90**	0.49	0.63	0.69	**0.64**	**0.78**
stack+sh+cor	TA+16S+RMDB	23S	0.85	0.47	0.60	0.66	0.57	0.74
RNAstructure	-	16S	0.34	**0.36**	0.35	0.35	-	-
PPfold	-	16S	0.64	0.22	0.33	0.43	-	-
GTFold	-	16S	0.40	0.42	**0.41**	0.41	-	-
RNAstructure+sh	23S	16S	0.75	**0.78**	**0.76**	**0.77**	-	-
PPfold+sh	16S+23S	16S	0.76	0.68	0.71	0.72	-	-
GTFold+sh	n.a.	16S	0.75	0.77	**0.76**	0.76	-	-
RNAstructure	-	23S	0.48	**0.52**	**0.50**	0.50	-	-
PPfold	-	23S	0.73	0.22	0.34	0.48	-	-
GTFold	-	23S	0.47	0.51	0.49	0.49	-	-
RNAstructure+sh	23S	23S	0.74	**0.76**	**0.75**	**0.75**	-	-
PPfold+sh	16S+23S	23S	0.70	0.59	0.64	0.65	-	-
GTFold+sh	n.a.	23S	0.65	0.68	0.67	0.67	-	-

Table S3 – Comparison of secondary structure prediction performance (PPV and SEN) of ProbFold, PPfold [1], RNAStructure [2], and GTFold [3] methods on small RNA structures, not using probing data (sequence-only predictions).

Structures	ProbFold		PPFold		RNAStructure		GTFold	
	PPV	SEN	*PPV*	SEN	*PPV*	SEN	*PPV*	SEN
5S RNA	0.360	0.264	***0.499***	**0.265**	0.220	**0.265**	0.231	**0.265**
Adenine riboswitch	0.000	0.000	***1.000***	0.286	***1.000***	**1.000**	***1.000***	**1.000**
cidGMP riboswitch	***1.000***	0.320	***1.000***	0.286	0.800	**0.960**	0.618	0.840
Glycin	***1.000***	0.375	***1.000***	0.200	0.609	**0.700**	0.619	0.650
P4-P6 domain (Tetrahymena ribozyme)	***1.000***	0.333	***1.000***	0.333	0.830	**0.917**	0.238	0.250
Ribonuclease P	0.900	0.537	***0.946***	0.522	0.603	0.567	0.831	**0.731**
tRNA phenylalanine (yeast)	***1.000***	0.100	***1.000***	0.333	0.950	**0.950**	0.712	0.833
M-Box riboswitch	***1.000***	0.686	***1.000***	0.702	0.739	0.354	0.913	**0.875**
Lysine riboswitch	***0.319***	0.230	0.313	**0.266**	0.237	0.215	0.271	0.246
Group I Intron, T. thermophila	0.911	0.394	***0.984***	0.462	0.649	**0.727**	0.583	0.614
Group II Intron, O. iheyensis	0.556	0.150	***0.857***	0.181	0.708	**0.639**	0.647	0.564

Table S4 – Comparison of secondary structure prediction performance (F-Values and Accuracy) of different models on small RNA structures, not using probing data (sequence-only predictions).

Structures	ProbFold		PPFold		RNAStructure		GTFold	
	F-Value	ACC	*F-Value*	ACC	*F-Value*	ACC	*F-Value*	ACC
5S RNA	0.305	0.312	***0.346***	**0.382**	0.240	0.242	0.247	0.248
Adenine riboswitch	0.000	0.000	0.444	0.643	***1.000***	**1.000**	***1.000***	**1.000**
cidGMP riboswitch	0.485	0.660	0.333	0.600	***0.873***	**0.880**	0.712	0.729
Glycin	0.545	**0.687**	0.182	0.550	***0.651***	0.654	0.634	0.635
P4-P6 domain (Tetrahymena ribozyme)	0.500	0.667	0.499	0.667	***0.871***	0.873	0.762	0.768
Ribonuclease P	0.673	0.719	0.673	0.734	0.585	0.585	***0.778***	**0.781**
tRNA phenylalanine (yeast)	0.182	0.550	0.320	0.500	***0.950***	**0.950**	0.243	0.244
M-Box riboswitch	0.815	0.844	0.720	0.771	0.479	0.547	***0.894***	**0.894**
Lysine riboswitch	***0.268***	**0.275**	0.266	0.272	0.226	0.226	0.258	0.259
Group I Intron, T. thermpphila	0.562	0.686	0.629	0.723	***0.686***	**0.688**	0.598	0.598
Group II Intron, O. iheyensis	0.237	0.353	0.298	0.519	***0.672***	**0.674**	0.602	0.605

Table S5 – Positive predictive value (PPV) of different methods on a set of small RNA structures using SHAPE Data.

Species	ProbFold	PPFold	RNAStructure	GTFold
		PPV		
5S RNA	0.586	**0.654**	0.225	0.237
Adenine riboswitch	**1.000**	0.913	0.913	**1.000**
cidGMP riboswitch	0.652	0.632	**0.667**	0.625
Glycin	0.824	0.650	**0.841**	0.800
P4-P6 domain (Tetrahymena ribozyme)	0.843	**1.000**	0.846	0.782
Ribonuclease P	**0.855**	0.171	0.576	0.382
tRNA phenylalanine (yeast)	**0.952**	0.400	**0.952**	0.909
M-Box riboswitch	**0.833**	0.541	0.690	0.762
Lysine riboswitch	**0.366**	0.297	0.304	0.286
Group I Intron, T. thermpphila	0.832	**0.875**	0.786	0.776
Group II Intron, O. iheyensis	0.554	0.594	**0.626**	0.603
Average	**0.754**	0.612	0.675	0.651

Table S 6 – Comparison of ΔPPV performance contributions for different methods on set of small RNA structures when including SHAPE data (ΔPPV = [PPV with probing data included] − [PPV of sequence only]).

Structure	ProbFold	PPFold	RNAStructure	GTFold
5S RNA	0.226	**0.245**	-0.040	-0.028
Adenine riboswitch	**1.000**	0.087	-0.087	0.000
cidGMP riboswitch	-0.348	**0.368**	-0.133	0.007
Glycin	-0.176	**0.350**	0.232	0.181
P4-P6 domain (Tetrahymena ribozyme)	-0.157	0.000	0.016	**0.080**
Ribonuclease P	-0.045	-0.774	**-0.027**	-0.449
tRNA phenylalanine (yeast)	-0.048	0.400	0.002	**0.671**
M-Box riboswitch	-0.167	-0.459	**-0.049**	-0.151
Lysine riboswitch	0.047	-0.016	**0.067**	0.015
Group I Intron, T. thermpphila	-0.079	-0.198	0.137	**0.195**
Group II Intron, O. iheyensis	**-0.002**	-0.263	-0.082	-0.044
Average	0.023	-0.022	-0.029	**0.043**

Table S7 – Sensitivity (SEN) of different methods on a set of small RNAs structures using SHAPE Data.

Species	ProbFold	PPFold	RNAStructure	GTFold
		SEN		
5S RNA	**0.500**	**0.500**	0.265	0.265
Adenine riboswitch	**1.000**	**1.000**	**1.000**	**1.000**
cidGMP riboswitch	0.600	0.480	**0.800**	**0.800**
Glycin	0.700	0.650	**0.925**	0.900
P4-P6 domain (Tetrahymena ribozyme)	**0.917**	0.677	**0.917**	0.896
Ribonuclease P	**0.702**	0.224	0.567	0.388
tRNA phenylalanine (yeast)	**1.000**	0.500	**1.00**	**1.000**
M-Box riboswitch	0.625	0.417	0.417	**0.667**
Lysine riboswitch	0.230	0.169	0.262	**0.231**
Group I Intron, T. thermpphila	0.750	0.530	**0.780**	0.735
Group II Intron, O. iheyensis	0.466	0.474	**0.579**	0.571
Average	0.681	0.511	0.683	**0.678**

Table S 8 – Comparison of ΔSEN performance contributions for different methods on set of small RNA structures when including SHAPE data (ΔSEN = [SEN with probing data included] − [SEN of sequence only]).

Structure	ProbFold	PPFold	RNAStructure	GTFold
5S RNA	**0.237**	0.235	0.000	0.000
Adenine riboswitch	**1.000**	0.714	0.000	0.000
cidGMP riboswitch	**0.306**	0.280	-0.160	-0.040
Glycin	0.387	**0.550**	0.225	0.250
P4-P6 domain (Tetrahymena ribozyme)	**0.536**	0.344	0.000	0.063
Ribonuclease P	**0.241**	-0.298	0.000	-0.343
tRNA phenylalanine (yeast)	**0.875**	0.300	0.050	0.705
M-Box riboswitch	-0.061	-0.285	0.064	-0.208
Lysine riboswitch	0.000	-0.097	0.047	-0.015
Group I Intron, T. thermpphila	**0.356**	0.068	0.053	0.121
Group II Intron, O. iheyensis	**0.316**	0.293	-0.060	0.007
Average	**0.381**	0.191	0.020	0.049

Table S 9 – Accuracy (ACC) of different methods on a set of small RNA structures using SHAPE Data.

Species	ProbFold	PPFold	RNAStructure	GTFold
		ACCURACY		
5S RNA	0.543	**0.577**	0.246	0.251
Adenine riboswitch	**1.000**	0.957	0.957	**1.000**
cidGMP riboswitch	0.626	0.556	**0.734**	0.713
Glycin	0.762	0.650	**0.883**	0.850
P4-P6 domain (Tetrahymena ribozyme)	0.869	0.833	**0.882**	0.839
Ribonuclease P	**0.778**	0.198	0.572	0.385
tRNA phenylalanine (yeast)	**0.976**	0.450	**0.976**	0.955
M-Box riboswitch	**0.729**	0.479	0.553	0.714
Lysine riboswitch	**0.298**	0.233	0.283	0.259
Group I Intron, T. thermpphila	**0.791**	0.703	0.783	0.755
Group II Intron, O. iheyensis	0.509	0.534	**0.602**	0.587
Average	**0.716**	0.561	0.679	0.664

Table S 10 – Comparison of ΔACC performance contributions for different methods on set of small RNA structures when including SHAPE data (ΔACC = [ACC with probing data included] − [ACC of sequence only]).

Structure	ProbFold	PPFold	RNAStructure	GTFold
5S RNA	**0.231**	0.195	0.004	0.003
Adenine riboswitch	**1.000**	0.314	0.043	0.000
cidGMP riboswitch	**0.129**	-0.044	-0.146	-0.016
Glycin	0.075	0.100	**0.229**	0.215
P4-P6 domain (Tetrahymena ribozyme)	**0.202**	0.166	0.009	0.071
Ribonuclease P	**0.059**	-0.545	-0.013	-0.396
tRNA phenylalanine (yeast)	0.426	-0.100	0.026	**0.711**
M-Box riboswitch	-0.115	-0.292	**0.006**	-0.180
Lysine riboswitch	0.023	-0.039	**0.057**	0.000
Group I Intron, T. thermpphila	0.105	-0.020	0.095	**0.157**
Group II Intron, O. iheyensis	**0.156**	0.015	-0.072	-0.018
Average	**0.208**	-0.023	0.022	0.050

Table S 11 – ProbFold performance on set of small RNA structures using DMS Data.

Species	PPV	SEN	F-Value	ACC
5S RNA	0.310	0.265	0.286	0.287
Adenine riboswitch	0.500	0.381	0.432	0.441
cidGMP riboswitch	0.923	0.480	0.632	0.702
Glycin	0.733	0.550	0.629	0.630
P4-P6 domain (Tetrahymena ribozyme)	0.864	0.396	0.5428	0.630
tRNA phenylalanine (yeast)	1.000	0.800	0.889	0.900
Average	0.722	0.479	0.544	0.584

Table S12 – ProbFold performance on a set of small RNA structures using CMCT Data.

Species	PPV	SEN	F-Value	ACC
5S RNA	0.346	0.265	0.300	0.305
Adenine riboswitch	0.444	0.191	0.267	0.318
cidGMP riboswitch	1.000	0.480	0.649	0.740
Glycin	0.940	0.450	0.610	0.699
P4-P6 domain (Tetrahymena ribozyme)	1.000	0.354	0.523	0.677
tRNA phenylalanine (yeast)	1.000	0.350	0.519	0.675
Average	0.788	0.348	0.478	0.569

Table S13 – Average performance on six small structural RNAs of the Multi-data version of ProbFold with step-wise inclusion of of CMCT, DMS and SHAPE structure probing data set, starting with sequence-only (Seq-only). The performance of the Multi-data version of RNAStructure [4] is shown for comparison.

Models	Data Type	PPV	SEN	F-Value	ACC
ProbFold	Seq-only	0.729	0.271	0.401	0.500
	Seq, CMCT	0.734	0.309	0.434	0.521
	Seq, CMCT, DMS	0.713	0.436	0.541	0.575
	Seq, CMCT, DMS, SHAPE	0.799	0.633	0.706	0.716
RNAStructure	Seq-only	0.687	0.771	0.727	0.729
	Seq, CMCT	0.818	0.883	0.849	0.851
	Seq, CMCT, DMS	0.816	0.894	0.853	0.855
	Seq, CMCT, DMS, SHAPE	0.795	0.846	0.820	0.821

References

[1] Sükösd,Z., Knudsen,B., Kjems,J. and Perdesen,C. (2012) PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, **28**, 2691-2692.

[2] Mathews,D.H., Dinsey,D.M., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *PNAS*, **101**, 7287-7292.

[3] Swenson,M.S.,Anderson, J.,Ash, A.,Gaurav, P.,Sükösd,Z., Bader,D.A., Harvey,S.C. and Heitsch,C.E. (2012) GTfold: enabling parallel RNA secondary structure prediction on multi-core desktops. *BMC Res. Notes*, **5**, 341.

[4] Cordero,P., Kladwang,W., VanLang,C.C. and Das,R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, **51**, 7037.

[5] Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad−Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and Classification of Conserved RNA Secondary Structures in the Human Genome, *PLoS Comput. Biol.*, **2**, e33.

[6] Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press.

[7] Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, **39**, 138.

[8] Bishop,M.C. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc. Secaucus, NJ, USA.

[9] Knudsen,B. and Hein,J.J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars *Nucleic Acids Res.*, **13**, 3423-3428.

[10] Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models, *Bioinformatics*, **22**, e90-e98.

[11] Nussinov,R. and Jacobson,A.B. (1980) Fast Algorithm for Predicting the Secondary Structure of Single Stranded RNA, *Proceedings of the National Academy of Sciences, USA*. **77**, 6309-6313.

[12] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (2007) *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press.

Declarations of co-authorship

AARHUS UNIVERSITY

Declaration of co-authorship

Full name of the PhD student: Michał Piotr Świtnicki

This declaration concerns the following article/manuscript:

Title:	PINCAGE: Probabilistic integration of cancer genomics data for perturbed gene identification and sample classification
Authors:	Michał P. Świtnicki, Malene Juul, Tobias Madsen, Karina D. Sørensen, Jakob S. Pedersen

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: Bioinformatics

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. No or little contribution
B. Has contributed (10-30 %)
C. Has contributed considerably (40-60 %)
D. Has done most of the work (70-90 %)
E. Has essentially done all the work

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	C
2. Planning of the experiments and methodology design and development	C
3. Involvement in the experimental work/clinical studies/data collection	E
4. Interpretation of the results	D
5. Writing of the first draft of the manuscript	D
6. Finalization of the manuscript and submission	D

Signatures of the co-authors

Date	Name	Signature
6/7/15	Malene Juul	
6/7/15	Tobias Madsen	
6/7/15	Karina D. Sørensen	
6/7/15	Jakob S. Pedersen	

1 of 1

AARHUS
UNIVERSITY

In case of further co-authors please attach appendix

Date: 6/7/15

Świtnicki Michał

Signature of the PhD student

AARHUS UNIVERSITY

Declaration of co-authorship

Full name of the PhD student: Michał Piotr Świtnicki

This declaration concerns the following article/manuscript:

Title:	Sample classification using a parameter-sparse probabilistic graphical model for integration of cancer genomics data
Authors:	Michał P. Świtnicki, Tobias Madsen, Jakob S. Pedersen

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☐ In preparation ☒

If published, state full reference:

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

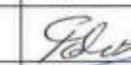
No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. No or little contribution
B. Has contributed (10-30 %)
C. Has contributed considerably (40-60 %)
D. Has done most of the work (70-90 %)
E. Has essentially done all the work

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	C
2. Planning of the experiments and methodology design and development	C
3. Involvement in the experimental work/clinical studies/data collection	D
4. Interpretation of the results	D
5. Writing of the first draft of the manuscript	C
6. Finalization of the manuscript and submission	D

Signatures of the co-authors

Date	Name	Signature
6/7/15	Tobias Madsen	
6/7/15	Jakob S. Pedersen	

1 of 1

AARHUS
UNIVERSITY

In case of further co-authors please attach appendix

Date: 6/7/15

Świtnicki Michał

Signature of the PhD student

AARHUS UNIVERSITY

Declaration of co-authorship

Full name of the PhD student: Michał Piotr Świtnicki

This declaration concerns the following article/manuscript:

Title:	ProbFold: A probabilistic method for integration of probing data in RNA secondary structure prediction
Authors:	Sudhakar Sahoo, Michal P. Świtnicki, Jakob S. Pedersen

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: Bioinformatics

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. No or little contribution
B. Has contributed (10-30 %)
C. Has contributed considerably (40-60 %)
D. Has done most of the work (70-90 %)
E. Has essentially done all the work

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments and methodology design and development	B
3. Involvement in the experimental work/clinical studies/data collection	B
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	B

Signatures of the co-authors

Date	Name	Signature
4/7/15	Sudhakar Sahoo	Sudhakar Sahoo
6/7/15	Jakob S. Pedersen	

AARHUS
UNIVERSITY

In case of further co-authors please attach appendix

Date: 6/7/15

Signature of the PhD student

Reference list

Ahn, J.Y.L., J. T. (2008) X Chromosome: X Inactivation. *Nature Education*.

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

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data, *Bioinformatics*, **31**, 166-169.

Atkins, J.F., Gesteland, R.F. and Cech, T. (2011) *RNA worlds : from life's origins to diversity in gene regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing, *J Roy Stat Soc B Met*, **57**, 289-300.

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

Bishop, C.M. (2006) *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Borza, T., Konijeti, R. and Kibel, A.S. (2013) Early detection, PSA screening, and management of overdiagnosis, *Hematology/oncology clinics of North America*, **27**, 1091-1110, vii.

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

Carpenter, J. and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, *Statistics in Medicine*, **19**, 1141-1164.

Carreiro, A.V.*, et al.* (2015) Integrative biomarker discovery in neurodegenerative diseases, *Wiley interdisciplinary reviews. Systems biology and medicine*.

Cech, T.R. and Bass, B.L. (1986) Biological catalysis by RNA, *Annual review of biochemistry*, **55**, 599-629.

Chan, C.M.*, et al.* (2013) A signature motif mediating selective interactions of BCL11A with the NR2E/F subfamily of orphan nuclear receptors, *Nucleic Acids Res*, **41**, 9663-9679.

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

Cordero, P.*, et al.* (2012) Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference, *Biochemistry-Us*, **51**, 7037-7039.

Cox, D.R. and Oakes, D. (1984) *Analysis of survival data*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York.

Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 14664-14669.

Dedeurwaerder, S.*, et al.* (2011) Evaluation of the Infinium Methylation 450K technology, *Epigenomics*, **3**, 771-784.

Deigan, K.E.*, et al.* (2009) Accurate SHAPE-directed RNA structure determination, *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 97-102.

Delong, E.R., Delong, D.M. and Clarkepearson, D.I. (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.

Durbin, R. (1998) *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK New York.

Fawcett, T. (2004) ROC graphs: Notes and practical considerations for researchers, *ReCALL*, **31**, 1--38.

Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861-874.

Felgueiras, J., Silva, J.V. and Fardilha, M. (2014) Prostate cancer: the need for biomarkers and new therapeutic targets, *Journal of Zhejiang University. Science. B*, **15**, 16-42.

Fisher, R.A. (1938) *Statistical methods for research workers*. Biological monographs and manuals,. Oliver and Boyd, Edinburgh,.

Gelfman, S.*, et al.* (2013) DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure, *Genome Res*, **23**, 789-799.

Gesing, A.*, et al.* (2011) Decreased expression level of apoptosis-related genes and/or proteins in skeletal muscles, but not in hearts, of growth hormone receptor knockout mice, *Experimental biology and medicine*, **236**, 156-168.

Gower, J.C. (1966) Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis, *Biometrika*, **53**, 325-&.

Grate, L. (1995) Automatic RNA secondary structure determination with stochastic context-free grammars, *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **3**, 136-144.

Gupta, M.R. and Chen, Y. (2010) Theory and Use of the EM Algorithm, *Foundations and Trends® in Signal Processing*, **4**, 223-296.

Haas, G.P.*, et al.* (2008) The worldwide epidemiology of prostate cancer: perspectives from autopsy studies, *The Canadian journal of urology*, **15**, 3866-3871.

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

Hanczar, B.*, et al.* (2010) Small-sample precision of ROC-related estimates, *Bioinformatics*, **26**, 822-830.

Hand, D. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Mach Learn*, **77**, 103-123.

Hand, D.J. (2012) Assessing the Performance of Classification Methods, *International Statistical Review*, **80**, 400-414.

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29-36.

Hartigan, J.A. (1975) *Clustering algorithms*. Wiley series in probability and mathematical statistics. Wiley, New York,.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) The elements of statistical learning : data mining, inference, and prediction. In, *Springer series in statistics,*. Springer, New York, NY, pp. xxii, 745 p.

Illumina, I. (2015) Paired-end sequencing.

Jeanmougin, M.*, et al.* (2010) Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies, *Plos One*, **5**, e12336.

Kalia, M. (2015) Biomarkers for personalized oncology: recent advances and future challenges, *Metabolism: clinical and experimental*, **64**, S16-21.

Koller, D. and Friedman, N. (2009) *Probabilistic graphical models : principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

Koller, D.F., Nir, Getoor, L. and Taskar, B. (2007) Graphical Models in a Nutshell. In, *Introduction to Statistical Relational Learning*. MIT Press.

Kopeikin, Z. and Chen, S.J. (2005) Statistical thermodynamics for chain molecules with simple RNA tertiary contacts, *The Journal of chemical physics*, **122**, 094909.

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

Kruskal, W.H. and Wallis, W.A. (1952) Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association*, **47**, 583-621.

Kullback, S.L., Richard A. (1951) On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86.

Kurotaki, N.*, et al.* (2005) Sotos syndrome common deletion is mediated by directly oriented subunits within inverted Sos-REP low-copy repeats, *Hum Mol Genet*.

Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC bioinformatics*, **12**.

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

Ling, C.X., Huang, J. and Zhang, H. (2003) AUC: a better measure than accuracy in comparing learning algorithms. *Proceedings of the 16th Canadian society for computational studies of intelligence conference on Advances in artificial intelligence*. Springer-Verlag, Halifax, Canada, pp. 329-341.

Ling, C.X., Huang, J. and Zhang, H. (2003) AUC: a statistically consistent and more discriminating measure than accuracy. *Proceedings of the 18th international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc., Acapulco, Mexico, pp. 519-524.

Lobo, J.M., Jiménez-Valverde, A. and Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models, *Global Ecology and Biogeography*, **17**, 145-151.

Lorenz, R.*, et al.* (2013) 2D meets 4G: G-quadruplexes in RNA secondary structure prediction, *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, **10**, 832-844.

Mann, H.B. and Whitney, D.R. (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, 50-60.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate analysis*. Probability and mathematical statistics. Academic Press, London ; New York.

Marioni, J.C.*, et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome Res*, **18**, 1509-1517.

Mathews, D.H.*, et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7287-7292.

Mathews, D.H., Moss, W.N. and Turner, D.H. (2010) Folding and finding RNA secondary structure, *Cold Spring Harbor perspectives in biology*, **2**, a003665.

McGettigan, P.A. (2013) Transcriptomics in the RNA-seq era, *Current opinion in chemical biology*, **17**, 4-11.

Morris, T.J.*, et al.* (2014) ChAMP: 450k Chip Analysis Methylation Pipeline, *Bioinformatics*, **30**, 428-430.

Neyman, J. (1933) On the problem of the most efficient tests of statistical hypotheses, *Philos T R Soc Lond*, **231**, 289-337.

Nieschlag, E*., et al.* (2014) New approaches to the Klinefelter syndrome, *Annales d'endocrinologie*, **75**, 88-97.

Ong, C.T. and Corces, V.G. (2012) Enhancers: emerging roles in cell fate specification, *EMBO reports*, **13**, 423-430.

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

Poulsen, L.D*., et al.* (2015) SHAPE Selection (SHAPES) enrich for RNA structure signal in SHAPE sequencing-based probing data, *Rna*, **21**, 1042-1052.

Rajkowitsch, L*., et al.* (2007) RNA chaperones, RNA annealers and RNA helicases, *RNA biology*, **4**, 118-130.

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 research*, **72**, 1170-1181.

Rice, G.M., Leonard, C.W. and Weeks, K.M. (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE, *Rna*, **20**, 846-854.

Rich, J.T*., et al.* (2010) A practical guide to understanding Kaplan-Meier curves, *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, **143**, 331-336.

Ritchie, M.E*., et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Res*, **43**, e47.

Robin, X*., et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC bioinformatics*, **12**, 77.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (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.

Sethi, S*., et al.* (2013) Clinical advances in molecular biomarkers for cancer diagnosis and therapy, *International journal of molecular sciences*, **14**, 14771-14784.

Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical applications in genetics and molecular biology*, **3**, Article3.

Sprinthall, R.C. (2012) *Basic statistical analysis*. Pearson Allyn & Bacon, Boston.

Stephen, J.K.*, et al.* (2013) Significance of p16 in Site-specific HPV Positive and HPV Negative Head and Neck Squamous Cell Carcinoma, *Cancer and clinical oncology*, **2**, 51-61.

Świtnicki, M.P.*, et al.* (2015) PINCAGE: Probabilistic integration of cancer genomics data for perturbed gene identification and sample classification.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105-1111.

Trapnell, C.*, et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature protocols*, **7**, 562-578.

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

Van Den Eynde, B.J.*, et al.* (1999) A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription, *The Journal of experimental medicine*, **190**, 1793-1800.

Venables, W.N., Ripley, B.D. and Venables, W.N. (2002) *Modern applied statistics with S*. Statistics and computing. Springer, New York.

Venkatraman, E.S. (2000) A Permutation Test to Compare Receiver Operating Characteristic Curves, *Biometrics*, **56**, 1134-1138.

Venkatraman, E.S. and Begg, C.B. (1996) A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment, *Biometrika*, **83**, 835-848.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature reviews. Genetics*, **10**, 57-63.

Weeks, K.M. (2012) 16S ans 23S E. coli data. *Personal Communication.*

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

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, **1**, 80-83.

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.