Network Models of Stochastic Processes in Cancer

Thomas Emlyn Bartlett

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Mathematics

University College London

September 19, 2015

I, Thomas Emlyn Bartlett, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Complex systems which can be modelled as networks are ubiquitous. Well-known examples include social and economic networks, as well as many examples in cell biology such as gene regulatory and protein signalling networks. Many cell biological processes are inherently stochastic and non-stationary, and this is the perspective from which I have developed novel mathematical and computational statistical models, focusing particularly on network models. These models are primarily motivated by cell biological processes relating to DNA methylation and stem cell and cancer biology, but can be generalised to other systems and domains. I have used these and other models to identify and analyse novel DNA-based cancer biomarkers.

Acknowledgements

My family: David, Denise and William Bartlett.

Melissa Panlasigui, for support and helpful comments about the work.

My supervisors: Sofia Olhede, Martin Widschwendter and Alexey Zaikin.

Everyone at UCL CoMPLEX, and particularly Geraint Thomas, Guy Moss and Lewis Griffin.

Those who funded me for this work via UCL CoMPLEX: ESPRC, MRC.

All specimen donors and research groups involved in providing the data used in this work, and all those who contributed to the open-source software used in this work.

Contents

List of Figures

List of Tables

Chapter 1

Introductory Material

1.1 Background to the work

1.1.1 Epigenomics and DNA methylation

Epigenetic information is stored in the genome in the form of heritable modifications to the chemical structure of DNA, such as methylation of particular bases, as well a variety of chemical modifications of the histone proteins which package the DNA. This epigenetic information is changed much more easily and more often than the genetic information contained in the sequence of nucleotides: genetic and epigenetic information have been likened to 'nature's pen and pencil set' (Gosden & Feinberg, 2007). Epigenetic information can be modulated during the lifetime of an organism by, for example, diet and other environmental cues (Teschendorff *et al.* , 2009) and these changes persist in subsequent mitosis, leading to an acquired change of phenotype.

DNA methylation is an epigenetic mark consisting almost entirely of the methylation of CpG dinucleotides (Bernstein *et al.* , 2007), and most CpGs in the genome are methylated (Bird, 2002). It is possible for one, both, or neither alleles at a particular genomic locus to be methylated (Li *et al.* , 2010), and methylation states of specific loci and alleles are propagated in mitotic cell division via 'maintenance' methyltransferase DNMT1 (Bernstein *et al.* , 2007). The methylation of CpGs in the promoter region of a gene is associated with a silencing of that gene, and this effect is particularly important in cancer, where such aberrant gene silencing is associated with functional changes important in every stage of tumour progression (Jones & Baylin, 2002).

CpGs tend to cluster together in short regions of around 1kb (Jones, 2012) with high C, G and CpG densities, termed CpG islands (CGI), and CpGs in these regions tend to be hypomethylated relative to the methylation level of CpGs outside. In humans, about 60% of gene promoter regions are associated with CGIs (Bernstein *et al.* , 2007). Hypermethylation of CpGs in the

gene promoter, i.e., the region close to the transcriptional start site (TSS), are incontrovertibly associated with silencing of the corresponding gene (Jones, 2012), although gene silencing associated with hypermethylation of a gene's promoter region is not exclusively associated with promoter regions aligned with CGIs (Blelloch *et al.* , 2006). It is also questionable whether promoter methylation initiates the gene silencing or *vice versa*, with the latest evidence in support of the second of these scenarios (Jones, 2012).

Polycomb group proteins (PcG) play a fundamental role in development. They maintain a class of genes known as polycomb group targets (PCGTs) in a repressed state in ES (embryonic stem) cells, to maintain pluripotency, and 'poised for activation' during differentiation (Lee *et al.* , 2006). The link between PCGTs and cancer has been discussed by many authors (Widschwendter *et al.* , 2006; Ohm *et al.* , 2007; Schlesinger *et al.* , 2006). It was recently shown that DNA hypermethylation in cancers preferentially targets PCGTs which are developmental regulators (Easwaran *et al.* , 2012), and that this may hence contribute to the stem-like characteristics of cancer. In further support of these ideas it has been noted that tumours which are particularly poorly differentiated tend to display expression patterns which are similar to ES cells, including repression of PCGTs (Ben-Porath *et al.* , 2008).

Polycomb group proteins maintain the repressed state of genes via chromatin (the DNA packaging). DNA in its compact state is wrapped around histone proteins (a main component of chromatin). PRC2 (polycomb repressive complex 2) is responsible for the trimethylation of lysine 27 of histone 3 (leading to the epigenetic mark H3K27me3), which is associated with this compact state (Jones, 2012). Genes occupied by PRC2 in ES cells mostly carry bivalent chromatin marks (Easwaran *et al.* , 2012). Bivalency includes the histone modification H3K4me3 (trimethylation of lysine 4 on histone 3), a mark which is associated with activation of the corresponding gene, in addition to the repressive H3K27me3 mark. It is thought that it is this bivalent state which maintains stemness, keeping the gene repressed, but poised for activation upon differentiation. Because DNA methylation is also associated with repression and activation of genes, it is of interest whether genes that carry the chromatin markings H3K27 and/or H3K4me3 in stem cells have altered methylation patterns in cancer, as this might be associated with a return or accentuation of stem-like cell characteristics.

1.1.2 Network models

Networks and other non-Euclidean relational datasets have become important applications in modern statistics. An important consideration is balancing statistical fidelity with computational tractability. For network data, much attention has been on parametric models, such as degree based models, and community based alternatives (Holland *et al.* , 1983; Bickel & Chen,

2009; Rohe *et al.* , 2011; Qin & Rohe, 2013; Wilson *et al.* , 2013). One of the most widely studied of these models is the stochastic blockmodel, in which, under the assortative assumption, there is a greater probability of observing an edge (or interaction) between a pair of nodes (or individuals) if they are in the same block, or community. The problem of finding communities in social and biological networks has been studied for many years (Girvan & Newman, 2002). Real life examples of this problem include identifying groups of friends in a social network, and identifying functional subnetwork modules in a biological network. In the biological setting, considering groups of genes defined together as subgraphs can lead to great increases in statistical power, aiding discovery of novel biological phenomena (Jacob *et al.* , 2012; Li & Li, 2010; Peng *et al.* , 2010). The solution to this problem is often based on maximising the Newman-Girvan modularity (Newman & Girvan, 2004). The Newman-Girvan modularity quantifies the extent to which edges are observed between community members, for a particular assignment of nodes to communities, compared with the expected number of edges between community members if there were no community structure present. It can be shown that fitting the stochastic blockmodel and maximising the Newman-Girvan modularity over a network are, under certain conditions, both equivalent to spectral clustering (Bickel & Chen, 2009; Newman, 2013).

It has been shown recently that the stochastic blockmodel can be used to represent any exchangeable network as a 'network histogram', even if the generative mechanism of the network is not that of the blockmodel (Olhede & Wolfe, 2014). Exchangeable means here that the ordering of its nodes carries no information (Diaconis, 1977; Bickel & Chen, 2009). The network histogram and the blockmodel in general are piecewise-constant approximations of an underlying function, called the 'graphon' (Wolfe & Olhede, 2013), in the sense that the graphon function can be thought of as the generative mechanism of the data. The network histogram also provides a method to estimate the optimal number of blocks, or communities, which a valid blockmodel representation of the network comprises, if there is a smooth function in the graphon equivalence class. This is important and useful, because it means that the blockmodel can be used to identify, for example, an unknown number of communities in a social network, or an unknown number of functional subnetwork modules in a biological network. The network histogram method (Olhede & Wolfe, 2014) can be used to estimate the optimal granularity at which communities, or functional subnetwork modules, can be identified and isolated in social and biological networks, by fitting the stochastic blockmodel.

Over the past few years in cell biology, much of the focus has shifted from investigation of individual genes, to pathways of genes, to gene networks. The need for novel methodology

for network analysis in cell biology results from this recognition that examining the way genes work in groups is often more successful in revealing biological principles. Further, by considering groups of genes together as communities, statistical significance can be obtained which would not be possible at the level of individual genes. Many genes regulate, directly or indirectly, the behaviour (such as expression level) of other genes. Hence, networks are a natural way to model this gene regulatory and associated behaviour (e.g., relating to cell signalling). As a cancer progresses, its signalling and control networks are re-arranged ('re-wired'), and this drives adaptive alterations in phenotype, which are advantageous for the cancer (Barabási & Oltvai, 2004). Previous research has found that patient survival outcome in breast cancer (BRCA) could be predicted well by network models of this re-wiring, based on gene expression data (Taylor *et al.* , 2009). It is well established that DNA methylation plays a major role in gene regulation, and therefore DNA methylation patterns often reflect patterns of gene regulation. It has been previously shown that DNA methylation can serve as a surrogate for activity at genomic-regulatory regions (Brocks *et al.* , 2014). Hence, DNA methylation measurements are well suited as a basis from which to infer information about the topology and behaviour of genomic regulatory and associated networks.

1.2 Aims for the work

Changes in DNA methylation are highly stochastic. The time-scale over which these changes take place is much quicker than mutations in the basic DNA code, but much slower than the transient and periodically varying expression level of individual genes. This time-scale is ideal for biomarker development. DNAm measurements are also taken directly from DNA, whereas gene expression measurements must come via RNA. Hence, DNAm patterns might be expected to lead to more reliable disease biomarkers than gene expression patterns. Further, differences in DNA methylation levels are among the earliest changes in human carcinogenesis (Feinberg *et al.* , 2006). Therefore, DNA methylation data are thought to be extremely promising as a basis for the development of novel biomarkers. A major aim of this work is to inform the development of DNA-based biomarkers, and this is the main reason for the focus in this work on DNA methylation patterns.

The study of network models is a fascinating area of mathematical statistics on the most abstract level, and is a topic of much current interest in that field. Network models are also very well suited to analyse many equally fascinating problems of current interest in cell biology. Both fields stand to gain from this situation: new questions are raised by the field of cell biology, which give rise to new directions in mathematical statistics. In the process, cell biology acquires

new models and techniques with which to approach some difficult questions. My main aim is that the work presented here is of interest and relevance both to mathematical statisticians, and to cell biologists.

This thesis is organised as follows. In chapter 2, I introduce statistical measures of stochastic processes in intra-gene DNA methylation patterns, and investigate the association of these patterns with cancer. In chapter 3, I develop time-series methodology to infer differential network patterns in time-course DNA methylation measurements, of differentiating healthy and cancer stem cells. In chapter 4, I develop novel statistical network methodology to infer networks, and communities (i.e., groups of entities within these networks, amongst which there is a high density of interactions), from a range of matrices which measure the strength of interactive behaviour, between pairs of entities or variables. Such matrices may include covariance and correlation matrices, and test-statistics from arbitrary distributions (which may be expressed as 'p-values'). The methodology of chapter 4 is relied upon for the work carried out in chapters 5, 6 and 7. In chapter 5, I develop novel statistical methodology to infer and represent groups (called 'co-communities'), of strongly interacting entities or variables which are of two fundamentally different types, in bipartite networks. In chapter 6, I draw on the findings of chapter 2 to develop DNA-based prognostic biomarkers, based on stochastic processes in DNA methylation. In chapter 6, I also draw on the findings of chapters 4 and 5, to find groups of genes which strongly interact or are otherwise highly co-associated in terms of DNA methylation stochasticity and gene expression, deriving further biological meaning and suggesting new directions for experimental investigation. In chapter 7, I develop a measure of the strength of network interaction between pairs of genes based entirely on DNA methylation data, and I draw on the findings of chapter 5 to infer groups of strongly interacting, or highly co-associated genes. These groups form potential DNA methylation network biomarkers for cancer.

1.3 Publications

The findings of chapter 2 have been published as the following journal article: *Bartlett, T. E., Zaikin, A., Olhede, S. C., West, J., Teschendorff, A. E., & Widschwendter, M. (2013). Corruption of the intra-gene DNA methylation architecture is a hallmark of cancer. PloS one, 8(7), e68285.*

Chapter 3 is based on methodology which contributed to a journal article currently under review at Cell Stem Cell: *Carn H., Stricker S. H., Gagrica S., Bartlett T. E., Feber A., Wilson G., Teschendorff A. E., Beck S., & Pollard S. M. BMP signalling does not trigger terminal differentiation of glioblastoma stem cells.*

The findings of chapter 4 are included in a journal article currently under review at the

Journal of Applied Statistics: *Bartlett, T. E. Community detection and network inference, based on covariance matrices and test statistics from arbitrary distributions.*

The findings of chapter 5 are being prepared for submission to the Journal of the Royal Statistical Society Series C: *Bartlett, T. E., & Olhede, S. C. Co-modularity and Co-community Detection in Large Networks.*

The findings of chapter 6 are included in a journal article currently under review at PLoS Medicine: *Bartlett, T. E., Jones, A., Goode, E. L., Fridley, B. L., Cunningham, J. M., Berns, E. M. J. J., Wik, E., Salvesen, H. B., Davidson, B., Trope, C. G., Lambrechts, S., Vergote, I., & Widschwendter, M. Intra-gene DNA methylation variability is a technically and clinically independent prognostic marker in women's cancers.*

Some of the findings of chapter 7 have been published as part of a journal article: *Bartlett, T. E., Olhede, S. C., & Zaikin, A. (2014). Detection of Epigenomic Network Community Oncomarkers. PloS one, 9(1), e84573.*

Some of the findings of chapter 7 have appeared as part of a conference paper: *Bartlett, T. E., Olhede, S. C., & Zaikin, A. (2014). Novel Statistical Network Methodology to Identify and Analyze Cancer Biomarkers. Joint Statistical Meeting Proceedings, Statistical Epidemiology Section. American Statistical Association, Boston, MA, U.S.A.*

The rest of the findings of chapter 7 are included in a journal article currently under review at the Annals of Applied Statistics: *Bartlett, T. E., & Zaikin, A. DNA Methylation Network Community Oncomarker Detection.*

Work I did during this PhD, but which is not included in this thesis, has also been published as part of a journal article: *Teschendorff A. E., Marabita F., Lechner M., Bartlett T. E., Tegner J., Gomez-Cabrero D., & Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450K DNA methylation data. Bioinformatics, 29(2):189-196.*

Chapter 2

Corruption of the Intra-Gene DNA Methylation Architecture Is a Hallmark of Cancer

2.1 Introduction

Differences in DNA methylation levels - in particular hypermethylation of genes with tumour suppressor function or involvement in stem cell biology - are among the earliest changes in human carcinogenesis, and hence offer novel strategies to identify individuals who might be at risk of developing such illnesses or individuals with early stage cancers. However, to proceed with developing such tests, measures of DNA methylation are needed which can be consistently linked to clinically relevant differences such as disease states of samples.

The role of early epigenetic changes in oncogenic transformation, including disruption of the healthy epigenotype of progenitor cells, the creation of an epigenetically permissible environment in which genetic aberrations can have tumorigenic effects, and phenotypic plasticity leading to tumour adaptation and associated with intra-tumour heterogeneity, was originally proposed by (Feinberg *et al.* , 2006). More recently, the effect of stochastic noise as an epigenetic phenomenon and its effect on phenotypic plasticity has been explored by (Pujadas & Feinberg, 2012). It is hypothesised that one way in which stochastic dysregylation of stem cell genes (such as PCGTs) and associated phenotypic heterogeneity might manifest, is in terms of changes to intra-gene methylation variability. As such, a change in intra-gene methylation variability may be closely linked to the creation of an epigenetically permissible environment for oncogenic transformation, and to tumourigenesis. Such changes would be expected to accompany the early stages or even precede the onset of the disease, and hence identifying reliable indicators of such changes might provide a valuable lead in the search for markers for use in screening programmes or early diagnosis.

Recently, evidence together with plausible biological mechanisms have been presented (Jaffe *et al.* , 2012) suggesting that variability of methylation at specific genomic locations is important in the development of cancer. It has been noted in particular that there is an increase in stochastic methylation variability in regions which are already known to have altered levels of methylation in cancers, leading to aberrant and varying gene expression, and providing an epigenetic mechanism for tumour heterogeneity (Hansen *et al.* , 2011). Intra-gene methylation variability is deemed to be a disruption of the normal methylation profile, or architecture, of a particular gene, and such a change may be more generally linked to the creation of an epigenetically permissible environment for oncogenic transformation, and to tumourigenesis. Such changes would be expected to accompany the early stages or even precede the onset of the disease, and hence identifying reliable indicators of such changes might provide a valuable lead for the development of DNA-based cancer biomarkers in bodily fluids.

Previous studies (Jaffe *et al.* , 2012; Teschendorff & Widschwendter, 2012; Teschendorff *et al.* , 2012) have focussed on the effects of sample to sample variability of methylation; here for the first time, I analyse the association of phenotype with intra-gene variability of methylation (IGV). Making use of DNA methylation data derived from the Illumina Infinium HumanMethylation450 platform, which interrogates > 485000 CpGs genome-wide including > 330000 with known gene annotations (corresponding to on average 17 CpGs per gene), I have analysed IGV in 681 normal and 3284 cancerous samples, taken from 14 different cancer entities.

2.2 Results

To investigate intra-gene methylation architecture, four gene-centric measures are considered, as follows:

1. The mean deviation of the sample methylation profile from the mean methylation profile of healthy phenotype control samples, for each gene. This mean methylation profile may fluctuate a lot within each gene, and so it is not the same as the mean methylation level of a gene. Because this mean deviation is normalised at every probe by dividing by the probe standard deviation across the healthy phenotype control samples, it is called the 'mean z-score' measure; this is illustrated in figure 2.1(a). An example of one of the genes found to be most significant according to this measure is shown in figure 2.1(b) and (c).

2. The mean derivative of the methylation measurements for each gene. The derivative of the methylation profile for a given gene and sample is approximated by the differences

between the methylation values measured at consecutive probes mapping to that gene. The mean of the absolute values of these differences is then calculated as the 'mean derivative' measure; this is the same as the sum total of all the increases and decreases in methylation level from one probe to the next across the gene. This is a self-calibrating measure of intra-gene methylation variability, because it is calculated for a given sample without reference to any other sample.

3. The mean of the methylation measurements for a particular genomic region for each gene. Typical mean methylation levels vary greatly from one genomic region to another; hence the mean methylation level for a particular genomic region was used as the 'mean methylation measure' for a gene, and the same region was used for each gene.

4. The variance for each gene of the methylation measurements for a particular genomic region. Because variance is calculated in relation to the mean, this measure was similarly calculated for each gene using only the probes mapping to a particular genomic region, again using the same genomic region for each gene. This is called the 'methylation variance' measure; it is another self-calibrating measure.

These four measures each seek to examine a different characteristic of intra-gene methylation architecture, and all are able to classify samples one-by-one, i.e., they are intra-gene or intra-sample measures, rather than sample to sample measures as has been investigated previously in the context of methylation variability.

As the mean z-score is calculated as a mean measure of methylation difference from the healthy methylation profile, strictly speaking it is a measure of methylation instability. The mean derivative and methylation variance measures are both measures of intra-gene methylation variability; however, the mean derivative is calculated with reference to the ordering of the probes (i.e., this measure would return a different number if the order of the probes was randomised) whereas the methylation variance would not; the mean derivative additionally considers all probes mapping to the gene, whereas the methylation variance measure only considers probes mapping to a particular genomic region. The mean methylation measure is unique here in that

	healthy	cancer	total
BRCA	98	586	684
UCEC	36	334	370
THCA	50	357	407
LUAD	32	306	338
BLCA	18	126	144
LUSC	43	227	270
COAD	38	258	296
HNSC	50	310	360
KIRC	160	283	443
LIHC	50	98	148
READ	7	96	103
PRAD	49	176	225
KIRP	44	87	131
PAAD	6	40	46

Table 2.1: *Number of samples in each data set*

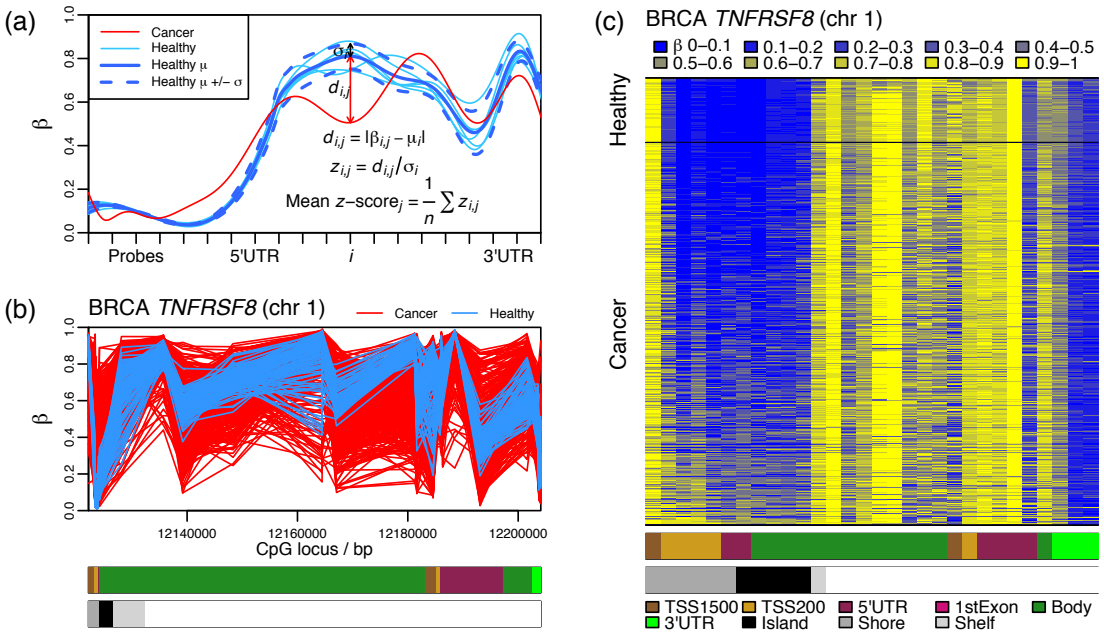


Figure 2.1: *The mean z-score measure*

(a) The mean z-score is calculated for tumour sample j (shown in red) for gene g (to which n probes map), from the mean, μ_i, and standard deviation, σ_i, of the healthy control samples at each probe i (b) The methylation profiles of 586 cancer (red) and 98 healthy (blue) samples across a gene, with probes spaced (unevenly) according to their genomic loci. Genomic regions are indicated under the gene with the colour code displayed at the bottom of the figure. (c) A heatmap illustrating the same gene, with probes evenly spaced; beta values for each sample and each probe are indicated by the colour code displayed at the top of the figure. Samples are plotted in order of mean z-score, such that the tumour sample with the smallest mean z-score and the healthy sample with the smallest mean z-score are adjacent. Genomic regions are indicated under the gene with the colour code displayed at the bottom of the figure. N.B., this gene has two transcriptional start sites (TSSs) in different locations.

it does not measure difference in methylation level and instead measures absolute methylation level; it is included here mainly for comparison.

The properties of these four measures were initially investigated in the context of fourteen Illumina Infinium Human Methylation 450 data sets, which were downloaded from The Cancer Genome Atlas (TCGA) (Collins & Barker, 2007). I applied these four measures to the fourteen TCGA data sets; in all, I analysed 450K DNAm data from 3284 tumour and 681 healthy samples; details of the number of samples of each phenotype and in each data set are shown in table 2.1 (for data set abbreviations, see 'methods and models'). I also carried out a meta-analysis of these data which is to my knowledge the largest meta-analysis performed in any DNA methylation study.

2.2.1 Comparison of intra-gene methylation measures

As a preliminary assessment of the relative merits of these four measures, I looked at their ability to distinguish between tumour and healthy tissue. The correlation of the tissue sample phenotype to the four methylation measures was considered in terms of distributions of per-

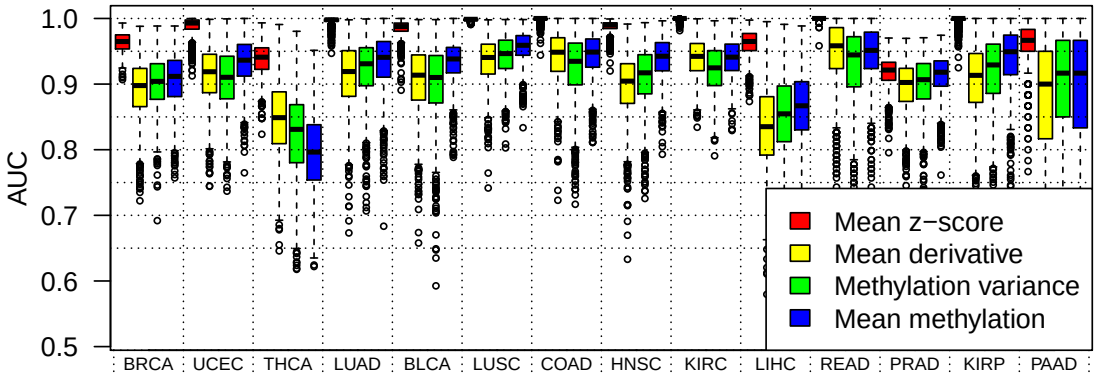


Figure 2.2: *Distributions of per-gene AUCs calculated from intra-gene methylation measures*
Each box displays the values of the AUCs for the 1000 most significant genes for a particular tumour type and intra-gene methylation measure. The mean z-score predicts phenotype better than the other three measures in all 14 tumour types. Tumour type abbreviations are as follows: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Colon Adenocarcinoma (COAD), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Renal Papillary Cell Carcinoma (KIRP), Liver (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Pancreatic Adenocarcinoma (PAAD), Prostate Adenocarcinoma (PRAD), Rectum Adenocarcinoma (READ), Thyroid Carcinoma (THCA), and Uterine Corpus Endometrioid Carcinoma (UCEC).

gene AUCs (area under curve, which is a measure of prediction accuracy, see 'methods and models' for details). These distributions are shown in box-plots in figure 2.2. For every data set, the mean z-score measure is significantly better at discriminating tumour from healthy tissue using these methylation data, than the mean derivative measure, the methylation variance measure, and the mean methylation measure (visual comparison of figure 2.2 was confirmed by Kolmogorov-Smirnov tests, data not shown); this is because the mean z-score measure is defined relative to the healthy mean methylation profile. Excluding the mean z-score measure, the mean methylation measure is significantly better at discriminating tumour from healthy tissue than the remaining two measures in ten of the remaining data sets, with the mean derivative discriminating significantly better in two data sets (READ and THCA), and inconclusive results for the remaining data sets (KIRC and PAAD, which has unstable results due to small sample size).

I investigated the overlap of the top-ranked genes according to each measure (figure 2.3). These four methylation measures are seen to be complimentary in that there are lower ranked genes according to each measure (genes further towards the bottom of each vertical bar) which are significant according to the other measures (indicated by colour coded lines); however the mean z-score appears to offer the most complimentary information because some of the top 1000 most significant genes according to this measure (coloured red) are frequently found among the lower ranked genes according to the other measures (towards the bottom of the vertical bars).

To directly compare the effectiveness of the mean z-score measure at predicting phenotype (cancer/healthy) independent of mean methylation level, a logistic regression model was fitted to each gene using mean z-score and mean methylation as covariates, leading to p-values for each gene for each of mean z-score and mean methylation. In every data set except two, for the large majority (80-100%) of those genes with at least one of the two covariates significant, the mean z-score covariate p-value was more significant than the corresponding mean methylation covariate p-value. In the remaining two data sets, the mean z-score covariate p-value was more significant for the majority (50-80%) of genes with at least one significant covariate (detailed results not shown). Hence, the mean z-score is a better predictor of phenotype than the mean methylation, even after adjustment for mean methylation level.

2.2.2 Meta-analysis and gene-set enrichment analysis

A meta-analysis of the fourteen data sets was carried out. Genes were assigned significance according to their mean AUC (based on the mean z-score measure) across all data sets by a permutation method (see 'methods and models' for details); this identified over 4000 significant genes which were associated with a consistent difference between cancer and healthy phenotypes across tissue types (FDR $q \leq 0.05$). These genes consistently show the biggest differences between healthy and cancer phenotypes (as the mean z-score measure is defined relative to healthy control samples), and as the mean z-score is a measure of methylation instability, they are termed the most unstable meta-analysis genes, 'MUs'. The mean z-scores for individual tumour and healthy samples for the 50 most significant of these MUs meta-analysis genes are displayed in figure 2.4. In particular, figure 2.4 shows the extent to which the instability is consistent (high mean z-score, red) across cancer patients as compared to healthy patients (low mean z-score, blue). Genes with a mean AUC close to 0.5 across most tumour types were also found; these are genes which tend to have the smallest differences between healthy and cancer phenotypes across tissue types and hence are marked as least unstable meta-analysis genes, 'LUs'. Over 2800 LUs genes were found to be significant by this permutation method (FDR $q \leq 0.05$).

To confirm the biological significance of the findings of this meta-analysis with reference to genes which are well known to be important in cancer biology, the MUs and LUs genes were tested for enrichment by genes which in ES cells carry the repressing/activating chromatin marks H3K27me3 (H3K27 ES genes), H3K4me3 (H3K4 ES genes) and bivalent (i.e., both H3K27me3 and H3K4me3 marks, Biv ES genes) and enrichment by PCGTs (ES cell polycomb group targets); MUs are highly enriched by Biv and H3K27 ES genes and PCGTs, and LUs genes are highly enriched by H3K4 ES genes (table 2.2). A more general gene-set enrichment

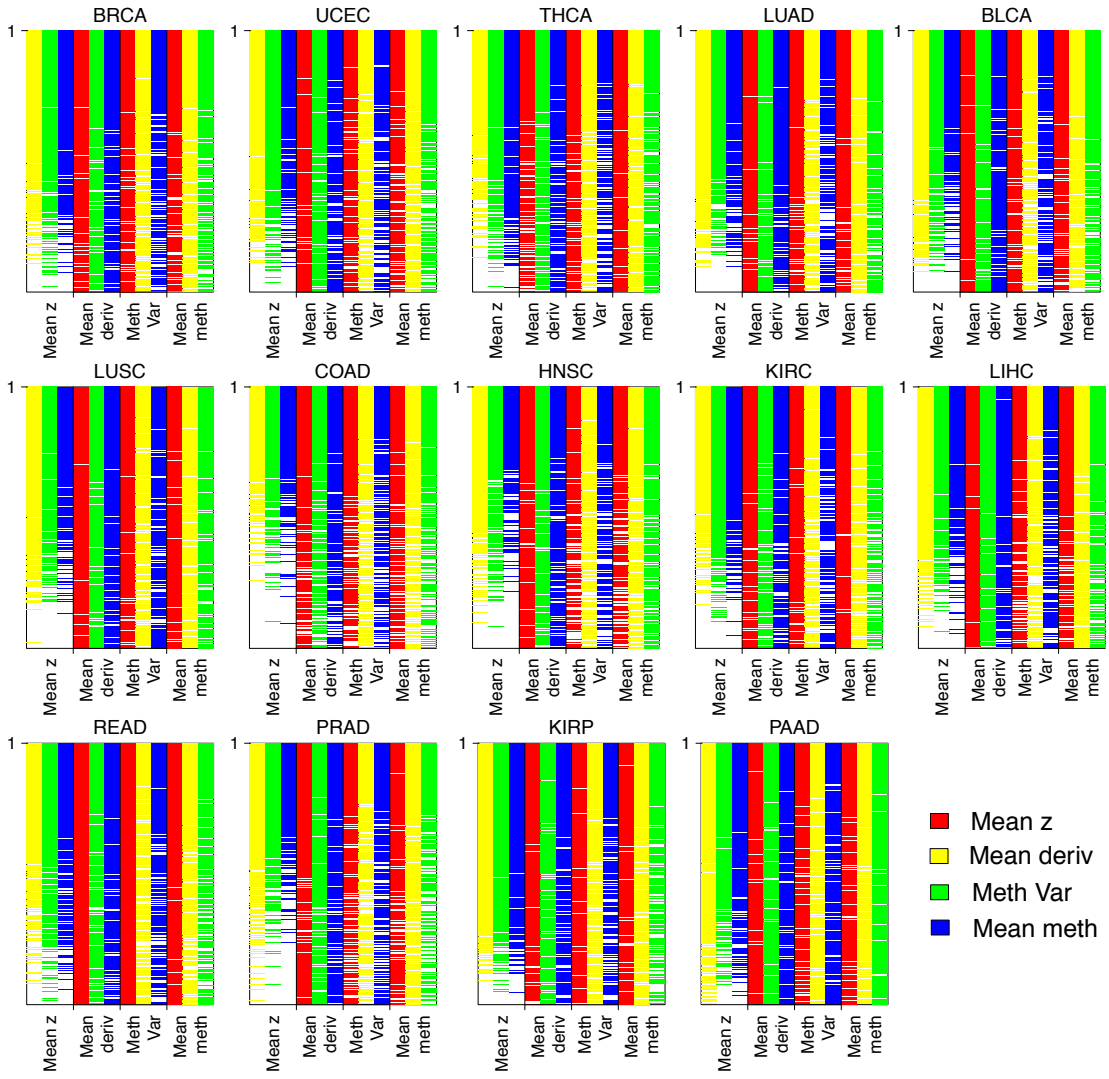


Figure 2.3: *Overlap of genes found as significant according to each of the intra-gene methylation measures*

For a given tumour type, for each methylation measure genes are ranked according to their significance as defined by the AUC measures summarised in figure 2.2. Then for each measure with genes ranked in this way, the locations of the top 1000 genes according to each of the other three measures are displayed with colour coded lines (these colours are as defined at the top of the figure). Hence, for each tumour type, there are 12 vertical bars of horizontal lines: for each of the four methylation measures, there is a set of three of these vertical bars, with one bar for each of the remaining three methylation measures. The horizontal lines in each of these three bars in a set have the same ordering (which is according to significance due to the methylation measure indicated under the set of three bars); each vertical bar is then coloured to indicate significance according to the three other methylation measures, with the 1000 most significant genes according to each of these methylation measures indicated with a coloured horizontal line. The four methylation measures are complimentary because there are lower ranked genes according to each measure (genes further towards the bottom of each vertical bar) which are significant according to the other measures; however the mean z-score appears to offer the most complimentary information because some of the top 1000 most significant genes according to this measure (coloured red) are frequently found among the lower ranked genes according to the other measures.

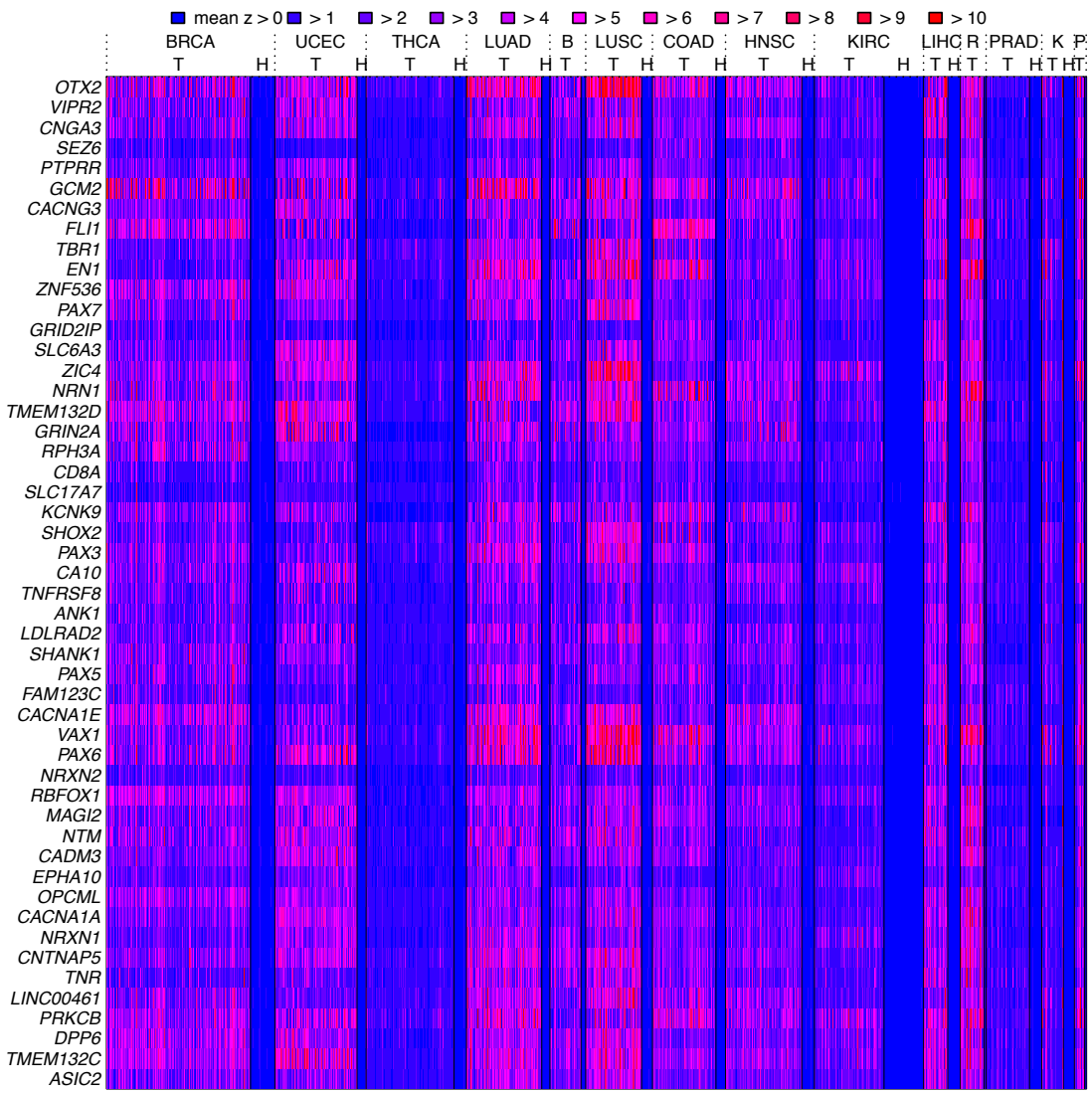


Figure 2.4: *Heatmap of the mean z-score for the top 50 genes found by the meta-analysis*

Mean z-scores for tumour (T) and healthy (H) samples are displayed in a heatmap according to the colour code for the top 50 meta-analysis genes (top 50 MUs genes). The heatmap shows the extent to which the instability is consistent (high mean z-score, red) across cancer patients as compared to healthy patients (low mean z-score, blue). For each tissue type healthy samples appear to the right of tumour samples; where no space is available the (H) label is omitted. Abbreviations: R (READ), B (BLCA), K(KIRP), P (PAAD).

	H3K27	H3K4	Biv	PCGT
MUs	1.43×10^{-28}	1	5.19×10^{-278}	1.77×10^{-234}
LUs	1	4.33×10^{-70}	1	1

Table 2.2: *Enrichment of MUs and LUs genes by stem cell genes*

P-values (one-sided Fisher's exact test) show enrichment of MUs (most unstable meta-analysis genes) and enrichment of LUs (least unstable meta-analysis genes) by genes in various SC categories. This confirms the biological significance of the findings of the meta-analysis with reference to these genes which are well known to be important in cancer biology.

analysis (GSEA) was also carried out, testing enrichment of the MUs and LUs genes by members of over 6000 gene sets (see 'methods and models' section for details). In particular, the MUs genes show enrichment by many developmental and cell signalling gene sets.

The MUs genes are associated with generally higher methylation levels than genes which are not significant according to the meta-analysis (i.e., genes which are neither MUs or LUs) for both tumour and healthy samples, for these genomic regions located closer to the promoter across all tissue types, however the MUs genes are also associated with a large variability of methylation levels. The LUs genes conversely are associated with consistently very low levels of methylation in both tumour and healthy samples for these genomic regions, and particularly for TSS200, 5'UTR and 1stExon, suggesting that the low methylation instability of these genes is associated with a lack of methylation in the most functionally important genomic regions in both diseased and normal tissues, and therefore that regulation of these genes is by mechanisms other than those involving DNA methylation, in particular the availability of transcription factors.

2.2.3 Correlation of tumour gene expression with intra-gene methylation architecture

In order to investigate the effect of intra-gene methylation architecture on gene expression, the 217 BRCA tumour samples with matched gene expression and methylation data available from TCGA were considered in more detail. For each gene a non-linear multivariate regression analysis was performed (see 'methods and models') of gene expression to intra-gene methylation architecture, for these matched tumour samples, taking gene expression as the response, and taking one of mean z-score, mean derivative and methylation variance as one covariate predictor, together with mean methylation as a second covariate predictor. The relative proportions of genes found as significant or not, and significant according to one covariate or the other, or both, are shown in figure 2.5; in particular there are many genes with expression not significantly predicted by mean methylation but significantly predicted by mean z-score, mean derivative, or methylation variance.

Enrichment by stem cell genes of genes with expression significantly predicted by only

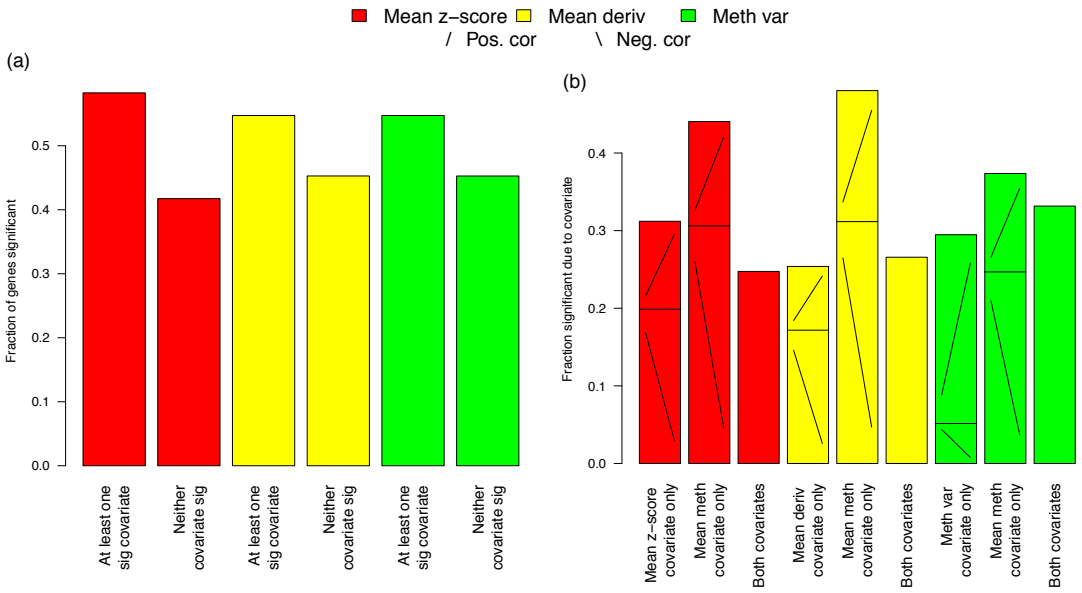


Figure 2.5: *Correlation of expression to intra-gene methylation architecture, for matched BRCA samples Expression was taken as the response variable, with one of mean z-score, mean derivative and methylation variance as one covariate predictor, together with mean methylation as a second covariate predictor. (a) The proportion of genes with at least one covariate significant (FDR $q \leq 0.05$), and the proportion of genes with neither covariate significant. (b) The proportion of significant genes (i.e., the proportion of the genes represented by the left of each pair of bars in a) which are significant due to one, or the other, or both covariates. For the genes which are significant due to only one covariate predictor, the proportions of these genes for which the significance is due to positive or negative correlation are indicated on the bars with / and \ respectively. There are many genes with expression not significantly predicted by mean methylation but significantly predicted by mean z-score, mean derivative, or methylation variance.*

one covariate was again tested to confirm the biological significance of findings with reference to genes which are well known to be important in cancer biology. It was found that genes with expression predicted by only the mean z-score covariate were significantly enriched by Biv ES genes and PCGTs ($p = 1.3 \times 10^{-3}$ and $p = 5.0 \times 10^{-3}$ respectively, Fisher's exact test), a result which is consistent with the findings here that Biv ES genes are enriched among MUs meta-analysis genes, i.e., those genes which are most consistently associated with the biggest difference in methylation pattern between cancer and healthy phenotypes. It was also found that, correspondingly, genes with expression predicted by only the mean methylation covariate in the multivariate regression with the mean z-score covariate were significantly enriched ($p = 9.0 \times 10^{-4}$, Fisher's exact test) by H3K4 ES genes, a result which is consistent with my findings that H3K4 ES genes are enriched among LUs meta-analysis genes, i.e., those genes which have consistently least difference in methylation pattern between cancer and healthy phenotypes. Similarly, it was found that genes with expression predicted by only the mean derivative covariate were significantly enriched by Biv ES genes and PCGTs ($p = 9.5 \times 10^{-4}$ and $p = 8.4 \times 10^{-4}$ respectively, Fisher's exact test) and that genes with expression predicted

only by the mean methylation covariate in the same multivariate regression were significantly enriched by H3K4 ES genes ($p = 3.1 \times 10^{-4}$, Fisher's exact test).

These findings extend to heterogeneous tumour phenotype, as defined by gene expression, the idea that differences in methylation patterns in stem cell genes are a hallmark of cancer, and shows that this can be measured by intra-gene methylation architecture in the form of intra-gene methylation variability (according to the mean derivative and methylation variance measures) and instability (according to the mean z-score measure) more accurately than by mean methylation level alone.

2.2.4 Association of genome-wide mean z-score with breast cancer intrinsic subtypes

Differences in intra-gene methylation architecture between heterogenous tumour phenotypes (as defined by gene expression) was further explored, in the context of breast cancer intrinsic subtypes. The same 217 BRCA samples with matched gene expression and methylation data available were each uniquely assigned to one of these disease subtypes, according to established molecular definitions, using the PAM50 classifier (Parker *et al.* , 2009). This was done by correlating the gene expression profile (Spearman correlation) for each sample to the PAM50 classifier canonical gene expression profiles for 5 different intrinsic subtypes, and for each sample choosing the subtype with the largest correlation coefficient, leading to 42 samples classified as Basal, 24 as Her2, 81 Luminal A, 54 Luminal B, and 16 classified as Normal. For each of these samples, a genome-wide mean z-score was also calculated, as a per-sample genome-wide measure of intra-gene methylation architecture. The distributions of these genome-wide mean z-scores for each intrinsic subtype are shown in figure 2.6; there are clear differences in the means and distributions between each of the subtypes. A Kruskal-Wallis test was carried out to check the significance of these differences, with a very significant result, $p = 1.4 \times 10^{-12}$. Removing the samples classified as Luminal B and Normal (as the distributions of genome-wide mean-z scores have larger and smaller variances, respectively, for these subtypes than the others), still resulted in a significant result in the Kruskal-Wallis test, $p = 0.023$. This ability to distinguish between heterogenous tumour phenotypes, in the context of established molecular definitions of disease subtypes, indicates that it may be possible to use intra-gene methylation architecture to develop new molecular classifiers of cancer, or make established ones more robust. This is particularly interesting, since methylation levels are typically more stable than gene expression levels.

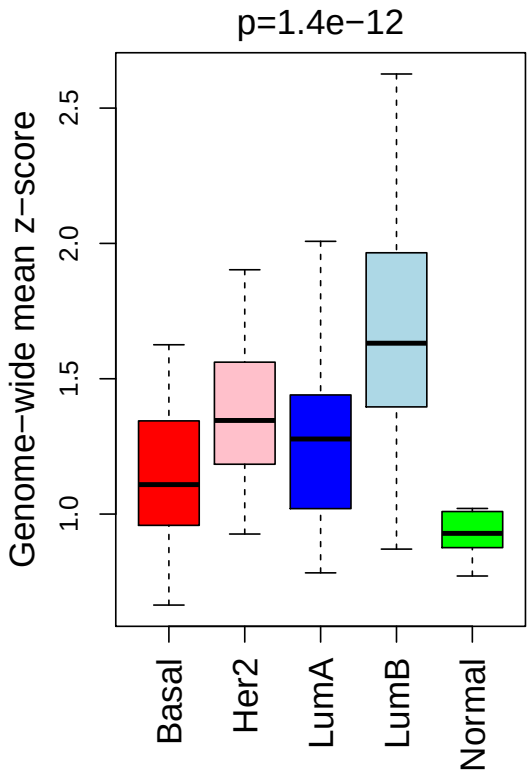


Figure 2.6: *Distributions of genome-wide mean z-score, for breast cancer intrinsic subtypes*
The mean across all genes of the mean z-scores was calculated for the 217 BRCA samples with matched expression and methylation data available. These samples were independently classified by correlation of their gene expression profiles (Spearman correlation) with those of the PAM50 breast cancer intrinsic subtype classifier (Parker et al. *, 2009). The distributions of these genome-wide mean z-scores, for each intrinsic subtype, are shown in the boxplots. Indicated significance was calculated using the Kruskal-Wallis test.*

2.2.5 Intra-gene methylation architecture as a predictor of clinical outcome

A preliminary test of the ability of intra-gene methylation architecture to predict clinical outcome was carried out, using a small, publicly available pilot data set, generated as part of a study of childhood B-cell acute lymphoblastic leukaemia (Sandoval *et al.* , 2013). This data set contains methylation data generated using the same Illumina Infinium HumanMethylation450 platform, with corresponding clinical outcome data in the form of binary recurrence / non-recurrance status (5 and 24 samples respectively), and was downloaded from Gene Expression Omnibus (GEO).

The data were split into training and test sets, and the Elastic Net algorithm (Zou & Hastie, 2005; Friedman *et al.* , 2010) was used to fit the model to the training set, automatically selecting the subset of features (genes or CpGs), out of all those available, which model the data best. This model fit was then used to blindly predict the outcome in the test set, and this was repeated multiple times as part of a 'leave two out cross-validation' strategy (Herzberg & Tsukanov, 1986), covering every possible division of test and training set (see 'Methods and models' for

further details). This analysis was carried out for the four methylation measures described, as well as probe-level CpG beta-values, and in the case of the methylation variance and mean methylation levels, for both gene body and TSS 200 (promoter) genomic regions; the resulting test-set AUCs calculated across all possible test/training set permutations are shown in table 2.3. The methylation variance performs particularly well, outperforming all other measures. On the other hand, the mean z-score performs particularly badly; this might be because corresponding healthy data for only four healthy samples was available, which could have lead to poor estimates of the healthy population methylation parameters which this measure is calculated in relation to.

2.3 Discussion

I have shown that the reorganisation of intra-gene methylation architecture is a fundamental characteristic of cancer cells, and that there are many ways to assess these differences, which can provide complimentary information. I have developed measures to detect some of these differences, including the first investigation of intra-gene variability of methylation (as opposed to sample to sample variability of methylation). I have shown that my mean z-score measure is consistently more effective at predicting cancer compared to healthy phenotype than mean methylation, even after adjustment for the mean methylation level, and I have found an indication that intra-gene methylation variability is more effective at predicting clinical outcome than mean methylation level or individual CpG methylation level, in a small pilot study.

	AUC
CpG Beta	0.78
Mean z-score	0.64
Mean derivative	0.79
Gene body variance	0.88
Gene body mean	0.77
TSS 200 variance	0.86
TSS 200 mean	0.84

Table 2.3: *Association of methylation measures with clinical outcome*

Association of methylation measures with disease recurrence, based on a small pilot data set from a study of childhood B-cell acute lymphoblastic leukaemia. AUCs were calculated by fitting the model to a training set; this model was then used for blind prediction in a test set. AUCs reported were calculated from all possible test/training set permutations.

I have carried out what is, to my knowledge, the largest meta-analysis performed in any DNA methylation study. In particular, over 4000 MUs genes were found to be significantly associated with a consistent difference between cancer and healthy phenotypes, demonstrating that, as a method for distinguishing cancer from healthy tissue, my mean z-score measure is robust to differences between tumour types. The most significant MUs genes according to this meta-analysis can be considered as particularly characteristic of a generalised and non tissue-specific cancer phenotype. These significant MUs meta-analysis genes are also significantly

enriched (table 2.2) by genes carrying H3K27 and bivalent chromatin marks in ES cells and by PCGTs, consistent with the idea that the tumour phenotype is associated with the acquisition of stem-like cell characteristics (Easwaran *et al.* , 2012). In this meta-analysis, over 2800 LUs genes were also found to be significantly associated with an absence of difference in methylation pattern from healthy to cancer, and these are significantly enriched by genes carrying the activating H3K4 chromatin mark in ES cells (table 2.2).

The correlation for tumour samples of gene expression to intra-gene methylation architecture (figure 2.5) shows that there are a substantial number of genes for which mean methylation is not significantly predictive of gene expression but other measures of intra-gene methylation architecture are. In particular, in the case of my mean z-score and mean derivative measures, genes with expression predicted by these measures and not by mean methylation are enriched by Biv ES genes and PCGTs, suggesting that the intra-gene methylation instability and variability are able to provide important information about heterogeneous tumour phenotype (as measured by gene expression), particularly in relation to stem-like cell characteristics, which is beyond the reach of measures based on mean methylation level alone.

The differences in the genome-wide mean z-scores across breast cancer intrinsic subtypes (figure 2.6) highlight the potential of intra-gene methylation architecture to distinguish between heterogenous tumour phenotypes in the context of established gene expression based definitions of distinct subtypes of this disease. This indicates that it may be possible to use intra-gene methylation architecture to develop new molecular classifiers of cancer, or make established ones more robust.

Further improvements in classification by my methods will be gained by the inclusion of complementary epigenetic data, in particular those which measure patterns of histone modification. As discussed, it is well established how crucial genes which carry important histone markings in stem cells are to understanding cancer biology. By extending the view of the epigenetic landscape beyond DNA methylation to consider also histone markings not just in stem cells but also in mature healthy cells and cancer cells, we will gain mechanistic insights into the interaction between intra-gene methylation architecture and histone modifications.

In summary, I have shown for the first time that generalised differences in intra-gene methylation architecture are a better predictor of phenotype than mean methylation level alone, and I have developed novel measures of these differences, which offer a considerable reduction in complexity from per CpG methylation measures (hundreds of thousands of features) to per gene methylation measures (tens of thousands of features). I have shown that there are many genes with expression predicted by measures of intra-gene methylation architecture other than

mean methylation level, and therefore that more general measures of intra-gene methylation architecture offer novel information about heterogeneous tumour phenotype (as defined by gene expression). I have also shown that intra-gene methylation architecture is able to distinguish between established molecular definitions of heterogenous cancer subtypes, and I have found an indication that intra-gene methylation architecture might be a better indicator of patients likely to suffer a recurrence of cancer, than more familiar mean or individual CpG methylation measures. Because it has been shown previously that differences in methylation pattern occur prior to the onset of disease (Zhuang *et al.* , 2012), I anticipate that my measures of intra-gene methylation architecture might also be able to efficiently find pre-disease methylation patterns. I therefore believe that my measures of intra-gene methylation architecture have potential for further development as DNA based cancer biomarkers.

2.4 Methods and models

2.4.1 Data source and preprocessing

Methylation data, collected via the Illumina Infinium HumanMethylation450 platform, were downloaded from The Cancer Genome Atlas (TCGA) project (Collins & Barker, 2007) at level 3. These data were obtained from fourteen different tumour types, as follows: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Colon Adenocarcinoma (COAD), Head and Neck Squamous Cell Carcinoma (HNSC), Kidney Renal Clear Cell Carcinoma (KIRC), Kidney Renal Papillary Cell Carcinoma (KIRP), Liver (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Pancreatic Adenocarcinoma (PAAD), Prostate Adenocarcinoma (PRAD), Rectum Adenocarcinoma (READ), Thyroid Carcinoma (THCA), and Uterine Corpus Endometrioid Carcinoma (UCEC).

These data were pre-processed by first removing probes with non-unique mappings and which map to SNPs (as identified in the TCGA level 3 data); probes mapping to sex chromosomes were also removed; in total 98384 probes were removed in this way from all data sets. After removal of these probes, 270985 probes with known gene annotations remained. Individually for each data set, probes were then removed if they had less than 95% coverage across samples; probe values were also replaced if they had corresponding detection p-value greater than 5%, by KNN (k nearest neighbour) imputation ($k = 5$).

Matched gene expression data were also downloaded for 217 samples for the BRCA data set, and were quantile normalised.

2.4.2 Intra-gene methylation measures

Four methylation measures were considered, and were calculated separately for each sample, for each gene:

- 'Mean z-score': the mean of the z-scores calculated from the methylation values for the probes mapping to the gene, with population parameters for each probe calculated from healthy control samples
- 'Mean derivative': the mean absolute derivative of the methylation profile across the gene
- 'Methylation variance': the variance of the methylation values for probes mapping to one genomic region of the gene
- 'Mean methylation': the mean of the methylation values for probes mapping to one genomic region of the gene

To calculate the mean of the z-scores for each gene, the *R* / *Bioconductor* package 'IlluminaHumanMethylation450k' (Triche & Jr., 2012) was used to identify the probes mapping to each gene. Then for each probe, the mean and standard deviation of the methylation values for that probe were found from healthy tissue samples, allowing a z-score $z_{i,j}$ for each probe i, and for each sample j, to be calculated according to equation 2.1. By taking the mean of the absolute $z_{i,j}$ for all probes i mapping to gene g, a single intra-gene methylation predictor value $x_j(g)$ was then calculated for each gene g, for each sample j, according to equation 2.2. A regularisation parameter, ξ, was added to each probe standard deviation when calculating probe z-scores to prevent very large values from occurring; ξ was chosen to be 0.01 after considering the distribution of probe standard deviations.

$$z_{i,j} = \frac{|\beta_{i,j} - \mu_i^{(\mathrm{h})}|}{\sigma_i^{(\mathrm{h})} + \xi} = \frac{d_{i,j}}{\sigma_i^{(\mathrm{h})} + \xi} \tag{2.1}$$

$$x_j(g) = \frac{1}{n(g)} \sum_{i \in P(g)} z_{i,j} \tag{2.2}$$

where $\beta_{i,j}$ is the methylation value for probe i and sample j, $\mu_i^{(\mathrm{h})}$ and $\sigma_i^{(\mathrm{h})}$ are the mean and standard deviation of the methylation values corresponding to the relevant healthy tissue samples for probe i, $n(g)$ denotes the number of probes mapping to gene g and $P(g)$ is the set of probes mapping to gene g.

To calculate the 'mean derivative' methylation measure, the 'IlluminaHumanMethylation450k' package was again used to find the probes mapping to each gene. Ordering the

probes $P(g) = \{i(1), ..., i\,(n(g))\}$ mapping to gene g as they are positioned along the DNA, the derivative of the methylation profile for gene g and sample j is estimated as the differences between the beta values at consecutive probes; hence the mean derivative for this gene and sample is estimated according to equation 2.3.

$$x_j\,(g) = \frac{1}{n(g)-1} \sum_{1 \leq k < n(g)} \left|\beta_{i(k+1),j} - \beta_{i(k),j}\right| \tag{2.3}$$

In this way, a single intra-gene methylation predictor value $x_j(g)$ was calculated for each gene g, for each sample j.

To calculate the 'methylation variance' and 'mean methylation' measures, first the most effective genomic region, for each of these measures, across which to calculate these measures for each gene, was selected. For this, annotation information for the probes used by the Illumina Infinium platform was obtained from Gene Expression Omnibus (GEO) (Edgar *et al.* , 2002). This annotation information details which probes map to one of six genomic regions for each gene, as follows: (1) TSS1500; probes annotated to distances greater than 200bp and less than 1500bp upstream from the TSS (transcriptional start site) of the gene. (2) TSS200; probes annotated to within 200bp upstream of the TSS of the gene. (3) 5'UTR; probes annotated to the 5-prime untranslated region of the gene. (4) 1stExon; probes annotated to the first exon of the gene. (5) Body; other probes annotated to the gene body. (6) 3'UTR; probes annotated to the 3-prime untranslated region of the gene.

Separately for each of these genomic regions, the variance of methylation levels for each gene for probes mapping to the genomic region in question was calculated. Then the effectiveness of each genomic region at discriminating between healthy and tumour tissue was compared, by considering the correlation of the tissue sample phenotype to the methylation variance measure in terms of distributions of per-gene AUCs; the 'Body' (gene body) genomic region was chosen for the methylation variance measure, as it performed best in 13 out of 14 data sets. This methylation variance was calculated for each gene for which there was 'Body' annotation information available, to give a single intra-gene methylation predictor value $x_j(g)$, for each gene g, for each sample j.

It should be noted, however, that in general for each gene there were far more probes annotated as 'Body' than for other genomic regions (table 2.4), leading to better estimates of the methylation variance for this region. Therefore, the relative greater effectiveness of this genomic region in this comparison does not necessarily imply biological significance. The minimum number of probes to be able to calculate the methylation variance for a given gene

and genomic region was set to be 3, and the methylation variance was not calculated for any gene with any fewer annotated probes than this for a given genomic region. As there were more genes with at least 3 probes annotated to the 'Body' region (table 2.4), it would be expected that there would be more genes which significantly associate with phenotype when this genomic region is used, which is likely to be another reason it performs better, without relevance to biological significance.

	TSS1500	TSS200	5'UTR	1stExon	Body	3'UTR
Mean no. probes	2.7	2.4	2.5	1.5	7	0.82
Median no. probes	2	2	1	1	3	1
No. probes, 95% CI	(0-10)	(0-7)	(0-13)	(0-6)	(0-39)	(0-4)
No. genes with min 3 probes	8512	7570	5258	3734	10029	958
No. genes with min 1 probe	14259	12979	11408	12194	15858	10291
No. genes with 0 probes	4013	5293	6864	6078	2414	7981

Table 2.4: *Number of probes per genomic region and gene, of 18272 annotated genes*

To choose which region to use to calculate the mean methylation measure, the same procedure was followed as for the methylation variance measure; the 'Body' genomic region was similarly chosen as this region correlated best with cancer/healthy phenotype in 10 out of 14 data sets. This mean methylation measure was calculated for each gene for which there was 'Body' annotation information available, to give a single intra-gene methylation predictor value $x_j(g)$, for each gene g, for each sample j. It is again worth noting that it is likely to be be due to the greater number of probes per gene annotated to 'Body', and the corresponding increase in accuracy of the calculated estimates of the mean methylation, which leads to this genomic region being more effective in this comparison, rather than there being any biological significance to this finding. In the case of mean methylation, it was only required that there be one probe annotated to a genomic region to allow a mean methylation level to be represented for that genomic region for that gene, as methylation levels of neighbouring CpGs within the same genomic region are expected to be highly correlated; again, there were more genes with at least one probe annotated to the 'Body' region than the other regions (table 2.4), similarly suggesting a reason for its better performance other than biological significance.

2.4.3 Comparison of intra-gene methylation measures

Methylation measures were assessed according to the distributions of their per-gene AUCs. The AUC is the ROC (receiver-operator characteristic) 'area under curve' and is defined as the probability that a randomly chosen item from the 'positive' class will be scored higher than a randomly chosen item from the 'negative' class (Fawcett, 2006).

The same procedure was used for the main comparison of intra-gene methylation mea-

sures, for the choice of genomic region used in the methylation variance measure, and for the choice of genomic region used in the mean methylation measure. In this procedure, each data set was split half and half into a training and test set, maintaining the same proportion of cancer and healthy samples in both sets. Using only the training set, AUCs were calculated for all genes, and the top 1000 genes were selected as those with the best AUC. Then using the test set, an AUC was calculated for each of these top 1000 genes identified in the training set. For the mean z-score measure, the mean healthy methylation profiles and healthy methylation standard deviations calculated from the training set were used to calculate the z-scores for both the cancer and healthy samples in the test set. The distributions of these test-set AUCs were compared in distribution density plots and using the Kolmogorov-Smirnov test (figure 2.2).

2.4.4 Meta-analysis and gene-set enrichment analysis

A meta-analysis of the fourteen data sets was carried out. The mean across all data sets of the per-gene AUCs generated from the mean z-score measure was calculated for each gene. Significance was then assigned to each of these per-gene mean AUCs by similarly calculating null mean AUCs after permuting AUCs within data sets. This resulted in 4267 significant most unstable (MUs) meta-analysis genes with FDR q-value (Benjamini & Hochberg, 1995) less than 5%, i.e., those genes corresponding to the upper tail of the null mean AUC distribution, which are associated with a consistent difference between cancer and healthy phenotypes across tissue types. This permutation method also resulted in 2818 significant (FDR $q \leq 0.05$) significant least unstable (LUs) meta-analysis genes, i.e., those genes corresponding to the lower tail of the null mean AUC distribution, which were associated with least difference from healthy to cancer phenotype across tissue types.

To confirm the biological significance of the findings of this meta-analysis with reference to genes which are well known to be important in cancer biology, the MUs and LUs genes were tested for enrichment by genes which in ES cells carry the repressing/activating chromatin marks H3K27me3 (H3K27 ES genes), H3K4me3 (H3K4 ES genes) and bivalent (i.e., both H3K27me3 and H3K4me3 marks, Biv ES genes) and enrichment by PCGTs (ES cell polycomb group target genes) using the one-tailed Fisher's exact test. A more general gene-set enrichment analysis (GSEA) was also carried out both on the MUs and LUs genes; 6811 gene set definitions were downloaded from the Broad Institute Molecular Signatures Database *http://www.broadinstitute.org/*, and each gene set was tested separately for enrichment among the significant genes. Enrichment was again tested using the one-sided Fisher's exact test, finding 1048 and 778 gene sets significantly (FDR $q \leq 0.05$) enriched by MUs and LUs meta-analysis genes respectively.

2.4.5 Correlation of tumour gene expression with intra-gene methylation architecture

For the 217 BRCA tumour samples for which matched gene expression and methylation data were available, for each gene a multivariate regression analysis of gene expression and intra-gene methylation architecture was carried out. Gene expression was used as the response, with one of mean z-score, mean derivative and methylation variance as one covariate predictor, and with mean methylation as a second covariate predictor. As it was expected that this relationship would be non-linear, and as for a non-specified non-linear monotonic function the ranks of data points in response and predictor variables are linearly related if there is a good association between these variables, the ranks of each of the variables across the samples were correlated to one another, as follows.

Defining for gene g the ranks of the samples according to the expression data as $\mathbf{r}^{(e)}(g)$, the ranks of the samples according to the mean z-score, mean derivative or methylation variance as $\mathbf{r}^{(x)}(g)$, and the ranks of the samples according to the mean methylation as $\mathbf{r}^{(m)}(g)$, the data were modelled according to equation 2.4:

$$\mathbf{r}^{(e)}(g) = \alpha(g)\mathbf{r}^{(x)}(g) + \gamma(g)\mathbf{r}^{(m)}(g) + \mu(g) + \boldsymbol{\epsilon} \tag{2.4}$$

where $\mu(g)$ is the intercept term for gene g, and $\boldsymbol{\epsilon}$ is the model error. Where $\mathbf{r}^{(e)}(g)$ is well-correlated with $\mathbf{r}^{(x)}(g)$, similar integer entries in these vectors (corresponding to similar ranks) will appear in similar positions in these vectors (N.B., these vectors are not themselves ordered). This will then be reflected as a small p-value for this comparison (calculated from the corresponding t-statistic for the linear model $\alpha(g)$ coefficient), and similarly for $\mathbf{r}^{(m)}(g)$ (and corresponding $\gamma(g)$ coefficient), if it is well-correlated with $\mathbf{r}^{(e)}(g)$.

This linear model was applied to the data for each gene present in the matched expression and methylation data for the BRCA dataset. 'Body' annotated probes were again used to calculate the methylation variance and mean methylation measures as used in this model, because probes annotated to this genomic region produced, in both cases, the greatest number of significant p-values (for the respective covariate), as compared to using probes annotated to each of the other genomic regions.

2.4.6 Association of clinical outcome with intra-gene methylation architecture

A childhood B-cell acute lymphoblastic leukaemia data set, with corresponding clinical outcome data in the form of binary recurrence / non-recurrance status (5 and 24 samples respectively, together with 4 healthy samples), was downloaded from the Gene Expression Omnibus *http://www.ncbi.nlm.nih.gov/geo/* under accession number GSE39141, and these data were sim-

ilarly pre-processed. The Elastic Net method (Zou & Hastie, 2005) was then used to predict recurrence status in a test set, following model fitting to a training set.

The data were split into test set and training set, with the model fitted to the training set and the accuracy of this fit assessed using the test set, according to the 'leave two out cross validation' strategy (Herzberg & Tsukanov, 1986). This method proceeds by systematically selecting the test set as one item from the 'positive' class (recurrence, in this case) and one item from the 'negative' class (non-recurrance, in this case), before fitting the model to the remainder of the data as the training set. This model fit is used to score the two test set items (with their actual recurrance/non-recurrance class unseen), resulting in two possibilities: either the item from the 'positive' class scores more highly than the item from the 'negative' class, or it doesn't. This process is repeated until all possible combinations of pairs of samples (as the test set) have been exhausted, each pair consisting of one sample from the 'positive' class and one from the 'negative' class. Then the proportion, of all test set combinations, in which the item from the 'positive' class scores higher than the item from the 'negative' class, is taken as the AUC, fulfilling the definition of the AUC as the probability that a randomly chosen item from the 'positive' class will be scored higher than a randomly chosen item from the 'negative' class (Fawcett, 2006). Viewed alternatively, it is valid to take what is in effect the mean of a large number of test set AUCs, because comparisons of sample scores are only ever made between items in the same test set which have been calculated using the same corresponding training set model fit.

Chapter 3

Time-series and Network Modelling of the DNA Methylation Epigenome of Differentiating Human Glioblastoma and Healthy Neural Stem Cells

3.1 Introduction

Glioblastoma is the most common type of brain tumour, and is invariably lethal (Surawicz *et al.* , 1999). Even after radiation and chemo therapies in combination with surgery, median life expectancy is only around 12 to 14 months (Surawicz *et al.* , 1998; Ballman *et al.* , 2007), and this very poor prognosis has not changed much over the last 20 years (Chen *et al.* , 2011). The stem-cell model of cancer (Kleinsmith & Pierce, 1964), or more correctly tumour initiating cell or stem-like cell as these cells may derive from differentiated cell types, has been demonstrated to be an applicable model for glioblastoma (Singh *et al.* , 2004), giving rise the to the study of glioblastoma stem-like cells. Differentiation of brain tumour initiating cells greatly reduces their tumorigenicity (Piccirillo *et al.* , 2006), indicating the need to study in more detail the dynamic changes in gene regulation of glioblastoma stem-like cells, as they differentiate. A genetic 'hit' does not necessarily lead to runaway cell proliferation (as in cancer) because of cell heritable gene transcriptional regulation, i.e., epigenetic mechanisms (Carén *et al.* , 2013). Hence, the study of dynamic changes in the DNA methylation epigenome of differentiating glioblastoma stem-like cells.

Time-series methods, in comparison to methods which only consider two experimental conditions (e.g., measurements taken at the beginning and end of an experiment) are able to analyse more precisely the dynamic behaviour of quantities of interest during the experiment. They are able to use the uncertainties across all time points (as well as experimental replicates)

for statistical inference, leading to more reliable estimates of which features (out of a potentially very large number measured) change significantly and in a regular/co-ordinated way during the experiment. The time-series modelling approach also allows smoothed estimates to be made of the inferred change in a quantity (such as methylation level) between any two time points during the experiment, based on the model as fitted to all the data points, rather than directly comparing the measured quantity at only those time points, thus reducing the uncertainty associated with such contrasts.

3.2 Methods and models

3.2.1 Data collection and preparation

Three human glioblastoma stem-like cell lines (G19, G26 and G144, referred to here as GNS1, GNS2 and GNS3 respectively) and one healthy human neural stem cell line as a control (Cb130, referred to here as NS) were induced to differentiate by treatment with bone morphogenetic protein (BMP). Average methylation level β values were collected at six time points during the experiment (from 0 to 64 days inclusive) using the Illumina Infinium HumanMethylation450 platform. These measurements represent time-series of the DNA methylation profiles of the differentiating glioblastoma and healthy neural stem cells.

These data were first pre-processed by removing individual data values with corresponding detection p-value greater than 0.05, before removing all data for CpGs with corresponding coverage less than 95% across samples. Any remaining missing values (in total, 0.24% of the data set) were then replaced by KNN (k nearest neighbour) imputation ($k = 5$). The data were also checked for batch effects by hierarchical clustering and correlation of the significant principle components with phenotype and batch: no significant batch effects (which would warrant further correction) were found.

3.2.2 Time-series modelling using spline curves

The method of Storey (Storey *et al.* , 2005) was used to model the time-series of the methylation profiles of the differentiating cells. Originally, this method was developed in the context of gene expression data, however due to the analogous nature of the problem, the core methodology was taken to be a suitable way to approach these data, and was re-implemented here from the original algebraic descriptions of the method, in the R programming language.

The foundation of Storey's method is the use of spline basis functions to model the data, and a brief summary of the method, as it is applied here, now follows; for more details the reader is referred to the original paper (Storey *et al.* , 2005). It is assumed that the biological replicates j lead to noisy measurements $y_{ij}(t)$ of a 'true' methylation time-course $\mu_i(t)$ for CpG

i, which can be modelled according to equation 3.1,

$$y_{ij}(t) = \mu_i(t) + \gamma_{ij} + \epsilon_{ij}(t) \tag{3.1}$$

where $\epsilon_{ij}(t)$ represents the noise, or random error, and γ_{ij} represents the systematic deviation from the 'true' methylation time-course $\mu_i(t)$ for replicate j, and CpG i. Equation 3.1 can be discretised as equation 3.2 to express more specifically the scenario modelled in this investigation, that of discrete observations, y_{ijk}, at particular points in time, k, with $k \in \{1, 2, ...K\}$.

$$y_{ijk} = \mu_{ik} + \gamma_{ij} + \epsilon_{ijk} \tag{3.2}$$

Noting that γ_{ij} is not time dependent (equations 3.1 and 3.2), i.e., it is assumed that the systematic deviation from the 'true' methylation time-course is constant for a particular CpG and replicate, this term can be subtracted and instead the mean-centred methylation time-course can be modelled, as in equation 3.3,

$$y^{(c)}_{ijk} = \mu^{(c)}_{ik} + \epsilon^{(c)}_{ijk} \tag{3.3}$$

where

$$y^{(c)}_{ijk} = y_{ijk} - \bar{y}_{ij},$$

$$\bar{y}_{ij} = \frac{1}{k}\sum_{k=1}^{K} y_{ijk}$$

and

$$\mu^{(c)}_{ik} = \mu_{ik} - \bar{\mu}_i,$$

$$\bar{\mu}_i = \frac{1}{K}\sum_{k=1}^{K} \mu_{ik}$$

where $\epsilon^{(c)}_{ijk}$ is random error. Note that $\bar{y}_{ij}$ and $\bar{\mu}_i$ are simply the time-averages of the measured and 'true' methylation profiles, respectively. This model only allows us to assign significance according to change in methylation (in comparison to a null model of 'no change in methylation'), a restriction which is appropriate for the purposes of this investigation. The advantage of this model is that now the measurements for each replicate can be expected to deviate from the 'true' mean-centred methylation time-course $\mu^{(c)}_{ik}$ by only the random error $\epsilon^{(c)}_{ijk}$, and hence measurements for all replicates can be combined into one model fit without the extra model parameter γ which would complicate inference. The discretised, mean-centred model equation

3.3 can be expressed more succinctly in vector form, equation 3.4.

$$\mathbf{y}_{ij}^{(c)} = \boldsymbol{\mu}_i^{(c)} + \boldsymbol{\epsilon}_{ij}^{(c)} \tag{3.4}$$

As per the method of Storey (Storey *et al.* , 2005), the 'true' mean-centred methylation profile, $\boldsymbol{\mu}_i^{(c)}$, is modelled by spline basis functions of order P, according to equation 3.5,

$$\mathbf{y}_{ij}^{(c)} = \mathbf{S}^{(c)}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij}^{(c)} \tag{3.5}$$

where $\mathbf{S}^{(c)}$ is a $K \times P$ matrix (which is the same for each CpG i) representing the values of the mean-centered spline basis functions at each time point, and $\boldsymbol{\alpha}_i$ are the model coefficients for CpG i. The $k \in \{1, 2, ..., K\}$ are the indices of the time-points t_k at which the measurements are taken, which for this investigation are $\{t_1...t_K\} = \{0, 8, 16, 24, 32, 48, 64\}$ for NS, GNS1 and GNS2 cell-lines and $\{t_1...t_K\} = \{0, 8, 15, 32, 47, 63\}$ for the GNS3 cell-line with all times measured in days. Also $j \in \{1, 2\}$ for NS, GNS1 and GNS2 cell-lines and $j \equiv 1$ for GNS3; for NS, GNS1 and GNS2 cell-lines the model is fitted to both replicates j simultaneously.

3.2.3 Determining the order of the spline basis function

To determine the order, P, of the spline basis functions used in the model (equation 3.5), for each cell-line a singular value decomposition was taken of the data matrix. The resulting right singular vectors can be thought of as 'eigen-CpGs' (Alter *et al.* , 2000), and those which are significant (i.e., are associated with the significant component of the variation in the data) were selected by estimating the dimensionality d of the data matrix (for each cell-line) using random matrix theory (Plerou *et al.* , 2002), and then selecting the eigen-CpGs associated with the d largest singular values. The order of the spline basis was then determined by testing values of P from 2 to 5, using these to model each significant eigen-CpG according to equation 3.5, and identifying the value of P which minimised the model error for each eigen-CpG and each cell-line $\boldsymbol{\epsilon}_{i1}^{(c)} \cdot \boldsymbol{\epsilon}_{i1}^{(c)} + \boldsymbol{\epsilon}_{i2}^{(c)} \cdot \boldsymbol{\epsilon}_{i2}^{(c)}$ (using the dot product notation $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b}$). The order, P, for the spline basis was then set to be the same for NS, GNS1 and GNS2 (as these have the same data structure, with two replicates j); P was set for these cell lines as the largest value identified in this way for any of these cell-lines and eigen-CpGs, $P = 4$. GNS3 has a different data structure with only one replicate, and so P was set independently for this cell line, $P = 3$.

3.2.4 Assigning significance

The alternative model (equations 3.4 and 3.5) is compared to the null model that there is no methylation change over the time-course, equation 3.6, to assign significance.

$$\mathbf{y}_{ij}^{(c)} = \boldsymbol{\epsilon}_{ij}^{(0)} \tag{3.6}$$

Significance is assigned for each CpG i by calculating an *F*-statistic, F_i (equation 3.7), comparing these null and alternative models,

$$F_i = \frac{RSS_i^{(0)} - RSS_i^{(c)}}{RSS_i^{(c)}} \tag{3.7}$$

where $RSS_i^{(c)}$ and $RSS_i^{(0)}$ are the residual sum of squares for the alternative and null models respectively, i.e. (again using the dot product notation $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b}$)

$$RSS_i^{(c)} = \boldsymbol{\epsilon}_{i1}^{(c)} \cdot \boldsymbol{\epsilon}_{i1}^{(c)} + \boldsymbol{\epsilon}_{i2}^{(c)} \cdot \boldsymbol{\epsilon}_{i2}^{(c)}$$

and

$$RSS_i^{(0)} = \boldsymbol{\epsilon}_{i1}^{(0)} \cdot \boldsymbol{\epsilon}_{i1}^{(0)} + \boldsymbol{\epsilon}_{i2}^{(0)} \cdot \boldsymbol{\epsilon}_{i2}^{(0)}$$

for NS, GNS1 and GNS2 cell-lines and

$$RSS_i^{(c)} = \boldsymbol{\epsilon}_{i1}^{(c)} \cdot \boldsymbol{\epsilon}_{i1}^{(c)}$$

and

$$RSS_i^{(0)} = \boldsymbol{\epsilon}_{i1}^{(0)} \cdot \boldsymbol{\epsilon}_{i1}^{(0)}$$

for the GNS3 cell-line.

Because the errors $\boldsymbol{\epsilon}_{ij}^{(c)}$ and $\boldsymbol{\epsilon}_{ij}^{(0)}$ can be expected to be correlated in time for many i, significance is assigned by a permutation method, instead of calculated directly from the *F*-statistic (equation 3.7). To generate a null *F*-statistic distribution, samples are randomly taken with replacement from the alternative model fit errors, as these will resemble a residual distribution irrespective of whether the null or alternative model is true (Storey *et al.* , 2005); for repetition b, this leads to null time-course $\mathbf{y}_{ijb}^{(B)}$ for CpG i. Hence the value of the null time-course, $y_{ijkb}^{(B)}$, at each time-point t_k, is a sample taken with replacement from $\epsilon_{ijk}^{(c)}$ with $j \in 1, 2$ for NS, GNS1 and GNS2 or $j \equiv 1$ for GNS3 and $k \in \{1, ..., K\}$. A null *F*-statistic $F_{ib}^{(B)}$ is then calculated for CpG i and repetition b from the null time-course $\mathbf{y}_{ijb}^{(B)}$ in exactly the same way as the observed

F-statistic F_i is calculated for the observed data. This is repeated 250 times, with $b = 1, .., 250$, to generate the null F-statistic distribution for CpG i, with significance then assigned according to equation 3.8,

$$p_i = \frac{1}{250} \sum_{b=1}^{250} \mathbb{1} \left(F_{ib}^{(B)} > F_i \right) \tag{3.8}$$

where p_i is a conventional p-value. For each cell-line, this process of modelling and assigning significance is carried out independently for 482421 CpGs, and the calculated values of p_i are then converted to FDR q-values, q_i, according to the method of Benjamini and Hochberg (Benjamini & Hochberg, 1995). CpGs are classified as significant if $q_i < 0.05$ and if modelled change in methylation level $\Delta\hat{\beta}_i > 0.2$, where

$$\Delta\hat{\beta}_i = \max_{k \in \{1,...,K\}} \hat{y}_{ik}^{(c)} - \min_{k \in \{1,...,K\}} \hat{y}_{ik}^{(c)}$$

and $\hat{\mathbf{y}}_i^{(c)}$ is the modelled mean-centred methylation time-course for CpG i, which at time t_k has the value $\hat{y}_{ik}^{(c)}$, with

$$\hat{\mathbf{y}}_i^{(c)} = \hat{\boldsymbol{\mu}}_i = \mathbf{S}^{(c)} \hat{\boldsymbol{\alpha}}_i$$

where $\hat{\boldsymbol{\alpha}}_i$ are the fitted model coefficients. The threshold of 0.2 was set here after consultation with the experimental biologist who generated these data. It is based on her best judgement of a threshold above which a change in methylation level would be expected to be biologically meaningful, in this context. This identifies the number of significant CpGs, for each cell line, as shown in table 3.1.

NS	GNS1	GNS2	GNS3
5678	1235	37656	11737

Table 3.1: *Number of Significant CpGs*
Comparison of the numbers of CpGs (out of 482421) found as significant ($q < 0.05$ and $\Delta\hat{\beta} > 0.2$)

3.2.5 Identification of a glioblastoma stem-like cell differential epigenotype

To understand how the tumourigenicity of glioblastoma stem-like cells is so greatly reduced when they are induced to differentiate (Piccirillo *et al.* , 2006), it is of particular interest and relevance to study CpGs with methylation levels which are different in glioblastoma stem-like cells compared to healthy stem cells. These tumour-associated methylation profiles are expected to change towards the levels of healthy cells, which will remain at similar methylation levels; i.e., the tumour-associated aberrations will be normalised against the healthy profiles. Such CpGs could be thought of as characterising a glioblastoma stem-like cell differential epigenotype, contrasted against the healthy epigenotype. If glioblastoma stem-like cells are the drivers of

the tumour, then differentiation of all glioblastoma stem-like cells would be expected to stop tumour progression. Therefore, identifying such a characteristic glioblastoma stem-like cell differential epigenotype could lead to novel insights into this disease.

A list of 100145 CpGs previously found as significantly differentially methylated ($q < 0.05$) between healthy cells (3 cell lines) and glioblastoma stem-like cells (11 cell lines) were additionally filtered according to the following criteria:

1. CpGs found as significant ($q < 0.05$) in the time-course modelling of glioblastoma stem-like cells

2. CpGs which change methylation level by more than 0.2 during the experiment, in glioblastoma stem-like cells

3. CpGs with methylation level different by no more than 0.3 in glioblastoma stem-like cells and healthy stem cells at the end of the experiment

4. CpGs which change methylation level by less than 0.2 during the experiment in healthy neural stem cells

Criteria 2-4 were similarly set here after consultation with the experimental biologist who generated these data, based on her best judgement of biologically meaningful thresholds for this context. In order to apply these criteria, for CpGs found as significant, methylation levels $\hat{\mathbf{y}}_i$ inferred from the alternative model fits of the time-courses for each significant CpG were used, calculated according to equation 3.9.

$$\hat{\mathbf{y}}_i = \hat{\mathbf{y}}_i^{(c)} + \frac{\bar{y}_{i1} + \bar{y}_{i2}}{2}, \tag{3.9}$$

$$\hat{\mathbf{y}}_i^{(c)} = \boldsymbol{\mu}_i^{(c)} = \mathbf{S}^{(c)}\hat{\boldsymbol{\alpha}}_i,$$

$$\bar{y}_{ij} = \frac{1}{k}\sum_{k=1}^{K} y_{ijk}.$$

For CpGs which were not found as significant, methylation levels $\hat{y}_i^{(0)}$ inferred from the null model of the time-courses were used, as defined in equation 3.10:

$$\hat{y}_i^{(0)} = \frac{\bar{y}_{i1} + \bar{y}_{i2}}{2}. \tag{3.10}$$

In order to take account of the uncertainty estimated by the time-course modelling, these criteria were applied also subject to an uncertainty envelope. It is difficult to predict how these model

errors are distributed, and hence no prediction is given of the quantile range of any underlying distribution covered by this envelope. Instead, a small number multiple of the estimated model standard deviation was used for this envelope, and as it lead to consistent and relevant results, this was retained as an acceptable approach. The uncertainty envelope is defined as $\hat{\mathbf{y}}_i \pm 2\hat{\sigma}_i$, where $\hat{\sigma}_i$ is the estimated model error, defined for significant CpGs (i.e., those for which the alternative model is chosen) as:

$$\hat{\sigma}_i^{(m)} = \sqrt{\frac{\boldsymbol{\epsilon}_{i1}^{(c)} \cdot \boldsymbol{\epsilon}_{i1}^{(c)} + \boldsymbol{\epsilon}_{i2}^{(c)} \cdot \boldsymbol{\epsilon}_{i2}^{(c)}}{2K}}, \tag{3.11}$$

and defined for not significant CpGs (i.e., those for which the null model is chosen) as:

$$\hat{\sigma}_i^{(0)} = \sqrt{\frac{\boldsymbol{\epsilon}_{i1}^{(0)} \cdot \boldsymbol{\epsilon}_{i1}^{(0)} + \boldsymbol{\epsilon}_{i2}^{(0)} \cdot \boldsymbol{\epsilon}_{i2}^{(0)}}{2K}}. \tag{3.12}$$

Hence, incorporating this uncertainty envelope, criteria 2 requires, for the GNS cell-line:

$$\left[\left(\max_{k\in\{1,\ldots,K\}} \hat{y}_{ik}^{(c)} - 2\hat{\sigma}_i^{(m)}\right) - \left(\min_{k\in\{1,\ldots,K\}} \hat{y}_{ik}^{(c)} + 2\hat{\sigma}_i^{(m)}\right)\right] > 0.2$$

Incorporating the uncertainty envelope, criteria 3 then requires:

$$\left|\left(\hat{y}_{iK}^{(GNS)} \pm 2\hat{\sigma}_i^{(GNS)}\right) - \left(\hat{y}_{iK}^{(NS)} \pm 2\hat{\sigma}_i^{(NS)}\right)\right| < 0.3$$

where $\hat{y}_{iK}^{(GNS)}$ and $\hat{\sigma}_i^{(GNS)}$ are the alternative model estimated methylation level $\hat{y}_{iK}^{(m)}$ at the end of the experiment (i.e., the time when $k = K$, $t_K = 64$ days) and model error $\hat{\sigma}_i^{(m)}$ for the GNS cell-line for CpG i, and $\hat{y}_{iK}^{(NS)}$ and $\hat{\sigma}_i^{(NS)}$ are the estimated methylation level at the end of the experiment and model error for CpG i for the NS cell-line according to the alternative model $\hat{y}_{iK}^{(m)}$ and $\hat{\sigma}_i^{(m)}$ if $q_i^{(NS)} < 0.01$ or according to the null model $\hat{y}_i^{(0)}$ and $\hat{\sigma}_i^{(0)}$ otherwise. Incorporating the uncertainty envelope, criteria 4 then requires for the NS cell-line, for $q_i^{(NS)} < 0.01$:

$$\left[\left(\max_{k\in\{1,\ldots,K\}} \hat{y}_{ik}^{(c)} + 2\hat{\sigma}_i^{(m)}\right) - \left(\min_{k\in\{1,\ldots,K\}} \hat{y}_{ik}^{(c)} - 2\hat{\sigma}_i^{(m)}\right)\right] < 0.2$$

and otherwise:

$$2 \times 2\hat{\sigma}_i^{(0)} < 0.2$$

which reflects the need for some certainty that when the null model is chosen, the methylation levels really do change in the required range over the duration of the experiment. The num-

bers of CpGs identified in this way as being part of the glioblastoma stem-like cell differential epigenotype, according to each GNS cell-line, are shown in table 3.2.

	GNS1	GNS2	GNS3
No. CpGs	28	1521	639
No. annot. to genes	17	978	425
No. genes	17	769	341

Table 3.2: *Number of CpGs of the glioblastoma stem-like cell differential epigenotype. Comparison of the numbers of CpGs identified as part of the glioblastoma stem-like cell differential epigenotype according to each cell-line, the subset of these which are annotated to genes, and the number of annotated genes represented among these CpGs.*

3.2.6 Network model of the glioblastoma stem-like cell differential epigenotype

A network model of the glioblastoma stem-like cell differential epigenotype was produced, by filtering a list of pairs of genes corresponding to known biochemical interactions in humans downloaded from *http://www.pathwaycommons.org*, with the list of genes identified as being part of the glioblastoma stem-like cell differential epigenotype. Each pair of genes in the downloaded list represents a pair of genes known to take part in a biochemical interaction or process (if a particular interaction or process involves three genes, this will appear as three pairs in the list, etc). Pairs of genes in the list were only retained in the filtered list if both genes of the pair have at least one CpG annotated which was identified as being part of the glioblastoma stem-like cell differential epigenotype. Each pair of genes in the filtered list was then cross-checked to see if it connected to any other pairs in the filtered list, forming connected components where possible. This filtering and connecting was carried out independently for each glioblastoma cell line, with no connected components larger than two genes found for either the GNS1 or GNS3 cell-lines (i.e., none of the pairs in the filtered list had any genes in common for these cell lines). However, a connected component consisting of 32 genes was found for the GNS2 cell line, and this was used as the network model of the glioblastoma stem-like cell differential epigenotype.

3.3 Results

The network model of the glioblastoma stem-like cell differential epigenotype, as identified for the GNS2 cell line, is shown in figure 3.1. A subnetwork containing the 18 most connected and relevant genes was produced by removing all ‘tendrils’ (by removing all genes which are connected to at most only one other gene and repeating this until no such genes remain) with the exception of those including *HOXD4*, *PAX6* and *EZH2* (as these genes are of particular interest and relevance to stem cell biology and DNA methylation); this subnetwork is shown in the same figure. The methylation time-courses for all CpGs identified as part of the glioblastoma

stem-like cell differential epigenotype for the GNS2 cell-line and which are annotated to genes which form part of the subnetwork (figure 3.1 (b)) are shown in figure 3.2, for all cell lines.

The subnetwork (figure 3.1 (b)) contains a number of genes which are of great interest and relevance to glioblastoma and developmental biology, in particular *WT1*, *STAT3*, *HOXD4*, *EZH2*, *P73*, *PAX6*, *VIMENTIN* (*VIM*), and *CBP* (*CREBBP*); a review of all genes and interactions appearing in this subnetwork now follows.

STAT3 is typically thought of as an oncogene; its protein-product up regulates growth promoting genes such as *MYC* as well growth inhibitors such as *P21WAF1*, with this balancing inhibition lost during oncogenic transformation in glioblastoma (Barré *et al.* , 2003). In particular mutational circumstances in glioblastoma however, STAT3 can also function as a tumour suppressor, for example in the context of *PTEN* loss (De La Iglesia *et al.* , 2008). STAT3 predominantly resides in the cytoplasm of unstimulated cells; when it is activated, it translocates to the cell nucleus, and binds to the target gene DNA, promoting their expression, a process which is likely to include recruitment of the histone acetyltransferase CBP as a co-activator (Wang *et al.* , 2005). CBP is crucial as a co-activator in relation to multiple genes involved in the glioblastoma stem-like cell differential epigenotype (figure 3.1), and one of these which is particularly relevant to glioblastoma is *WT1*.

As well as being co-activated by CBP (Wang *et al.* , 2001), *WT1* is an oncogene and its expression is required for the viability of multiple tumours of the brain. WT1 is a regulator of cell cycle and growth factors as well as apoptosis; it is expressed in at least 80% of glioblastoma specimens and is not expressed in healthy glial cells (Nakahara *et al.* , 2004; Oji *et al.* , 2005; Chen *et al.* , 2011), and a pharmaceutical therapy directly targeting *WT1* has reached phase II of clinical trials (Izumoto *et al.* , 2008). *WT1* is also involved in a mutually repressive interaction with the tumour suppressor *P73* (Scharnhorst *et al.* , 2000). *P73* is additionally thought to interact with the master developmental regulator *PAX6* (Lunardi *et al.* , 2010), and the transcriptional promoting functions of P73 again rely on co-acvivation by CBP (Lemasson & Nyborg, 2001). *CBP* itself is directly regulated by the master developmental regulator HOXD4, the activity of this and other Hox proteins having been shown to inhibit *CBP in vivo*, while not themselves being acetylated (and hence not co-regulated) by CBP (Shen *et al.* , 2001).

A key and recently uncovered link from *WT1* to developmental biology, highlighted here as an interaction in the glioblastoma stem-like cell differential epigenotype, concerns the polycomb-group (PcG) protein EZH2 (Xu *et al.* , 2011). EZH2 forms part of the polycomb repressor complex-2 (PRC-2), and PRC-2 is involved in gene repression, particularly in relation to stem cell genes which may or may not become activated later in development according to

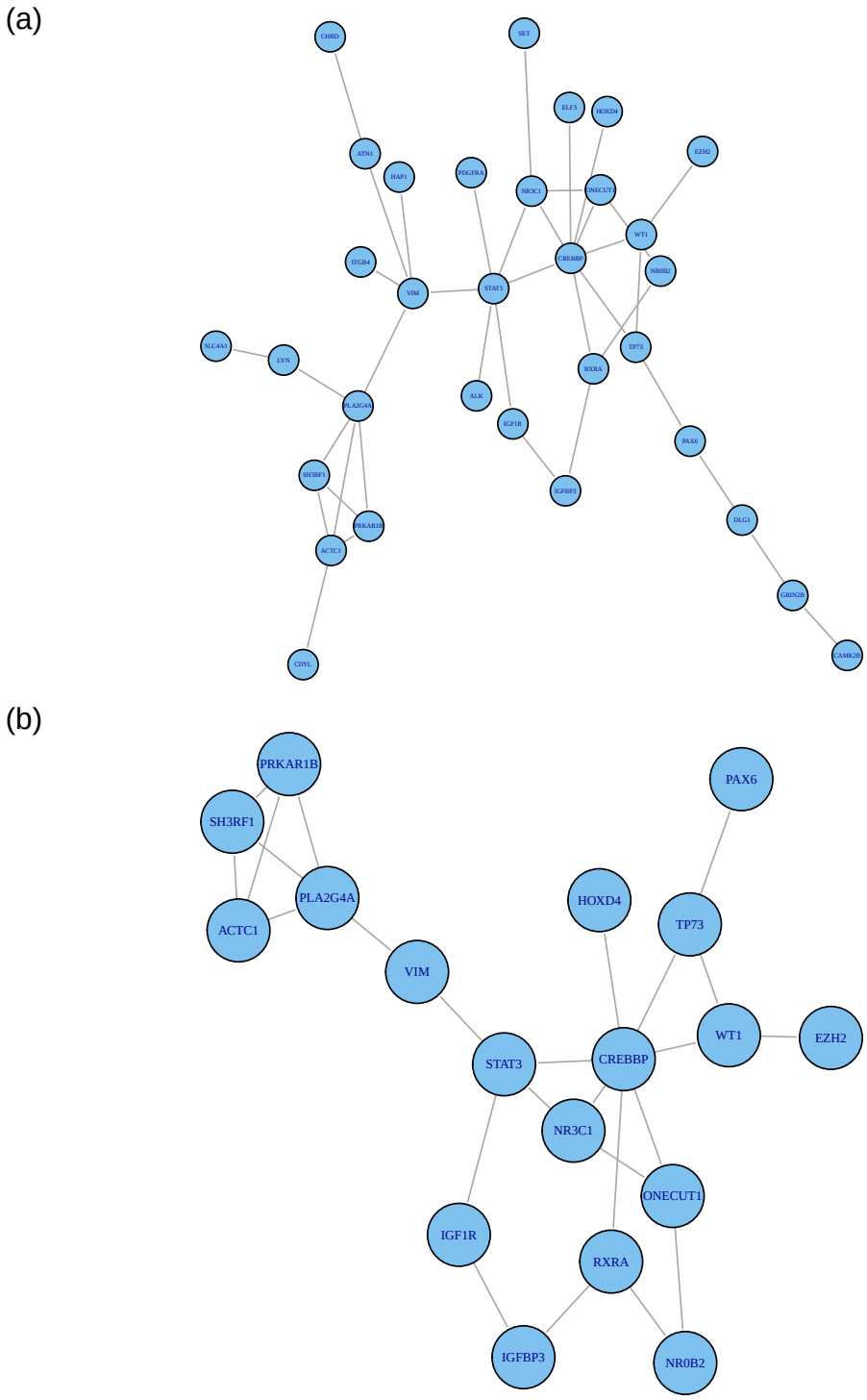


Figure 3.1: *Glioblastoma stem-like cell differential epigenotype network model*
(a) The network model of the glioblastoma stem-like cell differential epigenotype, as identified for the GNS2 cell line. (b) The same network model with tendrils removed, with the exception of those including HOXD4, PAX6 and EZH2.

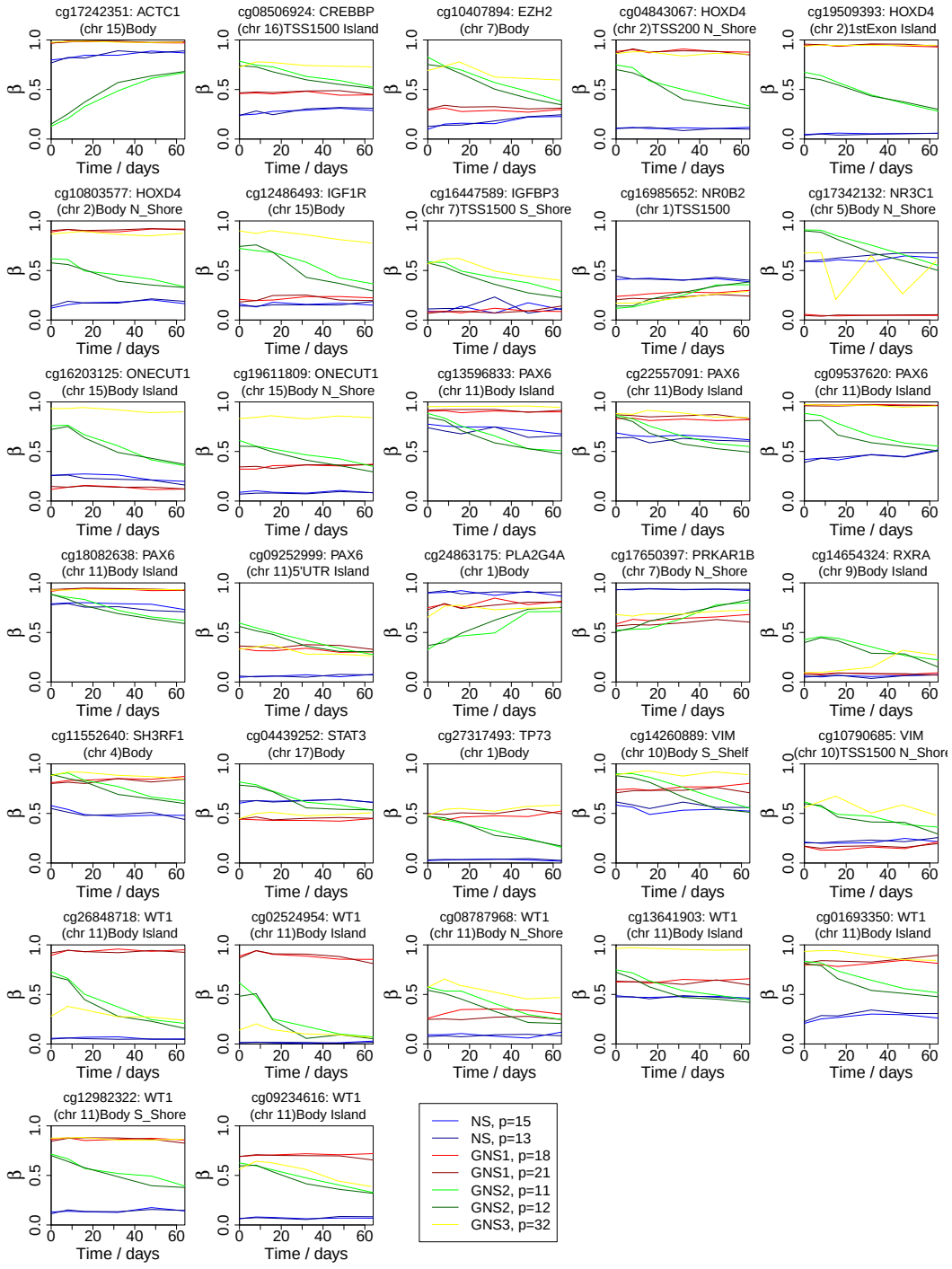


Figure 3.2: *Glioblastoma stem-like cell differential epigenotype methylation time-courses*

Methylation time-courses for all cell lines and experimental replicates, for all CpGs identified as being part of the glioblastoma stem-like cell differential epigenotype for the GNS2 cell-line, which are annotated to genes which are in the subnetwork shown in figure 3.1 (b). Annotation information for each CpG is shown above the time-course plots, relating to annotated gene, chromosome, genomic region location, and location relative to CpG island. The variable 'p' which appears in the legend indicates the number of 'passages' each experimental replicate went through during the experiment, and is referred to here only to distinguish between the experimental replicates.

cell type specialisation (Lee *et al.* , 2006). PRC-2 is responsible for the trimethylation of lysine 27 of histone 3, leading to the repressive epigenetic mark H3K27me3, which is then 'read' by polycomb repressor complex-1 (PRC-1), stabilising the chromatin in a compact state, such that the DNA does not get transcribed (Bickmore, 2012). WT1 interacts with both EZH2 and another component of PRC-2, SUZ12, as well as the DNA methyltransferase DNMT1, and EZH2 specifically methylates H3K27, with knockdown of *EZH2* resulting in down-regulated global H3K27me3 (Xu *et al.* , 2011).

DNMT1 is instrumental in CpG methylation (Yen *et al.* , 1992), and repressive CpG methylation of genes which are normally repressed in development by H3K27 methylation is known to be a key event in oncogenic transformation, leading to a return of stem-like cell characteristics (Easwaran *et al.* , 2012). So coupled with the finding that WT1 mediates interaction between DNMT1 and PRC-2 (Xu *et al.* , 2011), *WT1* is highlighted as fundamental to understanding glioblastoma biology, and the great reduction in the tumourigenicity of glioblastoma stem-like cells as they are induced to differentiate to specialised cell types which do not express *WT1*.

Further evidence for a link between *WT1* gene expression and CpG methylation of gene targets of PcG proteins (PCGTs) is provided by correlating *WT1* expression levels (Affymetrix data) to the methylation levels of CpGs annotated to PCGTs in TCGA glioblastoma samples (Spearman correlation test); there is a positive association for a significant number of CpGs for all genomic regions as shown by the concentration of p-values close to zero (figure 3.3).

IGFBP-3 regulates apoptosis, and has been shown to directly induce apoptosis in a number of different cancer cells (Sueoka *et al.* , 2000; Gill *et al.* , 1997), and RXR-α (RXRA) has a key role in the regulation of gene transcription (Solomin *et al.* , 1998), and is also co-activated by CBP (Gelman *et al.* , 1999). IGFBP-3 and RXR-α bind to each other within the nucleus and the RXR-α - IGFBP-3 interaction leads to modulation of the transcriptional activity of RXR-α and is essential for mediating the effects of IGFBP-3 on apoptosis (Liu *et al.* , 2000). Interestingly, the methylation levels of the CpGs identified for the genes encoding these proteins as part of the glioblastoma stem-like cell differential epigenotype are reduced over the course of the experiment in the GNS2 cell line (figure 3.2), towards the low methylation level in the NS cell lines, possibly indicating up regulation of these genes and restoring of normal tumour suppressive action in the differentiated state.

IGFBP-3 is thought to modulate signalling pathways activated by IGF-1R (Mohseni-Zadeh & Binoux, 1997), and IGF-1R is known to activate *STAT3* (Prisco *et al.* , 2001). GR (NR3C1) acts as a co-activator of *STAT3* (Zhang *et al.* , 1997; Lerner *et al.* , 2003), with *GR* again co-activated by CBP (Almlöf *et al.* , 1998; Peterson & Workman, 2000). HNF6 (ONECUT1) is

a cell-type specific transcription factor normally expressed only in the liver, which is again co-activated by CBP (Lannoy *et al.* , 2000), which inhibits the activity of GR (Pierreux *et al.* , 1999), and is repressed by SHP (NR0B2) (Lee *et al.* , 2008); SHP similarly functioning to repress *RXR-*α (Lee *et al.* , 2000).

STAT3 enhances the expression of the cytoskeletal protein VIMENTIN (VIM), which exhibits a complicated pattern of developmental and tissue specific expression (it is initially widely expressed in the embryo, but progressively restricted to fewer cell types during development), and is aberrantly expressed in most metastatic tumours, whatever their embryological origin (Wu *et al.* , 2004). VIMENTIN in turn regulates *PLA2G4A* (cytosolic phospholipase A2), and the remaining interactions indicated by that final branch of the subnetwork (figure 3.1) involving *ACTC1* (actin, alpha cardiac muscle 1), *PRKAR1B* and *SH3RF1*, are present in the subnetwork on the grounds of their protein products having been found to interact with c-MYC (which is one of the most potent activators of tumorigenesis, frequently overexpressed in diverse cancers (Agrawal *et al.* , 2010)). It is also worth noting that the CpG annotated to *ACTC1* which was identified as part of the glioblastoma stem-like cell differential epigenotype shows the largest beta change (0.54) of any CpG found as part of this glioblastoma stem-like cell differential epigenotype (figure 3.2), with methylation increasing in GNS2 from a low level at the start of the experiment to a high level similar to NS by the end of the experiment.

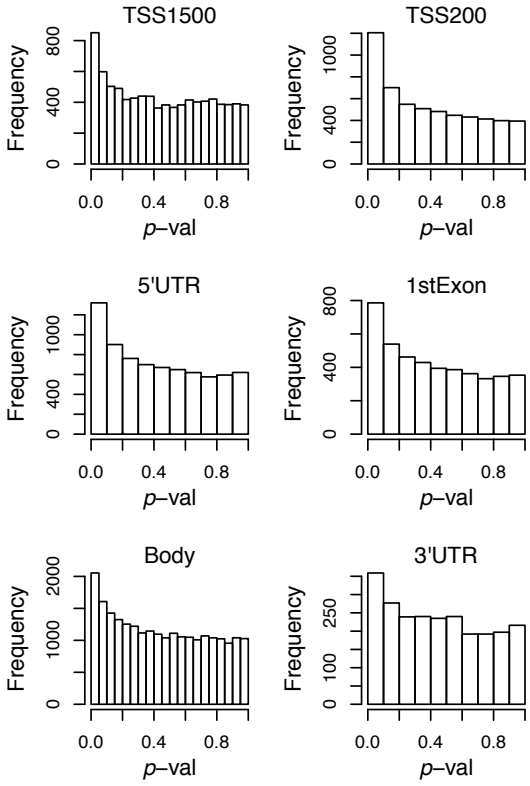


Figure 3.3: *Correlation of WT1 expression with PCGT methylation*

P-value histograms, showing the correlation of WT1 expression (Affymetrix data) to the methylation levels of CpGs annotated to PCGTs in TCGA glioblastoma samples (Spearman correlation test), for different genomic regions. There is a positive association for a significant number of CpGs for all genomic regions, as shown by the concentration of p-values close to zero.

3.4 Discussion

The time-series modelling approach used here has been effective, finding many CpGs with beta values which change in a significant and consistent way during the experiment, and rejecting as not significant CpGs with possibly large but more random changes in beta, such as can be seen in figure 3.2, NR3C1, for GNS3. The time-series modelling approach has also succeeded in finding significant CpGs on account of consistent behaviour throughout the experiment despite relatively large variance between experimental replicates, as in the case of the GNS1 cell line, for which 1235 CpGs were found as significant ($q < 0.05$) by this time-series approach. This contrasts with a LIMMA (Smyth *et al.* , 2004) analysis for this cell line of both experimental replicates but considering only two experimental conditions (the start and end of the experiment), which identified only 60 significant CpGs ($q < 0.05$).

The statistical methodology used to identify the glioblastoma stem-like cell differential epigenotype has also proven to be effective, with these results illustrated in the form of a 32 gene network model, which is further condensed to an 18 gene subnetwork (figure 3.1). This subnetwork contains a number of genes which are of great interest and relevance to glioblastoma and developmental biology, in particular *WT1*, *STAT3*, *HOXD4*, *EZH2*, *P73*, *PAX6*, *VIMENTIN* (*VIM*), and *CBP* (*CREBBP*).

WT1 is already well known for its oncogeneic function in glioblastoma; it is expressed in at least 80% of glioblastoma specimens and is not expressed in healthy glial cells (Nakahara *et al.* , 2004; Oji *et al.* , 2005; Chen *et al.* , 2011). A pharmaceutical therapy directly targeting *WT1* has also reached phase II of clinical trials (Izumoto *et al.* , 2008). WT1 has also recently been shown to mediate the interaction between the CpG methylating enzyme DNMT1 and PRC-2 (polycomb repressive complex-2), particularly relevant as CpG methylation of gene targets of polycomb group proteins (PCGTs) is thought to lead to the return of stem-like cell characteristics in cancer (Easwaran *et al.* , 2012).

IGFBP-3 regulates apoptosis including in cancer cells (Sueoka *et al.* , 2000; Gill *et al.* , 1997), and binding to RXR-α is necessary for this function (Liu *et al.* , 2000). Interestingly, the methylation levels of the CpGs identified for *IGFBP-3* and *RXR*-α as part of the glioblastoma stem-like cell differential epigenotype are reduced over the course of the experiment in the GNS2 cell line towards the low methylation level in the NS cell lines (figure 3.2), possibly indicating up regulation of these genes and restoring of normal tumour suppressive action in the differentiated state.

A network model such as figure 3.1 highlights biochemical interactions which might be of particular interest and relevance. Sequences and pathways of such interactions may act in

combination to form larger systems, and might be thought of in certain circumstances as genetic circuits. For example, IGFBP-3 is thought to modulate signaling pathways activated by IGF-1R (Mohseni-Zadeh & Binoux, 1997), and IGF-1R is known to activate *STAT3* (Prisco *et al.* , 2001) which is primarily an oncogene. On the other hand, *STAT3* is coactivated by GR (NR3C1) (Zhang *et al.* , 1997; Lerner *et al.* , 2003), but the activity of GR is paradoxically inhibited by HNF6 (ONECUT1) which is a cell-type specific transcription factor normally expressed only in the liver. Both *RXR*-α and *HNF6* are repressed by SHP (NR0B2) (Lee *et al.* , 2000, 2008), the first of these seeming to oppose apoptosis induced by IGFBP-3, and the second seeming to promote tumourigenesis via GR - STAT3. If *HNF6* is repressed by SHP this might be tumourigenic because HNF6 would then presumably be unavailable to repress *GR*, the protein product of which could then co-activate *STAT3*. However, as HNF6 is a liver specific transcription factor, it would be expected to be expressed only in cancerous cells of the brain (which might take on phenotypic characteristics not expected for their location in the organism), if at all. One possibly explanation is that in certain circumstances, STAT3 can act in the opposite sense, as a tumour suppressor (De La Iglesia *et al.* , 2008); hence, HNF6 expression in cancer cells could theoretically have a tumourigenic effect by downregulating *GR - STAT3*. However, this might imply a tumour suppressive function for SHP suppression of *HNF6*, in contradiction to the oncogenic function of SHP suppression of *RXR*-α. This alludes to the complexity of possible genetic circuits involved, and also the limitations and dangers of making mechanistic inferences based on a network model such as presented here, where interactions are drawn from a wide range of literature involving many different experimental circumstances.

It must be emphasised that these are only indications of potential gene circuits at work, as there is no supporting evidence from gene expression in this analysis, which would be needed to prove the existence of any gene regulation patterns in these cell lines. Additionally, although all these genes are identified as being in some way part of this glioblastoma stem-like cell differential epigenotype, there is no evidence presented here that the interactions represented take place in the sequence in time which would be necessary to constitute a regulating genetic circuit: the biochemical processes which are represented in figure 3.1 take place over the course of seconds, minutes or hours, whereas the sample frequency of the time-series as observed in these experiments is on the scale of days. Further, no evidence is provided here that the interactions shown in figure 3.1 even take place in the same cells at any time during the experiment; as many of the beta changes are in the region 0.3, we would expect only a sub-population of the cells in any particular cell line and experimental replicate to actually have real methylation changes at these CpGs, as the only possible values of beta for a single cell are 0, 0.5 and 1.

As would be expected in this stem cell context, there is a strong developmental link among the genes identified as part of the glioblastoma stem-like cell differential epigenotype. Although *WT1* is not expressed in healthy mature glial cells, it is expressed in parts of the spinal chord and brain of the developing mammalian embryo (Armstrong *et al.* , 1993). *P73* is intimately involved in differentiation and development (Scharnhorst *et al.* , 2000), and *VIMENTIN* (*VIM*) exhibits a complicated pattern of developmental and tissue specific expression (it is initially widely expressed in the embryo, but progressively restricted to fewer cell types during development), and is aberrantly expressed in most metastatic tumours, whatever their embryological origin (Wu *et al.* , 2004). As well as the identification of *HNF6* in this analysis which would normally be expressed only in the liver, the CpG annotated to *ACTC1* which was identified as part of the glioblastoma stem-like cell differential epigenotype shows the largest beta change (0.54) of any CpG found as part of this glioblastoma stem-like cell differential epigenotype (figure 3.2). The full name of *ACTC1* is 'actin, alpha, cardiac muscle 1', which would normally be expected to be expressed only in the heart, and so is another gene which is presumably only expressed in mature brain cells which are cancerous. In support of this, the methylation level of the relevant CpG increases in GNS2 from a low level at the start of the experiment to a high level similar to NS by the end of the experiment, suggesting expression in the glioblastoma stem-like cells, and repression in the differentiated cells. ACTC1 has also been found to interact with c-MYC, which is one of the most potent activators of tumorigenesis, frequently overexpressed in diverse cancers (Agrawal *et al.* , 2010).

Hox genes are master developmental regulators, and in *drosophila*, polycomb group proteins maintain the repressed state of Hox genes, which act as 'molecular address markers', after they have not been activated at a critical time in development; silencing of certain Hox genes will cause segments further back in a fly to take on characteristics of segments usually found further forward in the animal (Alberts, 2002). It is believed that the Hox genes are similarly important in mammalian development, participating, for example, in the growth and organisation of limbs, where certain Hox genes are sequentially activated as the limb develops (Zakany & Duboule, 2007). The methylation levels of the CpGs annotated to *HOXD4* which are identified as part of the glioblastoma stem-like cell differential epigenotype show methylation levels decreasing in GNS2 towards a low level in NS during the experiment, particularly in the region close to the transcriptional start site (TSS), indicating an increasing pattern of activation of *HOXD4* as the glioblastoma stem-like cells differentiate. HOXD4 inhibits the activity of *CBP*, the protein product of which co-activates the oncogenes *STAT3* and *WT1*, suggesting another possible mechanism for the reduction in tumorigenicity of the glioblastoma stem-like cells as

they differentiate.

The statistical methods used here, and the resulting glioblastoma stem-like cell differential epigenotype (figure 3.1), have highlighted a number of genes already known for their role in glioblastoma and developmental biology. Further evidence has been provided for the importance of certain biochemical interactions which are involved in this glioblastoma stem-like cell differential epigenotype, and a very interesting next stage would be to develop a mathematical model the dynamic behaviour of a number of these and other interactions, during differentiation, to gain further insights into the reduction in tumorigenicity of glioblastoma stem-like cells as they are induced to differentiate.

Chapter 4

Network Inference and Community Detection, Based on Covariance Matrices, Correlations and Test Statistics from Arbitrary Distributions

4.1 Introduction

In this chapter I present methodology which enables estimation of binary adjacency matrices, from a range of measures of the strength of association between pairs of network nodes, or more generally pairs of variables. This strength of association can be quantified in terms of sample covariance / correlation matrices, and more generally by test-statistics / hypothesis test p-values from arbitrary distributions. Binary adjacency matrices inferred in this way are then ideal for community detection, for example by fitting the stochastic blockmodel. I show that this methodology works well in a simulation study, and several gene expression data-sets. This methodology performs well on large datasets, and is based on commonly available and computationally efficient algorithms.

Community detection and clustering are, strictly speaking, different problems. Communities are composed of entities that have some interaction in a real-world sense (such as communication in a social network, and gene-regulation in a biological network). On the other hand, a cluster may simply consist of correlated variables. However, in practice, inference using network models can yield identical approaches to both these problems. The stochastic blockmodel is an effective and efficient method to detect communities in networks, and more generally, to cluster together variables with correlated observations. However, the stochastic blockmodel assumes a binary relationship between the network nodes, and by extension, the variables to be clustered: either there is an edge between a pair of nodes, or there isn't. Such

binary relationships are normally expressed in the form of an adjacency matrix.

If a binary adjacency matrix is used to define pairs of variables which are correlated, and other pairs of variables which are not correlated, then the zero entries may be used to define pairs of variables which are independent. This relates closely to the 'probabilistic graphical model' (Koller & Friedman, 2009) paradigm, in which a joint probability distribution over a large number of variables is made tractable, by taking advantage of independencies between pairs of variables, as specified in the graphical model. These ideas are also closely related to thresholding a covariance matrix to a sparse representation using regularisation techniques (Bickel & Levina, 2008), where again zeros in the sparse representation imply independent pairs of variables. A variety of other methods have also been presented to infer networks from measures of association, such as network inference from multiple node attributes in cell biological data (Katenka *et al.* , 2012).

This chapter is organised as follows. In section 4.2, I define the notation and models, and present the main methodology used throughout the chapter. Then in section 4.3, I present examples to illustrate the performance of the methods, including simulated datasets, and eight gene expression data-sets.

4.2 Model definition

I start by specifying a model by which we can estimate the adjacency matrix $\mathbf{A}$.

Definition 1. *For $m \in \mathbb{N}^+$, define the set of nodes $\{1, ..., m\}$, and for each node i, define a corresponding variable x_i. Let $\hat{z}_{ij}$ represent an observed measure of association/dependence between variables x_i and x_j, where:*

$$\hat{z}_{ij} \sim \mathcal{N}\left(\mu_{ij}, \sigma^2\right).$$

Let $\mathbf{A} \in \{0,1\}^{m \times m}$ be an adjacency matrix, the elements of which satisfy:

$$A_{ij} = \begin{cases} 0, & \textit{if there is no edge between nodes } i \textit{ and } j\textit{, which represents} \\ & \textit{the variables } x_i \textit{ and } x_j \textit{ being independent}, \\ 1, & \textit{if there is an edge between nodes } i \textit{ and } j\textit{, which represents} \\ & \textit{the variables } x_i \textit{ and } x_j \textit{ not being independent}, \end{cases}$$

and let $w = p\left(A_{ij} = 1\right)$. Then, the observed measures of association $\hat{z}_{ij}$ may be modelled

using the mixture distribution:

$$\hat{z}_{ij} \sim (1-w) \cdot \mathcal{N}\left(0, \sigma^2\right) + w \cdot \mathcal{N}\left(\mu_{ij}, \sigma^2\right). \tag{4.1}$$

Next, I describe how to calculate the observed measures of association/dependence $\hat{z}_{ij}$ from sample covariance/correlation matrices, and from test statistics from arbitrary or unknown distributions. After that, I describe how to fit the model of Definition 1, and how to infer the estimated adjacency matrix $\hat{\mathbf{A}}$ from the fitted model. Then, I describe how to carry out community detection on $\hat{\mathbf{A}}$, and at the end of this section I talk about a significant model mis-specification which can arise.

4.2.1 Applying the model to a covariance/correlation matrix

We can estimate an adjacency matrix, from a sample covariance or correlation matrix, by fitting the model of Definition 1, as follows. Equation 4.2 defines the sample covariance matrix $\hat{\mathbf{\Sigma}}$, for the m variables represented by the vector $\mathbf{x}$, $x_1, ..., x_m$, for samples $\mathbf{x}(k)$, $k = 1, ..., n$:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} \left(\mathbf{x}(k) - \bar{\mathbf{x}}\right)\left(\mathbf{x}(k) - \bar{\mathbf{x}}\right)^T, \qquad \text{where} \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}(k). \tag{4.2}$$

By dividing each row and each column of $\hat{\mathbf{\Sigma}}$ by the square roots of the corresponding elements of the leading diagonal, we obtain the sample correlation matrix, $\hat{\boldsymbol{r}}$:

$$\hat{\boldsymbol{r}} = \left(\mathrm{diag}(\hat{\mathbf{\Sigma}})\right)^{-1/2} \hat{\mathbf{\Sigma}} \left(\mathrm{diag}(\hat{\mathbf{\Sigma}})\right)^{-1/2}.$$

The $(i,j)^{\text{th}}$ element of $\hat{\boldsymbol{r}}$, i.e., $\hat{r}_{ij}$, is the Pearson correlation coefficient between variables x_i and x_j. If x_i and x_j are jointly normally distributed, and the $\{x_i(k), x_j(k)\}$, $k = 1, ..., n$ samples are independent, the Fisher transform (Fisher, 1915) converts $\hat{r}_{ij}$ to the normally distributed variable $\hat{z}_{ij}$:

$$\hat{z}_{ij} = \frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right), \tag{4.3}$$

where

$$\hat{z}_{ij} \overset{approx}{\sim} \mathcal{N}\left(\frac{1}{2}\ln\left(\frac{1+r_{ij}}{1-r_{ij}}\right), \frac{1}{\nu-3}\right),$$

where r_{ij} is the true correlation coefficient between variables x_i and x_j, and ν is the degrees of freedom. Hence, we can model the Fisher-transformed correlation coefficients $\hat{z}_{ij}$ with the

mixture model of equation 4.1, with:

$$\mu_{ij} = \frac{1}{2}\ln\left(\frac{1+r_{ij}}{1-r_{ij}}\right) \quad \text{and} \quad \sigma^2 = \frac{1}{\nu-3}. \tag{4.4}$$

4.2.2 Applying the model to test statistics from arbitrary distributions

We can also estimate an adjacency matrix by fitting the model of Definition 1, when the association between variables x_i and x_j is assessed by a test-statistic from an arbitrary distribution, which may be expressed in terms of a hypothesis-test *p*-value. Such *p*-values may result from test-statistics from any known distribution, or may even be derived from an unknown distribution, for example by Monte-Carlo simulation. We can represent these *p*-values in the matrix $\hat{\mathbf{P}}$, where $\hat{p}_{ij}$ is the estimated probability for the pair of variables x_i and x_j, under the null hypothesis H_0 that there is no association between x_i and x_j (i.e., that they are independent). Assuming these *p*-values arose from two-tailed tests, we can apply the inverse-normal transformation as follows:

$$\hat{z}_{ij} = \Phi^{-1}\left(1-\hat{p}_{ij}\right), \tag{4.5}$$

with equivalent expressions available for one-tailed tests. Applying this transformation is equivalent to applying quantile normalisation, mapping the null distribution of p_{ij} onto the standard Normal $\mathcal{N}(0,1)$ distribution. Hence, after applying this transformation, we can again fit the mixture model of Definition 1, and use this model fit to infer the estimated adjacency matrix $\hat{\mathbf{A}}$. I now describe how to carry out this model fitting and inference of $\hat{\mathbf{A}}$.

4.2.3 Model fitting and adjacency matrix inference

I fit the model of Definition 1 using an empirical Bayes procedure (Johnstone & Silverman, 2004). The method is based on a mixture prior over μ_{ij}, with a Laplace density for the non-zero mean component.

Definition 2. *With μ_{ij} and w given by Definition 1, let $\gamma(\cdot)$ represent the Laplace distribution probability density function, with spread parameter a:*

$$\gamma(\mu_{ij}) = \frac{a}{2}\exp\left(-a\left|\mu_{ij}\right|\right).$$

Then, the mixture prior over μ_{ij} is defined as:

$$f_{prior}(\mu_{ij}) = (1-w_i)\,\delta(\mu_{ij}) + w_i\gamma(\mu_{ij}).$$

Typically the Laplace spread parameter is taken as $a = 0.5$. If the mixture components have

Gaussian likelihoods $f_{\mathcal{N}}\left(\cdot\middle|\mu_{ij},\sigma^2\right)$ as in Definition 1, it follows from Definition 2 that the posterior density over the observed measures of association $\hat{z}_{ij}$ is:

$$f_{\text{posterior}}\left(\mu_{ij}|\hat{z}_{ij}\right)=\frac{\left(1-w_i\right)\delta\left(\mu_{ij}\right)f_{\mathcal{N}}\left(\hat{z}_{ij}\middle|0,\sigma^2\right)+w_i\gamma\left(\mu_{ij}\right)f_{\mathcal{N}}\left(\hat{z}_{ij}\middle|\mu_{ij},\sigma^2\right)}{f_{\text{marginal}}\left(\hat{z}_{ij}\right)},$$

where the marginal density is:

$$f_{\text{marginal}}\left(\hat{z}_{ij}\right)=(1-w_i)f_{\mathcal{N}}\left(\hat{z}_{ij}\middle|0,\sigma^2\right)+w_i g\left(\hat{z}_{ij}\right), \tag{4.6}$$

where $g\left(\mu_{ij}\right)$ is the convolution of the Laplace density with the standard normal density. Comparing the expression for $f_{\text{marginal}}\left(\hat{z}_{ij}\right)$ in equation 4.6 with equation 4.1, we see that the normally-distributed non-zero mixture component in equation 4.1, is replaced with the convolution of the Laplace and normal densities in equation 4.6. If a Gaussian prior were used instead of the Laplace prior, then the marginal density in equation 4.6 would be exactly the same as equation 4.1. However, as noted in (Johnstone & Silverman, 2004), this empirical Bayes procedure requires a prior with tails that are exponential or heavier. Hence I use, as they do, the Laplace rather than a Gaussian prior. I note that this is a slight model mis-specification.

This procedure results in a separate model being fitted to each pair of variables (x_i, x_j), based on the corresponding observed statistic $\hat{z}_{ij}$. However, a common weight w_i is used for all models corresponding to each x_i. This estimate of w_i is found as the value which maximises the marginal likelihood (equation 4.7) of the observed statistics $\hat{z}_{ij}$ over all the pairwise comparisons of x_i with x_j, $j \neq i$. This allows the model for each pairwise comparison (x_i, x_j) to 'borrow strength' from all the other comparisons (x_i, x'_j), $j' \neq i$, $j' \neq j$:

$$\hat{w}_i=\arg\max_{w}\sum_{j\neq i}\log\left\{(1-w)\phi\left(\hat{z}_{ij}\right)+w g\left(\hat{z}_{ij}\right)\right\}. \tag{4.7}$$

For a particular x_i, if the $\hat{z}_{ij}$ are mostly close to zero, then w_i will be set low, which means that fewer edges ($A_{ij}=1$) will be detected; this therefore corresponds to i being a low-degree node. If for a different x_i, the $\hat{z}_{ij}$ are generally further from zero, then $\hat{w}_i$ will be set high, which corresponds to more edges being detected; this therefore corresponds to i being a high-degree node. Hence, setting $\hat{w}_i$ separately for each variable x_i allows adaptation to a heterogenous degree distribution in $\mathbf{A}$. The marginal likelihood of equation 4.7 assumes that the $\hat{z}_{ij}$ are independent, however this will not be true in practice. The use of the Laplace prior rather than a Gaussian prior tends to mitigate the effect of this mis-specification.

As in (Johnstone & Silverman, 2004), I use the posterior median to calculate $\hat{\mu}_{ij}$. This

means that:

$$P\left(|\mu_{ij}| > 0|\hat{z}_{ij}\right) > P\left(\mu_{ij} = 0|\hat{z}_{ij}\right) \quad \Longrightarrow \quad |\hat{\mu}_{ij}| > 0,$$
$$P\left(|\mu_{ij}| > 0|\hat{z}_{ij}\right) < P\left(\mu_{ij} = 0|\hat{z}_{ij}\right) \quad \Longrightarrow \quad \hat{\mu}_{ij} = 0.$$

We can then estimate the corresponding adjacency matrix entry A_{ij} as follows:

$$\hat{A}_{ij} = 1 \quad \text{if} \quad |\hat{\mu}_{ij}| > 0,$$
$$\hat{A}_{ij} = 0 \quad \text{otherwise.}$$

However, because I apply the method separately to all comparisons x_i with x_j, $j \neq i$ and to all comparisons x_j with x_i, $i \neq j$, this may not always lead to consistent inference of the form: $\hat{A}_{ij} = \hat{A}_{ji}$. I therefore make a conservative estimate of A_{ij} as follows:

$$\hat{A}_{ij} = 1 \quad \text{if} \quad |\hat{\mu}_{ij}| > 0 \quad \text{and} \quad |\hat{\mu}_{ji}| > 0, \tag{4.8}$$
$$\hat{A}_{ij} = 0 \quad \text{otherwise.}$$

The spread parameter a in the Laplace prior is typically set as $a = 0.5$. However, for additional model flexibility where needed, a can also be estimated by marginal maximum likelihood, in which case I estimate a_i separately for each variable x_i, simultaneously with w_i.

4.2.4 Community detection

Having inferred $\hat{\mathbf{A}}$, community detection (Girvan & Newman, 2002) may then proceed by fitting the degree-corrected stochastic blockmodel (Holland *et al.* , 1983; Bickel & Chen, 2009; Rohe *et al.* , 2011; Qin & Rohe, 2013) directly to $\hat{\mathbf{A}}$. However to fit the degree-corrected stochastic blockmodel, the number of communities in the model must first be specified; this number can be estimated as in (Olhede & Wolfe, 2014). Using this estimate of the number of communities, I infer the set of communities $\hat{C}$ in $\hat{\mathbf{A}}$, such that a community $\hat{c} \in \hat{C}$ is a group of variables x_i, $i \in \hat{c}$. Such a community $\hat{c}$ would correspond to an unexpectedly large number of non-zero entries $\hat{\Sigma}_{ij}$, of the sample covariance matrix $\hat{\mathbf{\Sigma}}$, for pairs of variables x_i and x_j, where $i \in \hat{c}$ and $j \in \hat{c}$. Alternatively, the community $\hat{c}$ would correspond to an unexpectedly large number of significant *p*-values $\hat{p}_{ij}$, in the matrix $\hat{\mathbf{P}}$, for pairs of variables x_i and x_j again with $i \in \hat{c}$ and $j \in \hat{c}$.

4.2.5 Model mis-specification

A practical point of note, is that sometimes directional information may not be available in the measures of association/dependence between x_i and x_j. Such directional information determines whether the correlation/covariance is positive or negative, such as the sign on the sample Pearson correlation coefficient. This scenario might arise if the sign of the correlation or test-statistic has been discarded at an earlier stage in the data-processing which cannot be repeated, or if the measure of association is a *p*-value resulting from a two-tailed test. This would result in all calculated $\hat{z}_{ij} \geq 0$, which causes a model mis-specification, because the small values of $\hat{z}_{ij}$ which correspond to H_0, and which hence originate from $\mu_{ij} = 0$, will all be positive. Hence, under these circumstances, the small $\hat{z}_{ij}$ will originate from a half-normal rather than a normal distribution.

We can informally explore under what circumstances this model mis-specification will have a significant effect. The mean of a half-normal distribution is $\sigma\sqrt{2/\pi}$, where σ^2 is the variance of the corresponding normal distribution. Therefore, if the $\hat{z}_{ij}$ are derived from covariance/correlation matrices (as described in Section 4.2.1), the mean of the $\hat{z}_{ij}$ which originate from the zero-mean component (i.e., correspond to $\mu_{ij} = 0$) under this model mis-specification will be $\sqrt{2/\left(\pi \times (\nu - 3)\right)}$, where ν is the degrees of freedom. Hence, when $\hat{z}_{ij}$ is calculated from $n = 50$, $n = 100$ and $n = 200$ samples, the corresponding half-normal distribution means are 0.12, 0.081 and 0.057, respectively. The standard deviation of a half-normal is approximately $0.6\times$ that of a full normal distribution, and hence the corresponding standard deviations are 0.088, 0.061, and 0.043, respectively. Maclaurin expanding the Fisher transformed expression for $\hat{z}_{ij}$ in terms of $\hat{r}_{ij}$ (equation 4.3) up to first, third and fifth orders, gives $\hat{r}_{ij}$, $\hat{r}_{ij} + (\hat{r}_{ij})^3/3$ and $\hat{r}_{ij} + (\hat{r}_{ij})^5/5$ respectively. Hence, we can take $\hat{z}_{ij} \approx \hat{r}_{ij}$ for this assessment. We can reasonably hypothesise that the region in which we expect this model mis-specification will become problematic, is when the observed $\hat{z}_{ij}$ fall in the region between the mean, and the mean plus one standard deviation, of these half-normal distributions. This is because in this region, much of the distribution of the mixture component which arises from $\mu_{ij} \neq 0$ will overlap with much of the distribution of the component arising from $\mu_{ij} = 0$. Such regions correspond to $0.12 < r_{ij} < 0.21$ if $\nu = 50$, $0.081 < r_{ij} < 0.14$ if $\nu = 100$, and $0.057 < r_{ij} < 0.1$ when $\nu = 200$. This point will be examined further in the context of the simulation study, in the next section.

4.3 Examples

I now present results of applying the above methodology to simulated data, and to several gene-expression data-sets. I carry out network inference as described, resulting in the binary adjacency matrix, to which I fit the degree-corrected stochastic blockmodel by regularised spectral clustering (Holland *et al.* , 1983; Bickel & Chen, 2009; Rohe *et al.* , 2011; Qin & Rohe, 2013). I note that spectral clustering could be expected to be computationally intensive, as it requires a singular value decomposition (SVD) of a large matrix. However, efficient computational methods exist to find the top few components in the singular value decomposition of large sparse matrices (Sørensen, 1992; Lehoucq & Sørensen, 1996). Binary adjacency matrices such as those considered here tend to be very sparse, and we only require as many components as the number of communities or clusters we are trying to find, a number which tends to be two or more orders of magnitude smaller than the dimension of the adjacency matrix, m. Hence, these efficient computational methods are applicable here. Implementations of these efficient computational methods are included in *Matlab* and *R*, meaning that this methodology is practical for large data-sets, and are accessible to a wide range of users.

4.3.1 Simulation study

I carried out a simulation study, to evaluate the effectiveness of this network inference methodology against generated networks with known ground-truth community structure. A generative model for exchangeable random networks with heterogenous degrees is the logistic-linear model (Perry & Wolfe, 2012). I use a version of that model here with community structure added. This additional community structure takes the form of 'blocks'. This block structure is very general: as noted in (Olhede & Wolfe, 2014) it can be used as a model for community structure in relation to many real data-sets for which the true generative mechanism of the community structure is not exactly such block structure. The generative model for this simulation study is defined as:

$$\text{Logit}\left(p_{ij}\right) = \alpha_i + \alpha_j + \theta_{ij}$$

where p_{ij} defines the probability of an edge being observed between nodes i and j. I choose to use this model, because the parameters can take any real values, and the edge probabilities p_{ij} will still be between 0 and 1. This model only deviates from the equivalent log model when the parameter values become very large, which is what prevents p_{ij} from reaching (and exceeding) 1. The node-specific parameters α_i, $i \in 1, ..., m$ are elements of the parameter vector $\boldsymbol{\alpha}$ which defines a power-law degree-distribution for the nodes. Each α_i is generated as the logarithm of a sample taken from a bounded Pareto distribution as in (Olhede & Wolfe, 2012). I note that

because the α_i are chosen to be random, the generated networks are exchangeable (Kallenberg, 2005), whereas if the elements of $\boldsymbol{\alpha}$ were defined deterministically, these networks would instead be generated under the inhomogenous random graph model (Bollobás *et al.* , 2007). The community parameter θ_{ij} is allowed to take two values: $\theta_{ij} = \theta_{\text{in}}$ if i and j are in the same community, and $\theta_{ij} = \theta_{\text{out}}$ otherwise. I choose to do this because it is a simple way of adding community structure, and it is equivalent to a modelling constraint which improves parameter identifiability in some formulations of the stochastic blockmodel (Newman, 2013). After generating the p_{ij}, the network is generated by sampling each A_{ij} according to:

$$A_{ij} \sim \text{Bernouilli}\left(p_{ij}\right).$$

The communities themselves are planted in the network as randomly chosen groups of 150 nodes. I set the number of communities $k = 20$, and hence the generated networks each comprise $m = 3000$ nodes.

Having generated a network with known ground-truth community structure in this way, I use it to randomly generate a sample correlation matrix $\hat{\mathbf{r}}$, from which I attempt to reproduce the known community structure. To do this, I first generate a random sample covariance matrix $\hat{\mathbf{S}}_{ij}$ for each pair of nodes i and j, according to:

$$\hat{\mathbf{S}}_{ij} \sim \text{Wishart}\left(\mathbf{S}, \nu\right)$$

where

$$\mathbf{S} = \begin{pmatrix} 1 & r_{\text{gen}} \\ r_{\text{gen}} & 1 \end{pmatrix}$$

if $A_{ij} = 1$, where r_{gen} is the model generative correlation coefficient, and

$$\mathbf{S} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

if $A_{ij} = 0$, and ν is the degrees of freedom. I then calculate the estimate of the sample Pearson correlation coefficient $\hat{r}_{ij}$ for nodes i and j as $\hat{r}_{ij} = \left(\hat{\mathbf{S}}_{ij}\right)_{12} / \sqrt{\left(\hat{\mathbf{S}}_{ij}\right)_{11} \times \left(\hat{\mathbf{S}}_{ij}\right)_{22}} = \left(\hat{\mathbf{S}}_{ij}\right)_{21} / \sqrt{\left(\hat{\mathbf{S}}_{ij}\right)_{11} \times \left(\hat{\mathbf{S}}_{ij}\right)_{22}}$. With all elements of $\hat{\mathbf{r}}$ generated in this way, with $\hat{r}_{ij} = \hat{r}_{ji}$ and $\hat{r}_{ii} = 0$ for $i, j \in \{1, ..., m\}$, I proceed with network inference and community detection according to the methods presented above.

I test the methods on networks generated with values of $\theta_{\text{in}} \in \{50, 30, 20, 10\}$, which cor-

responds to within-community edge density $\rho_{\text{in}} \in \{0.81, 0.34, 0.15, 0.039\}$. For all networks, I set $\theta_{\text{out}} = 1$, corresponding to between-community edge density $\rho_{\text{out}} = 0.0013$. I generate sample covariance matrices with $r_{\text{gen}} \in [0, 0.8]$, and degrees of freedom $\nu \in \{50, 100, 200\}$. For each combination of parameters, I carry out 50 repetitions of network generation followed by network inference and community detection, to enable assessment of the variability of the accuracy of the network inference. To compare detected communities in the inferred network with the ground-truth planted communities, I use the normalised mutual information (NMI) (Danon *et al.* , 2005). The NMI assesses the numbers of nodes which appear together in the detected communities, compared with whether they appeared together in the planted communities (adjusted for group sizes). The NMI takes the value 1 if the communities are perfectly reproduced in the community detection, and 0 if they are not reproduced at all, and somewhere in between if they are partially reproduced.

The results of the simulation study are shown in Figure 5.2. The accuracy of reproduction of the ground-truth community structure is high, as long as the generative correlation coefficient r_{gen} is high enough. Below this threshold the performance quickly deteriorates, as the method described in Section 4.2.3 no longer detects any edges. This is because the non-zero mean component of the generative mixture model becomes centred too close to zero, and so the $\hat{z}_{ij}$ from this component become categorised together with those from the zero-mean mixture component, with the model fitting effectively assigning all $\hat{z}_{ij}$ to the zero-mean component. However, as long as the generative correlation coefficient r_{gen} is high enough, the method performs well even with fairly sparse within-community edge density in the ground-truth planted communities. Typically, the method fails when r_{gen} falls below 0.45, 0.35 and 0.25 for $\nu = 50$, $\nu = 100$ and $\nu = 200$, respectively. I also note that the performance actually decreases for high ρ_{in}, as r_{gen} becomes high. This can be explained by a combination of the limitation of the asymptotic normality of the Fisher transformation, and a limitation of the performance of the edge inference method described in Section 4.2.3 when there are many extreme values of $\hat{z}_{ij}$.

I then repeated the simulation study, with one change to the generative model, to examine the inferential performance under the model mis-specification described in Section 4.2.5. This change is to replace the generated $\hat{r}_{ij}$ with $|\hat{r}_{ij}|$, thus discarding the directionality information. With this change, we now expect the calculated $\hat{z}_{ij}$ which arise from $A_{ij} = 0$ to follow the half-normal distribution described in Section 4.2.5. The results of this model mis-specification simulation study are shown in Figure 4.2. When the within community edge-density, ρ_{in}, is highest, interestingly the performance actually improves for lower values of the generative correlation coefficient r_{gen}, and is then only limited by the model mis-specification. Typically, the

method now fails when r_{gen} falls below about 0.2, 0.15 and 0.1 for $\nu = 200$, $\nu = 100$ and $\nu = 50$, respectively. The vertical brown lines show the mean and the mean plus one standard deviation, of the half-normal distribution corresponding to the full-normal distribution of the zero-mean component specified by the mixture model; the mean plus one standard deviation line similarly corresponds to about 0.2, 0.15 and 0.1 for $\nu = 200$, $\nu = 100$ and $\nu = 50$, respectively. When ρ_{in} is highest, the NMI goes from approximately 1 to approximately 0 in this region. However, as ρ_{in} decreases, the performance quickly becomes poor, for all values of r_{gen}. The reason that (as long as ρ_{in} is high enough) performance is better in the case of the mis-specified model, is that now as r_{gen} decreases, the non-zero mean mixture component overlaps significantly with the half-normal component (which arises from the zero-mean component), and the resulting density is not centred on zero. Hence, the model fit categorises most $\hat{z}_{ij}$ as being from a non-zero-mean component, rather than being from the zero-mean component. Although a reasonable number of edges are almost always detected when the model is mis-specified, if the generative correlation coefficient r_{gen} leads to a distribution which has much overlap with the distribution of the half normal arising from the zero-mean mixture component, many of the edges which are detected are false positives.

4.3.2 Comparison with popular clustering methods

The clustering problem is fundamentally different to that of community detection, although there are nevertheless many similarities. The basic task of clustering is to group together entities (usually variables or samples) which share some attributes, which may lead to more highly correlated behaviour within the groups than between groups. When the entities being grouped are nodes in a network, the problems of clustering and community detection become essentially the same problem. In this study, I infer binary networks from continuous data before carrying out community detection. However, a number of popular methods provide alternative means of clustering entities into groups (which may be considered equivalent to communities), based on continuous data.

A method of clustering which is very popular across the biological and social sciences, is hierarchical clustering. In that method, variables or samples are grouped together according to their 'distance' from one another. A popular measure of distance between a pair of such variables or samples is simply $1 - r$, where r is the absolute value of the Pearson correlation coefficient between the pair. Hence, this method can be easily applied to data of the type presented here (without carrying out the network inference presented in Section 4.2.3). I tested this method on the simulated data presented in Section 4.3.1, by applying hierarchical clustering to the generated sample correlation matrix $\hat{\mathbf{r}}$ before comparing the detected clusters with the

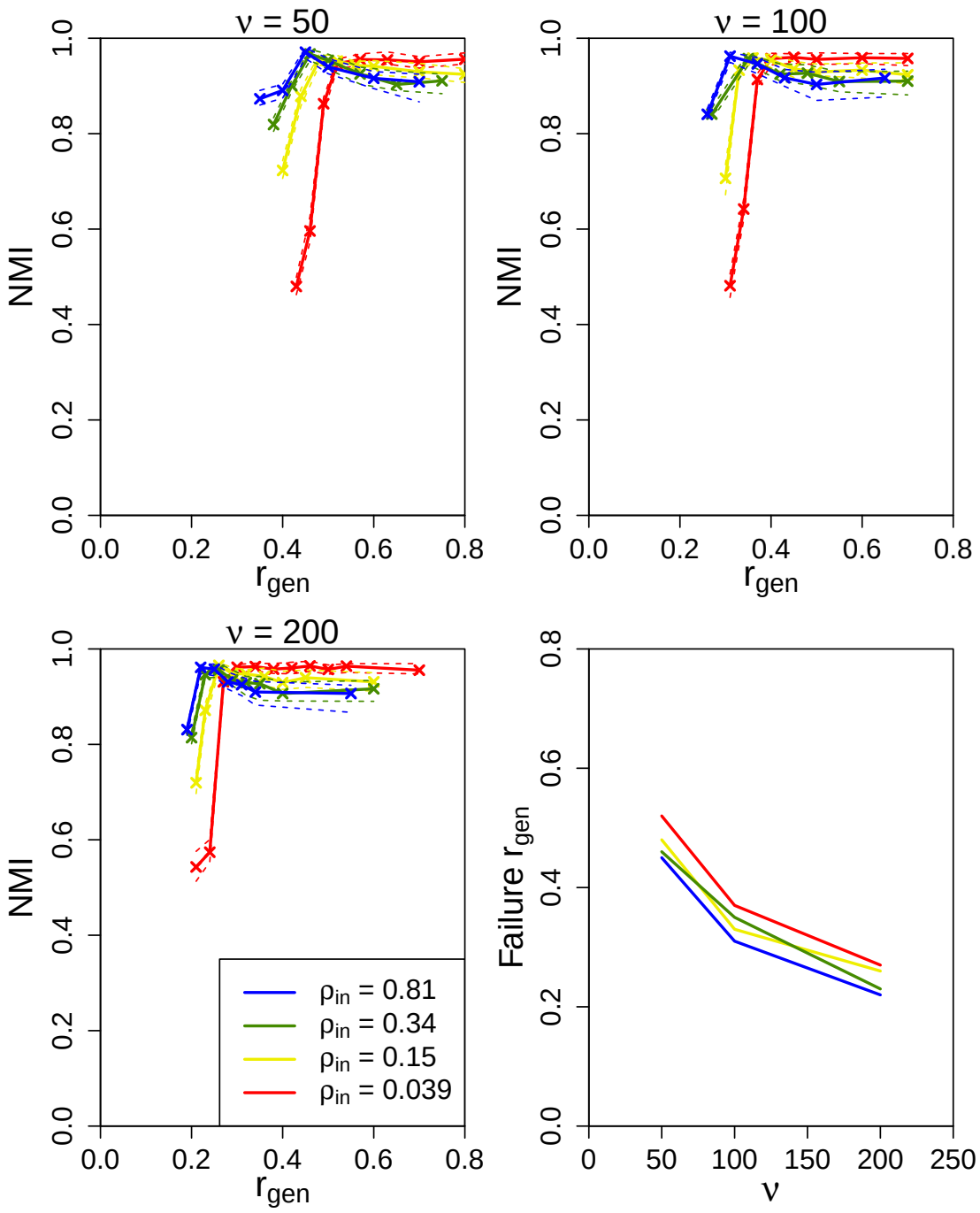


Figure 4.1: *Simulation study.*
Normalised mutual information (NMI) compares detected community structure with ground-truth planted communities. Each line corresponds to a different within-community edge-density; these are set as $\rho_{in} \in \{0.81, 0.34, 0.15, 0.039\}$ *by setting* $\theta_{in} \in \{50, 30, 20, 10\}$. *The degrees of freedom,* ν, *are set as* $\nu \in \{200, 100, 50\}$. *For each network, the number of nodes* $m = 3000$, *the ground-truth number of communities is* $k = 20$, *and the between-community edge density is set as* $\rho_{out} = 0.0013$ *by setting* $\theta_{out} = 1$. *Dashed lines indicated quartiles.*

planted communities. However, I found that in every case, the result of this comparison was a value of the NMI close to 0. Therefore, we may conclude that hierarchical clustering performs significantly worse than the methods presented here, on problems of this type.

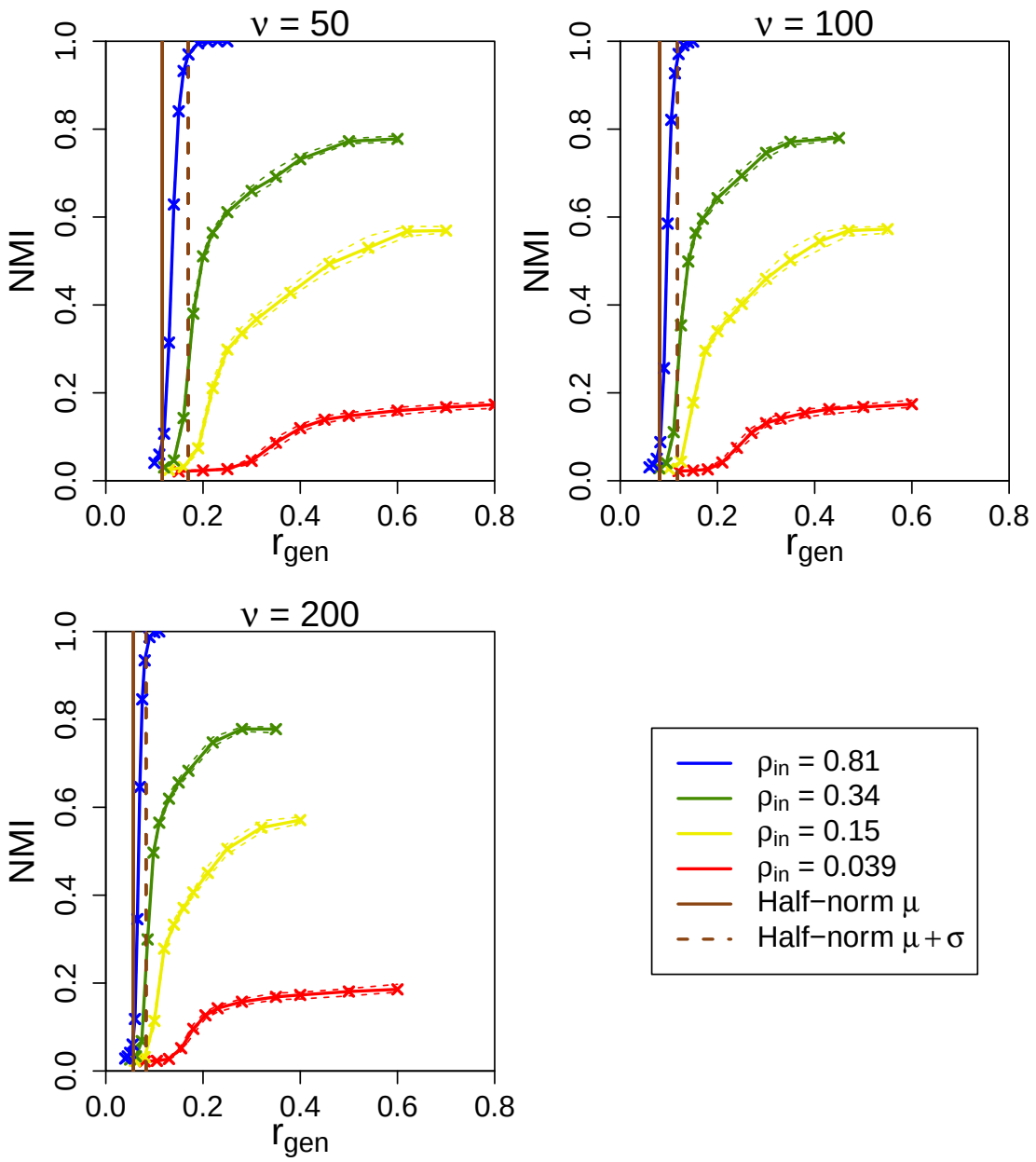


Figure 4.2: *Simulation study, with model mis-specification.*
Normalised mutual information (NMI) compares detected community structure with ground-truth planted communities. Each line corresponds to a different within-community edge-density; these are set as $\rho_{in} \in \{0.81, 0.34, 0.15, 0.039\}$ *by setting* $\theta_{in} \in \{50, 30, 20, 10\}$. *The degrees of freedom,* ν, *are set as* $\nu \in \{200, 100, 50\}$. *For each network, the number of nodes* $m = 3000$, *the ground-truth number of communities is* $k = 20$, *and the between-community edge density is set as* $\rho_{out} = 0.0013$ *by setting* $\theta_{out} = 1$. *Dashed coloured lines indicated quartiles. Vertical brown lines show the theoretical mean and mean + 1 standard deviation for the half-normal distribution corresponding to the full-normal distribution of the zero-mean component in the mixture model. Hence, they are an illustration of the effect of the model mis-specification discussed.*

One of the most popular clustering methods is K-means, in which samples (which may be thought of as equivalent to network nodes) are grouped into K clusters based on their location in N-dimensional space. On its own, this method is fundamentally ill-suited to network data,

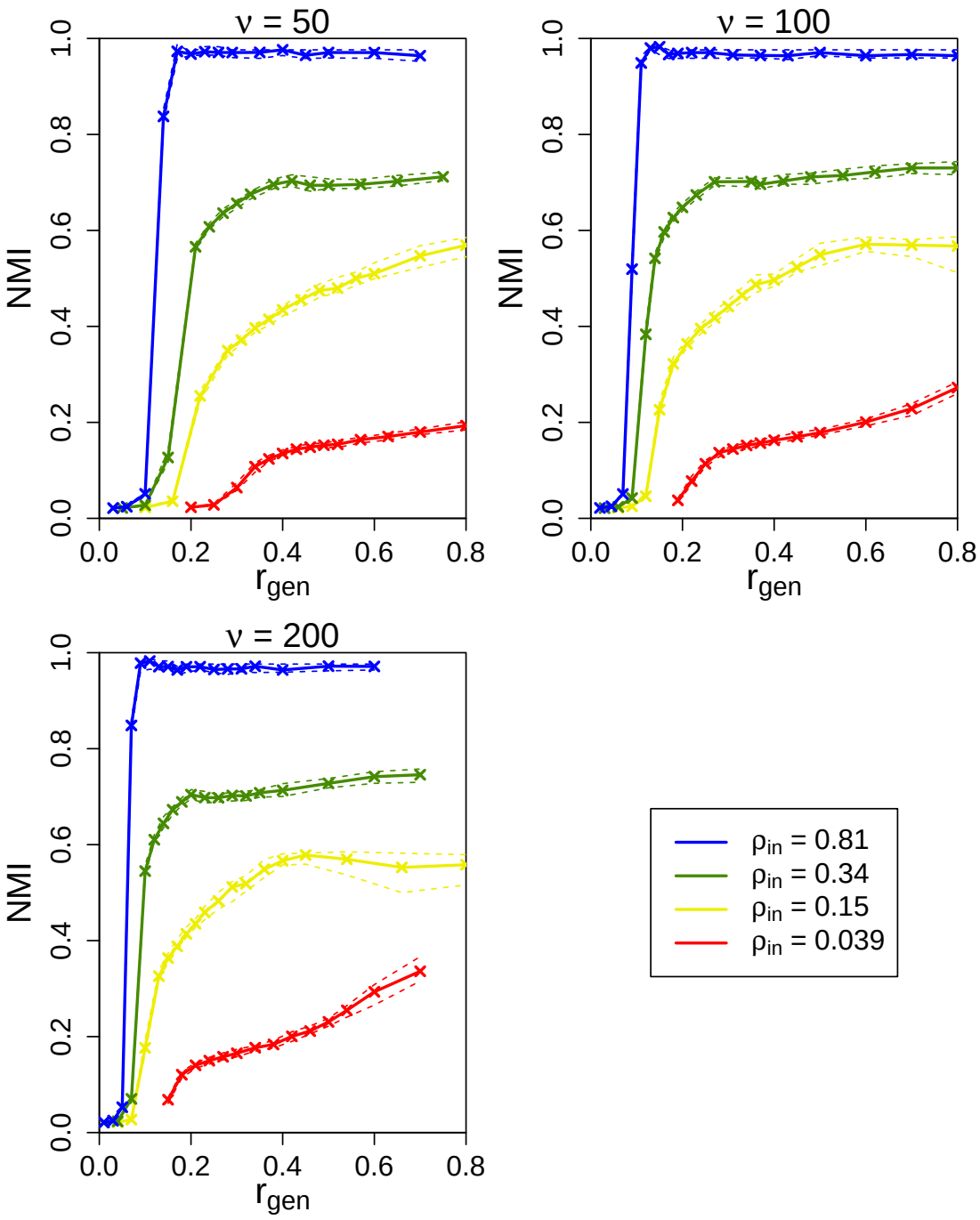


Figure 4.3: *Simulation study: spectral clustering without network thesholding. Normalised mutual information (NMI) compares detected community structure with ground-truth planted communities. Each line corresponds to a different within-community edge-density; these are set as* $\rho_{in} \in \{0.81, 0.34, 0.15, 0.039\}$ *by setting* $\theta_{in} \in \{50, 30, 20, 10\}$*. The degrees of freedom,* ν*, are set as* $\nu \in \{200, 100, 50\}$*. For each network, the number of nodes* $m = 3000$*, the ground-truth number of communities is* $k = 20$*, and the between-community edge density is set as* $\rho_{out} = 0.0013$ *by setting* $\theta_{out} = 1$*. Dashed lines indicated quartiles.*

because of the high dimensionality of the problem. However, K-means clustering is often used as the final stage in spectral clustering, which is the most common way of fitting the stochastic blockmodel - and it is used by us here, for that purpose. Spectral clustering can also be used

to cluster continuous data, and so for comparison, I have applied regular spectral clustering (without carrying out the network inference presented in Section 4.2.3) to the simulated data presented in Section 4.3.1. To do this, I applied spectral clustering as described at the start of Section 4.3 directly to the absolute of the generated sample correlation matrix $\hat{\mathbf{r}}$ (i.e., continuous data). The absolute values are used to ensure that the data is non-negative, as required for spectral clustering (Von Luxburg, 2007). The results appear in Figure 4.3.

The result of applying spectral clustering applied directly to $\hat{\mathbf{r}}$ is somewhat successful (Figure 4.3). However, it does not perform as well as when the network inference/thresholding of Section 4.2.3 is first applied. The most natural comparison with those results, relates to when the model mis-specification described in Section 4.2.5 occurs (Figure 4.2), as in both cases the absolute of the sample correlation matrix $\hat{\mathbf{r}}$ is considered. When spectral clustering with and without the network thresholding of Section 4.2.3 both perform comparably well, it is when the expected within-group correlation is highest ($\rho_{\text{in}} = 0.81$): i.e., when the presence of an edge (i, j) in the generative network will likely lead to an entry $\hat{r}_{ij}$ close to 1. However when ρ_{in} is set lower, the performance becomes relatively worse for spectral clustering on continuous data without network thresholding. Hence, we may conclude that the network inference presented in Section 4.2.3 provides an improvement in performance over and above regular spectral clustering, on problems of the type presented here. Further, due to the restriction that the data must be non-negative, when applying spectral clustering without such network thresholding/inference, model mis-specification similar to that described in Section 4.2.5 becomes unavoidable.

4.3.3 Gene-expression example

I present an example of a practical application of these methods to a standard problem in gene-expression analysis. Community detection can be used to infer groups of genes which comprise functional subnetwork modules, or groups of co-regulated genes. Examples of such groups are found in gene regulatory networks and protein signalling networks (Shen-Orr *et al.* , 2002). Defining $\mathbf{x}(k)$ to be gene expression measurements for sample k for the genes $x_1, x_2, ..., x_m$, I calculate the covariance matrix according to equation 4.2, and carry out network inference as described. I note that the network edges detected in this way may be transitive edges, i.e., they do not necessarily represent physical interactions between genes and gene products. To determine this would require additional functional data, such as that relating to DNA binding by gene products (e.g., transcription factors) (Jojic *et al.* , 2013). However, in general, the groups of genes detected in this way can be expected to form biologically meaningful subnetwork modules, generating biological hypotheses which may warrant further investigation by experimental scientists.

I carried out this process of network inference and community detection in gene expression data from 8 different types of cancer: brain, breast, colon, kidney, lung, ovarian, rectal and uterine (data source: The Cancer Genome Atlas (Hampton, 2006)). Each data set comprises gene expression measurements for 17505 genes (i.e., $m = 17505$). Figure 4.4 shows the inferred adjacency matrix, after community detection, for the lung cancer data-set. The number of communities is estimated as 105 by the network histogram method (Olhede & Wolfe, 2014) for this data-set, and the edge density is $\rho = 0.062$ (which is typical of all 8 gene expression datasets).

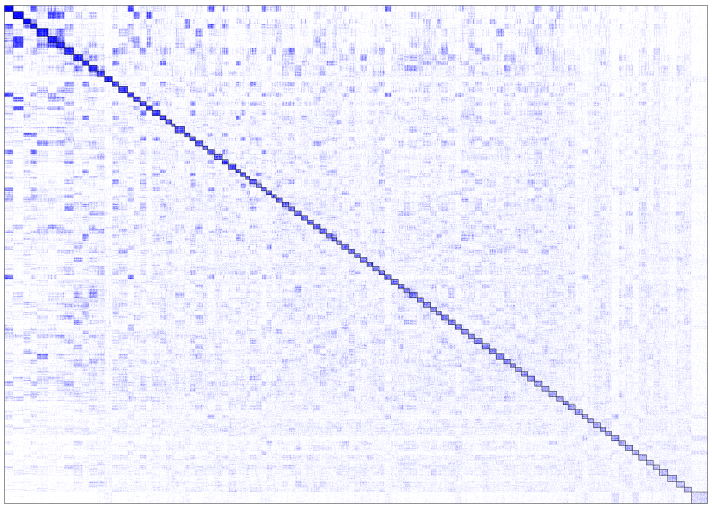


Figure 4.4: *Detected communities in a lung cancer gene expression data set.*
Entries in the adjacency matrix equal to 1 (representing a network edge) are coloured blue, and detected communities are outlined in black.

I also tested the domain-relevance of the communities detected in the inferred networks. I tested the overlap of the genes of each detected community, separately with each of 10295 known gene-groups (data source: *http://www.broadinstitute.org/gsea/msigdb/*). This is known as 'gene set enrichment analysis' (GSEA) (Subramanian *et al.* , 2005). Table 4.1 shows the percentage of the communities detected in each cancer data-set, which overlapped significantly (Fisher's exact test, FDR-adjusted $p < 0.05$) with at least one of these known gene-groups. As a benchmark, I also sampled random groups of genes, from the 17505 genes represented

Breast	Colon	Brain	Kidney	Lung	Ovarian	Renal	Uterine
97%	86%	87%	76%	89%	96%	76%	66%

Table 4.1: *Domain-relevance of detected communities.*
The table shows the percentage of the communities, detected in each cancer data-set, which overlap significantly (Fisher's exact test, FDR-adjusted $p < 0.05$) with at least one known gene group.

in the cancer data-sets, and tested them for overlap with the same 10295 known gene-groups. The number of genes in each random sample was itself randomly sampled from the distribution of the sizes of the communities detected in the cancer data-sets. I took 1000 randomly sampled groups of genes like this, of which 2% overlapped significantly (Fisher's exact test, FDR-adjusted $p < 0.05$) with at least one of the known gene-groups. These results show a

high level of domain-relevance of the detected communities, in all 8 gene-expression data-sets analysed here.

4.4 Discussion

In this chapter, I have presented a method which combines estimation of adjacency matrices with community detection via stochastic blockmodel, based on sample covariance and correlation matrices, and more generally matrices of arbitrary test statistics between pairs of variables. I have described the theory behind this method, and provided practical details for its implementation. I have shown examples of successful applications of this methodology to a simulation study, and to multiple gene-expression datasets. I have also shown that this methodology performs better than popular clustering methods, for discovering latent groupings in data of the type presented here. An important point to note, is that some network edges inferred from the correlation structure of data as in the methodology proposed here, may be what are often referred to as 'transitive edges'. I.e., an inferred edge may not correspond to a direct physical real-life interaction, instead deriving from some indirect interaction which may alternatively be mediated via a less direct route through the network, possibly also involving unobserved variables. An interesting extension to this methodology would be to consider overlapping blocks in the stochastic blockmodel (Latouche *et al.* , 2011). Another interesting extension would be to develop an online version of the method, as a computationally efficient approach to large and growing data-sets (Zanghi *et al.* , 2010). This methodology would be expected to work equally well in many other networks contexts. It could also be expected to work well in more general contexts where the aim is to cluster together correlated variables. The number of communities or clusters can be estimated automatically using the network histogram method (Olhede & Wolfe, 2014), allowing fully automated processing. This methodology is based on commonly available and computationally efficient methods, and performs well on large datasets.

Chapter 5

Co-modularity and Co-community Detection in Large Networks

5.1 Introduction

This chapter introduces the notion of co-modularity, to co-cluster observations of bipartite networks into co-communities. The task of co-clustering is to group together nodes of one type, whose interaction with nodes of another type are the most similar. The novel measure of co-modularity is introduced to assess the strength of co-communities, as well as to arrange the representation of nodes and clusters for visualisation. The existing non-parametric understanding of co-clustering is generalised in this chapter, by introducing an anisotropic graphon class for realisations of bipartite networks. By modelling the smoothness of the anisotropic graphon directly, it is possible to obtain a quantitative measure to determine the number of groups to be used when fitting co-communities, subsequently using the co-modularity measure to do so. I illustrate the power of the proposed methodology on simulated data, as well as an example based on linked DNA methylation and gene-expression data.

Studying relationships between variables of the same type is naturally of great utility; its simplest generalisation is to study relationships between variables of a different type; this is known as the co-clustering problem (Flynn & Perry, 2012; Choi *et al.* , 2014; Madeira & Oliveira, 2004). This problem can also be approached non-parametrically, as is made clear in (Choi *et al.* , 2014). To achieve consistent estimation, assumptions have to be made regarding the properties of the graphon function, where smoothness is standard (Olhede & Wolfe, 2014) and stronger assumptions (Airoldi *et al.* , 2013) yield better estimation procedures when the stronger assumptions are justified. Assumptions made for the symmetric graphon function in the clustering problem need extension to the asymmetric graphon lying behind the biclustering problem (Aldous, 1985). To enable understanding of nonparametric estimation, I introduce the model of an anistropic graphon, called the anisotropic graphon model.

Having provided a model for the set of relationships between two types of variables, we now need to infer them. I shall start from the modularity approach to recognising communities (Girvan & Newman, 2002), realising that extending such understanding to variables of different types is nontrivial (Aldous, 1985; Madeira & Oliveira, 2004). Having recognised communities in both types of variables, we need to transform the clustering or grouping of both types of variables, into an ordering of groups. This is not inherent to the formulation of the Aldous-Hoover representation of the generating mechanism of the random array we are modelling, but is important for visualisation purposes. I also use the modularity to make this choice of visualisation.

To be able to use the modularity, we need to decide how many groups we are using in both variable types. This will be based on the model of the anistropic graphon, and a choice of smoothness for the graphon function. I extend the work of (Olhede & Wolfe, 2014) to select the number of groups, adjusted for the anisotropic graphon model. All parameter choices are determined from the data, and a fully specified method of group allocation is given.

Finally, to demonstrate the power of the newly proposed method, I carry out a simulation study, and I analyse a relevant network data set which is based on linked DNA methylation and gene expression data. These analyses show the power of the proposed analysis methods, and enable us to discover both known and hitherto unknown characteristics of such data sets.

This chapter is organised as follows: Section 5.2 defines the stochastic block model, and gives the representation of an arbitrary separately exchangeable array. It also defines the co-modularity, and explains how the array data will be analysed. Section 5.3 describes how to choose the number of co-communities, and section 5.4 shows how to determine them from data. Section 5.5 gives examples to illustrate the performance of the proposed method, and the derivations section provides all proofs of the chapter.

5.2 Co-modularity and co-community detection

I begin this section by defining the degree-corrected stochastic co-blockmodel (Rohe & Yu, 2012; Flynn & Perry, 2012; Choi *et al.* , 2014) together with notation; I then define a generalisation of this model based on the notion of the graphon. Following these model definitions, I give a definition of the Newman-Girvan modularity, and by analogy, I define a quantity which I term the 'co-modularity', and I specify an algorithm for maximising this quantity. I then show that under certain conditions, maximising the co-modularity in this way is equivalent to maximising the model likelihood of the specified degree corrected stochastic co-blockmodel.

Definition 3 (Degree-corrected stochastic co-blockmodel)**.** *For $m, l \in \mathbb{N}^+$, define the set of X-nodes $\{1, ..., m\}$, and the set of Y-nodes $\{1, ..., l\}$. Denote an X-node grouping as $g_p^{(X)} \in G^{(X)}$, $p \in \left\{1, ..., k^{(X)}\right\}$, and a Y-node grouping as $g_q^{(Y)} \in G^{(Y)}$, $q \in \left\{1, ..., k^{(Y)}\right\}$, where $G^{(X)}$ and $G^{(Y)}$ are exhaustive lists of mutually exclusive X and Y-node groupings, respectively. Define map functions $z^{(X)}(i)$ and $z^{(Y)}(j)$, such that $g_p^{(X)} = \left\{i : z^{(X)}(i) = p\right\}$, and $g_q^{(Y)} = \left\{j : z^{(Y)}(j) = q\right\}$. Define co-community connectivity parameters $\boldsymbol{\theta} \in [0,1]^{k^{(X)} \times k^{(Y)}}$, where $\theta_{z^{(X)}(i), z^{(Y)}(j)}$ is the propensity of X-node i in group $z^{(X)}(i)$ to form a connection with Y-node j in group $z^{(Y)}(j)$. Define also node-specific connectivity parameters $\boldsymbol{\pi}^{(X)} \in \mathbb{R}_{\geq 0}^m$ and $\boldsymbol{\pi}^{(Y)} \in \mathbb{R}_{\geq 0}^l$. Let the elements of the adjacency matrix $\mathbf{A} \in \{0,1\}^{m \times l}$ follow the law of:*

$$A_{ij} \sim \text{Bernouilli}\left(\pi_i^{(X)} \pi_j^{(Y)} \theta_{z^{(X)}(i), z^{(Y)}(j)}\right), 1 \leq i \leq m, 1 \leq j \leq l. \tag{5.1}$$

Then, A_{ij} is generated under the degree corrected stochastic blockmodel.

I note that the terminology 'X-nodes' and 'Y-nodes' is non-standard; I introduce it here, to increase clarity. To improve identifiability of parameters of the model in Definition 5.2, I introduce a specification favoured by many other authors (Newman, 2013), that $\theta_{z^{(X)}(i), z^{(X)}(j)}$ may take only two values:

$$\theta_{p,q} = \begin{cases} \theta_{\text{in}}, & \text{if the pairing of } X\text{-node grouping } g_p^{(X)} \text{ with } Y\text{-node} \\ & \text{grouping } g_q^{(Y)} \text{ is a co-community,} \\ \theta_{\text{out}}, & \text{otherwise.} \end{cases} \tag{5.2}$$

We can also replace the Bernoulli model likelihood with a Poisson likelihood: because the Bernoulli success probability is typically small, and the number of potential edges (i.e., pairings of nodes) is large, a Poisson distribution with the same mean behaves very similarly, and so it makes little difference in practice (Zhao *et al.* , 2012; Perry & Wolfe, 2012). Its usage greatly simplifies the technical derivations. Hence, I calculate the model log-likelihood as follows (assuming $A_{ij} \in \{0,1\}$ and therefore $A_{ij}! = 1$ for all i, j):

$$\begin{aligned} &\ell\left(\boldsymbol{\theta}, \boldsymbol{\pi}^{(X)}, \boldsymbol{\pi}^{(Y)}; G^{(X)}, G^{(Y)}\right) \\ &\qquad = \sum_{i=1}^{m} \sum_{j=1}^{l} A_{ij} \ln\left(\pi_i^{(X)} \pi_j^{(Y)} \theta_{z^{(X)}(i), z^{(Y)}(j)}\right) - \pi_i^{(X)} \pi_j^{(Y)} \theta_{z^{(X)}(i), z^{(Y)}(j)}. \end{aligned} \tag{5.3}$$

If we want to let the network grow, it would be impractical to fully specify more complicated versions of the parametric model of Definition 3, which completely account for all effects. In-

stead, we can make a non-parametric generalisation of this model incorporating more smoothing, based on the notion of the graphon. The graphon is a latent, smooth function which sets the probability between each pair of nodes, of a connection forming between that pair of nodes (Wolfe & Olhede, 2013). In this setting, the graphon is not symmetric, due to the two different types of nodes modelled.

Definition 4. *For the Lipschitz-continuous graphon $f \in L\left((0,1)^2\right)$, with $\mathbf{A}$ defined according to Definition 3, define connectivity functions $\phi^{(X)} \in L(0,1)$ and $\phi^{(Y)} \in L(0,1)$, and define latent orderings $\xi_i^{(X)} \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$ and (independently) $\xi_j^{(Y)} \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$ on the graphon margins of X and Y-nodes $i \in \{1,...,m\}$ and $j \in \{1,...,l\}$ respectively. Then,*

$$\mathbb{E}(A_{ij}) = f\left(\xi_i^{(X)}, \xi_j^{(Y)}\right) \cdot \phi^{(X)}\left(\xi_i^{(X)}\right) \cdot \phi^{(Y)}\left(\xi_j^{(Y)}\right). \tag{5.4}$$

The graphon f (Definition 4) can be considered an infinite-dimensional equivalent to $\theta_{p,q}$ (Definition 3), up to a re-ordering of the nodes defined by the orderings $\xi_i^{(X)}$ and $\xi_j^{(Y)}$ (which are always, to some extent, unidentifiable). The connectivity functions $\phi^{(X)}$ and $\phi^{(Y)}$ (Definition 4) are then similarly equivalent to the node-specific connectivity parameters $\boldsymbol{\pi}^{(X)}$ and $\boldsymbol{\pi}^{(Y)}$ (Definition 3). These functions $\phi^{(X)}$ and $\phi^{(Y)}$ model the general variability of connectivity strength throughout the network, whereas the graphon f models the tendency for regions of the network to aggregate into specific co-communities. The model of Definition 4 is a more general model which is specified similarly for any network size. However, as the networks I consider here are of fixed size, the degree corrected stochastic co-blockmodel (Definition 3) may be a more parsimonious choice. To estimate the generating mechanism of a bipartite network stably, Definition 4 must be replaced by a model with a limited number of parameters, i.e., Definition 3.

The Newman-Girvan modularity (Newman & Girvan, 2004) measures, for a particular partition of a network into communities, the observed number of edges between community members, compared to the expected number of edges between community members without the community partition. The Newman-Girvan modularity may be defined as follows:

Definition 5 (Newman-Girvan modularity)**.** *Define $\mathbf{A} \in \{0,1\}^{n \times n}$ as a symmetric adjacency matrix representing a unipartite network with nodes $i \in \{1,...,n\}$, define $\mathbf{d}$ as the degree vector of the nodes of this network, $d_i = \sum_{j=1}^{n} A_{ij}$, and define the normalising factor d^{++} as the total number of edges, $d^{++} = \sum_{i=1}^{n} d_i$. Define a community, or grouping, of nodes as $g \in G$, where G represents the set of all such groupings of nodes, define the map function $z(i)$ such that $g_a = \{i : z(i) = a\}$, and let $\mathbb{I}[z(i) = z(j)]$ specify whether nodes i and j appear together*

in any community g, such that:

$$\mathbb{I}\left[z(i)=z(j)\right]=\begin{cases}1, & \textit{if nodes } i \textit{ and } j \textit{ are grouped together}\\ & \textit{in any community } g\in G,\\ 0, & \textit{otherwise}.\end{cases}$$

Then, the Newman-Girvan modularity Q_{NG} is defined as:

$$Q_{NG}=\frac{1}{d^{++}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[A_{ij}-\frac{d_i d_j}{d^{++}}\right]\cdot\mathbb{I}\left[z(i)=z(j)\right]. \tag{5.5}$$

The co-modularity is then defined by analogy with the Newman-Girvan modularity (Definition 5) as follows:

Definition 6 (Co-modularity)**.** *With $\mathbf{A}$ given by Definition 3, define $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ as the degree vectors of the X and Y-nodes of the network, $d_i^{(X)}=\sum_{j=1}^{l}A_{ij}$ and $d_j^{(Y)}=\sum_{i=1}^{m}A_{ij}$, and define the normalising factor d^{++} as the total number of edges, $d^{++}=\sum_{i=1}^{m}d_i^{(X)}=\sum_{j=1}^{l}d_j^{(Y)}$. With $g^{(X)}$ and $g^{(Y)}$, $z^{(X)}$ and $z^{(Y)}$ also defined according Definition 3, let $c_t=\{p,q\}\in C$, $t=\{1,...,T\}$, if $T\neq 0$. The enumeration of the pair $\{p,q\}$ is arbitrary, and is to facilitate ease of access of the co-blocks in a chosen order. If $T=0$, then by definition, $C=\varnothing$. The co-block c_t specifies that the X-node grouping $g_p^{(X)}$ is paired with the Y-node grouping $g_q^{(Y)}$; I refer to such a pairing as a 'co-community'. Furthermore, let $\Psi\left(C;G^{(X)},G^{(Y)};i,j\right)\in\{0,1\}$ specify whether nodes i and j appear together in any co-community $c\in C$, such that:*

$$\Psi\left(C;G^{(X)},G^{(Y)};i,j\right)=\begin{cases}1, & \textit{if } \left\{z^{(X)}(i),z^{(Y)}(j)\right\}=c:c\in C,\\ 0, & \textit{otherwise}.\end{cases}$$

Then, the co-modularity Q_{XY} is defined as:

$$Q_{XY}=\frac{1}{d^{++}}\sum_{i=1}^{m}\sum_{j=1}^{l}\left[A_{ij}-\frac{d_i^{(X)}d_j^{(Y)}}{d^{++}}\right]\Psi\left(C;G^{(X)},G^{(Y)};i,j\right). \tag{5.6}$$

I note that for the co-modularity (unlike the Newman-Girvan modularity), we require a set of pairings of X-node groupings with Y-node groupings C, such that each $c_t\in C$ is a pairing of an X-node grouping $g_p^{(X)}\in G^{(X)}$ with a Y-node grouping $g_q^{(Y)}\in G^{(Y)}$. Also, due to the asymmetry of the co-clustering problem, $c_t=\{p,q\}\neq\{q,p\}$. This separately specified set of parings C is not required in the case of the Newman-Girvan modularity, because in the unipartite

network setting, there is only one type of node, and hence node groupings already 'match-up' with one another. This can be visualised, in the unipartite network setting, as community structure present along the leading diagonal of the adjacency matrix, if the nodes are ordered by community. In the co-community setting, an X-node grouping $g^{(X)}$ may be paired in C with many, with one, or with no Y-node groupings $g^{(Y)} \in G^{(Y)}$, and equivalently a Y-node grouping $g^{(Y)}$ may be paired in C with many, with one, or with no X-node groupings $g^{(X)} \in G^{(X)}$. Further, if the X-nodes and Y-nodes of the network are arranged in the adjacency matrix according to the groupings $g^{(X)}$ and $g^{(Y)}$, there is no reason co-communities should appear along the leading diagonal. Hence, the function Ψ in Equation 5.6 generalises the role of the indicator function in Equation 5.5. I also note that sometimes in practice, we must relax the requirement of $\Psi \in \{0, 1\}$; the reason for this becomes clear in the technical derivations in Derivation A which relate to Algorithm 1 (which follows next).

Community detection of k communities can be performed by fitting the degree-corrected stochastic blockmodel. This is equivalent, under many circumstances, to spectral clustering (Bickel & Chen, 2009; Riolo & Newman, 2012; Newman, 2013), which may be carried out by grouping the nodes into k clusters in the space of the eigenvectors corresponding to the 2nd to k^{th} greatest eigenvalues of the Laplacian $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the diagonal matrix of the degree distribution. Co-community detection in a bipartite network of nodes attributed to the variables X and Y (respectively, X-nodes and Y-nodes), can equivalently be performed by degree-corrected spectral clustering (Dhillon, 2001).

A procedure to find an assignment of X and Y-nodes to $k^{(X)}$ X-node groupings ('row clusters') and $k^{(Y)}$ Y-node groupings ('column clusters') respectively, which finds a (possibly locally) optimum value of the co-modularity Q_{XY}, is specified in Algorithm 1:

Algorithm 1. *With* $\mathbf{A}$ *and* Q_{XY} *defined as in Definition 3, and* $\mathbf{d}^{(X)}$ *and* $\mathbf{d}^{(Y)}$ *defined as in Definition 6:*

1. *Calculate the co-Laplacian* $\mathbf{L}_{XY}$ *(Dhillon, 2001) as:*

$$\mathbf{L}_{XY} = \left(\mathbf{D}^{(X)}\right)^{-1/2} \mathbf{A} \left(\mathbf{D}^{(Y)}\right)^{-1/2}, \tag{5.7}$$

   where $\mathbf{D}^{(X)}$ *and* $\mathbf{D}^{(Y)}$ *are the diagonal matrices of* $\mathbf{d}^{(X)}$ *and* $\mathbf{d}^{(Y)}$*, respectively.*

2. *Calculate the singular value decomposition (SVD) of the co-Laplacian* $\mathbf{L}_{XY}$*.*

3. *Separately cluster the* X *and* Y*-nodes in the spaces of the left and right singular vectors corresponding to the 2nd to* $k^{(X)th}$ *and 2nd to* $k^{(Y)th}$ *greatest singular values, respectively,*

of this SVD of $\mathbf{L}_{XY}$.

Technical derivations relating to Algorithm 1 appear in Derivation A, and are based on arguments made previously in the context of unipartite (symmetric) community detection (Newman, 2013), extending them to this context of (asymmetric) co-community detection. I note in particular, that the notion of modularity assumes that within-community edges are more probable than between-community edges, and therefore modularity maximisation is only consistent if constraints are applied to ensure this assumption holds (Zhao *et al.* , 2012). In the community detection setting, under suitable constraints, the solutions which maximise model likelihood and modularity are identical (Bickel & Chen, 2009).

Proposition 1. *The solution which maximises the model likelihood specified in equation 5.3, subject also to the constraint of equation 5.2, is equivalent to the maximum co-modularity assignment obtained via Algorithm 1.*

Proof. The proof appears in Derivation B. It extends arguments made previously in relation to community detection (Newman, 2013) to this context of co-community detection. □

5.3 Selecting the number of co-communities

In order to use Algorithm 1 to carry out co-community detection, we must specify the number of X-node groupings $k^{(X)}$, and the number of Y-node groupings $k^{(Y)}$. The network histogram method of fitting the stochastic blockmodel (Olhede & Wolfe, 2014) in the unipartite/symmetric community detection setting provides a rule-of-thumb method for selecting the optimal number of communities, or blocks, in the model. Fitted in this way, the blockmodel is a valid representation of a network, whatever the generating mechanism of that network, as long as this generating mechanism results in an exchangeable network. The network histogram approximates the graphon, which is a continuous function: the nodes correspond to discrete locations along the graphon margins, ordered in an optimal way to satisfy the smoothness requirement of the graphon. The graphon oracle (Wolfe & Olhede, 2013; Olhede & Wolfe, 2014) defines a good ordering of the nodes, according to graphon smoothness, and community structure. This information is not available in practice, but it can be used to bound the mean integrated squared error of the network histogram approximation to the graphon. This ordering naturally corresponds to community assignments, and the number of communities, or blocks, is determined by the smoothness of the graphon. An intuition for this is by analogy with a wave: if there are many peaks over a fixed distance (i.e., short wavelength), the maximum gradient of the wave will be large, whereas if there are few peaks over the same fixed distance (i.e., long wavelength),

the maximum gradient will be small. Similarly, the more communities, or peaks, that there are in the graphon, the greater the maximum gradient of the graphon will be, and, correspondingly, the less smooth it will be.

5.3.1 Finding the optimal numbers of X and Y-node groupings

In this section I define the anisotropic graphon, which allows us to determine an optimal number of X and Y-node groupings, $k^{(X)}$ and $k^{(Y)}$, from which co-communities can be identified. This relates closely to the network histogram method in the symmetric unipartite community detection setting (Olhede & Wolfe, 2014). In the unipartite community detection setting, the graphon is a symmetric limit object bounded on $(0,1)^2$. It is symmetric because the in that setting, the set of X-nodes is the same as the set of Y-nodes, and hence the smoothness is the same with respect to the corresponding orthogonal directions on the graphon. In contrast, in this co-community detection setting the graphon is asymmetric, having different smoothnesses with respect to the X and Y-nodes. Hence, I refer to this as the 'anisotropic graphon', which is similarly a limit object bounded on $(0,1)^2$. To aid the analyses, we can stretch the anisotropic graphon so that it has the same smoothness with respect to the X-nodes, and with respect to the Y-nodes. It is easy to see that such a transformation exists for all anisotropic graphons. I refer to the result of stretching the anisotropic graphon in this way, as the 'equi-smooth graphon'. Without loss of generality, this transformation can be expressed as a stretch of scale-factor γ with respect to the X-nodes, and a simultaneous stretch of scale-factor $1/\gamma$ with respect to the Y-nodes. I refer to γ as the anisotropy factor. This is formalised as follows.

Definition 7. *For the Lipschitz-continuous anisotropic graphon $f \in L\left((0,1)^2\right)$ defined according to Definition 4, let the anisotropy factor γ define the linear-stretch transformation which maps f onto the Lipschitz-continuous equi-smooth graphon $\widetilde{f} \in L\left((0,\gamma)\times(0,1/\gamma)\right)$. Then,*

$$f(x,y) = \widetilde{f}\left(\gamma x, y/\gamma\right). \tag{5.8}$$

Lipschitz-continuity, in this context, means that the smoothness of the graphon (anisotropic or equi-smooth) is upper-bounded, and I use this bound to calculate the optimal number of X and Y-node groupings.

To determine the optimal number of X and Y-node groupings, $k^{(X)}$ and $k^{(Y)}$, I set these $k^{(X)}$ and $k^{(Y)}$ so as to minimise the mean integrated squared error (MISE) of the blockmodel approximation of the graphon. Following a methodology which is closely related to the network histogram estimator in the symmetric (unipartite) community detection setting (Olhede & Wolfe, 2014), making use of the graphon oracle estimator, an upper bound can be calculated on

this MISE, from a bias-variance decomposition, as follows:

Lemma 1. *With* **A**, m, l, $g^{(X)} \in G^{(X)}$, *and* $g^{(Y)} \in G^{(Y)}$ *defined according to Definition 3, let* ρ *be a deterministic scaling constant which specifies the expected number of edges in the network, such that:*

$$\rho = \mathbb{E}\left(\frac{1}{ml}\sum_{j=1}^{l}\sum_{i=1}^{m} A_{ij}\right),$$

and define piecewise block-approximations to the adjacency matrix, for each pairing of a set of X*-nodes* $g^{(X)}$ *with a set of* Y*-nodes* $g^{(Y)}$*, as:*

$$\bar{A}_{p,q} = \frac{\sum_{i\in g_p^{(X)}, j\in g_q^{(Y)}} A_{ij}}{\left|g_p^{(X)}\right|\left|g_q^{(Y)}\right|}$$

where $|\cdot|$ *represents cardinality. With* $z^{(X)}$ *and* $z^{(Y)}(j)$ *defined according to Definition 3,* $\xi^{(X)}$ *and* $\xi^{(Y)}$ *defined according to Definition 4, and* f *defined according to Definition 7, define alternative map functions* $\widetilde{z}^{(X)}(i')$*,* $i' \in \{1, ..., m\}$*, and* $\widetilde{z}^{(Y)}(j')$*,* $j' \in \{1, ..., l\}$*, which take the ordered locations of the* X *and* Y*-nodes respectively along the graphon margins, as specified by* $\xi^{(X)}$ *and* $\xi^{(Y)}$*, and return the corresponding* X *and* Y*-node groupings, such that* $\widetilde{z}^{(X)}\left(\left\lceil m\cdot\xi_i^{(X)}\right\rceil\right) = z^{(X)}(i)$*, and* $\widetilde{z}^{(Y)}\left(\left\lceil l\cdot\xi_j^{(Y)}\right\rceil\right) = z^{(Y)}(j)$*. Define the graphon oracle estimator as:*

$$\hat{f}(x,y) = \hat{\rho}^{-1}\bar{A}_{\widetilde{z}^{(X)}(\lceil lx\rceil),\widetilde{z}^{(Y)}(\lceil my\rceil)}, \tag{5.9}$$

and let:

$$\iint_{(0,1)^2} f(x,y)dx\,dy = 1. \tag{5.10}$$

With $\widetilde{f}$ *and* γ *defined as in Definition 7, let* $\widetilde{M}$ *be the maximum gradient of* $\widetilde{f}$*, and let* $h^{(X)}$ *and* $h^{(Y)}$ *be 'bandwidth' model parameters with respect to the* X *and* Y *nodes respectively. Then, the graphon oracle upper bound on the MISE of the blockmodel estimate of the graphon function* $\hat{f}$ *is:*

$$\begin{aligned}\mathrm{MISE}\left(\hat{\mathrm{f}}\right) \leq\ & \widetilde{M}^2\left\{\gamma^2\cdot\frac{\left(h^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2}\cdot\frac{\left(h^{(Y)}\right)^2}{l^2}\right\} \\ & + 2\widetilde{M}^2\left\{\gamma^2\cdot\frac{1}{4m} + \frac{1}{\gamma^2}\cdot\frac{1}{4l}\right\}\{1+o(1)\} + \frac{1}{\rho\cdot h^{(X)}\cdot\ h^{(Y)}}\{1+o(1)\}.\end{aligned} \tag{5.11}$$

Proof. See Derivation C. □

I note that the sets of nodes represented by the groupings $g^{(X)} \in G^{(X)}$ and $g^{(Y)} \in G^{(X)}$ are contiguous along the graphon margins (corresponding to the canonical graphon ordering,

Airoldi *et al.* (2013); Chan & Airoldi (2014)), but that these nodes are not contiguous along the adjacency matrix margins. Thus, we need to specify how nodes map to the groupings $g^{(X)}$ and $g^{(Y)}$ in a different way for the graphon, as compared to the adjacency matrix. This difference is accounted for by using different mapping functions: $\widetilde{z}^{(X)}(i')$ and $\widetilde{z}^{(Y)}(j')$ for the graphon, and $z^{(X)}(i)$ and $z^{(Y)}(j)$ for the adjacency matrix. I.e., $\widetilde{z}^{(X)}(i')$ and $\widetilde{z}^{(Y)}(j')$ are required to specify the (contiguous) ranges and locations of the X and Y-node groupings $g^{(X)}$ and $g^{(Y)}$ on the graphon margins, and equivalently $z^{(X)}(i)$ and $z^{(Y)}(j)$ for their (non-contiguous) locations on the adjacency matrix margins.

Using the MISE formulation of Lemma 1, we can estimate the optimal numbers of X and Y-node groupings, $k^{(X)}$ and $k^{(Y)}$.

Proposition 2. *With m and l defined as in Definition 3, and $\widetilde{M}$ and ρ defined as in Lemma 1, the optimal number of X and Y-node groupings, $k^{(X)}$ and $k^{(Y)}$ respectively, are:*

$$k^{(X)} = \gamma \cdot (ml)^{\frac{1}{4}} \cdot \left(2\rho\widetilde{M}^2\right)^{\frac{1}{4}} \tag{5.12}$$

and

$$k^{(Y)} = \frac{1}{\gamma} \cdot (ml)^{\frac{1}{4}} \cdot \left(2\rho\widetilde{M}^2\right)^{\frac{1}{4}}. \tag{5.13}$$

Proof. The proof of this proposition is developed from the equivalent proof for the case of the isotropic graphon (corresponding to community detection in unipartite networks) (Olhede & Wolfe, 2014). The optimal bandwiths $h^{(X)*}$ and $h^{(Y)*}$ can be found by differentiating the expression for the MISE of equation 5.11 with respect to $h^{(X)}$ and setting to zero, and doing the same with respect to $h^{(Y)}$, and combining the resulting equations. To calculate $k^{(X)}$ and $k^{(Y)}$, substitute these optimal bandwiths $h^{(X)*}$ and $h^{(Y)*}$ into $k^{(X)} = m/h^{(X)*}$ and $k^{(Y)} = l/h^{(Y)*}$, which leads to equations 5.12 and 5.13. □

I note that the above proof of Proposition 2 implies constant group sizes for the X-nodes, and constant group sizes for the Y-nodes. This assumption is relaxed in the practical implementation of this methodology I propose: this point is discussed further in Section 5.3.2.

5.3.2 Practical estimation of the number of X and Y-node groupings

I implement spectral clustering by including a standard k-means step, to group the X and Y-nodes in the spaces of the left and right singular vectors corresponding to the 2nd to $k^{(X)\text{th}}$ and 2nd to $k^{(Y)\text{th}}$ greatest singular values, respectively, of the singular value decomposition of the co-Laplacian $\mathbf{L}_{XY}$ (equation 5.7). This k-means step does not produce identical group sizes, however I note that the estimates of $k^{(X)}$ and $k^{(Y)}$ defined according to equations 5.12 and

5.13 assume that the X and Y node groupings are the same size (i.e., that the blocks in the blockmodel are all the same size with respect to the X-nodes, and separately with respect to the Y-nodes). I relax this requirement in practice, because after examining several empirical data-sets of the type presented in the next section, I observed that the group sizes produced by this type of regularised degree-corrected spectral clustering, tend not to vary significantly in size (there are no 'giant clusters'). Further, this requirement of identical group sizes is not physically realistic in the practical examples I present in the next section, and in many other real scenarios.

To estimate $\widetilde{M}$ and γ, I approximate the maximum slope of the graphon separately in the directions corresponding to the X and Y-nodes, by considering the top component of the singular value decomposition of the adjacency matrix $\mathbf{A}$. This is equivalent to the rule-of-thumb procedure in the network histogram method, in the symmetric/unipartite community detection scenario (Olhede & Wolfe, 2014). The top left and right singular vectors are ordered, and their gradients and values at their midpoints (the expected points of maximum slope) are estimated as $\hat{p}_X$ and $\hat{b}_X$ respectively for the X-nodes, and $\hat{p}_Y$ and $\hat{b}_Y$ respectively for the Y-nodes. By thinking of this singular value decomposition as a factorisation of the scaled, discrete-sampled graphon (i.e., the ordered adjacency matrix), denoting the greatest singular value as ν, leads to the linear approximations for the maximum gradient of the isotropic graphon M in the directions of the X and Y-nodes, M_X and M_Y respectively:

$$\hat{M}_X = \frac{\nu}{\rho}\hat{p}_X\hat{b}_Y m, \qquad \hat{M}_Y = \frac{\nu}{\rho}\hat{b}_X\hat{p}_Y l,$$

where m and l are the number of X and Y-nodes respectively (as previously defined). These factors m and l take account of the fact that the isotropic graphon margins are bounded on $[0, 1]$, whereas the adjacency matrix margins take the values $\{1, ..., m\}$ and $\{1, ..., l\}$, and the edge density factor ρ (defined as in Lemma 1) normalises with respect to the adjacency matrix realisation, such that the above estimates are independent of edge density ρ. The linear stretch transformation γ defines the maximum gradients of the equi-smooth graphon as $\widetilde{M}_X = \gamma M_X$ and $\widetilde{M}_Y = M_Y/\gamma$ respectively, and hence an estimate of the squared maximum gradient of the isotropic graphon can be found as:

$$\hat{\widetilde{M}}^2 = \gamma^2 \cdot \hat{M_X}^2 + \frac{1}{\gamma^2} \cdot \hat{M_Y}^2 = \frac{\nu^2}{\rho^2}\left(\gamma^2 \cdot \hat{p}_X^2\hat{b}_Y^2 m^2 + \frac{1}{\gamma^2} \cdot \hat{b}_X^2\hat{p}_Y^2 l^2\right).$$

Using the assumption that the equi-smooth graphon is Lipschitz-continuous, with the same upper-bound on its smoothness with respect to both the X and Y nodes, i.e., $\widetilde{M}_X = \widetilde{M}_Y$, $\implies$

$\gamma M_X = M_Y/\gamma$, we can estimate γ as:

$$\hat{\gamma}^2 = \frac{\hat{M}_Y}{\hat{M}_X}. \tag{5.14}$$

5.3.3 Model simplifications

We can draw an analogy between bandwidth estimation in the anisotropic graphon, and the anisotropic kernel (Wand & Jones, 1993; Duong & Hazelton, 2003). Similar to bivariate kernel density estimation, we may be able to achieve a more parsimonious model, if we can justifiably assume that the smoothness of the anisotropic graphon is the same with respect to both the X and Y nodes. This is the same as saying that the anisotropy factor $\gamma \approx 1$, and that $M_X \approx M_Y$.

Proposition 3. *With γ defined as in Definition 7, testing the following null and alternative hypotheses:*

$$H_0: \quad \gamma = 1, \qquad H_1: \quad \gamma \neq 1,$$

under the null, the estimated anisotropy constant $\hat{\gamma}$ follows the law of:

$$\gamma^2 \sim \mathcal{N}\left(1, \tau^2\right),$$

where the variance τ^2 is estimated from the linear model estimates of $\hat{b}_X$, $\hat{p}_X$, $\hat{b}_Y$ and $\hat{p}_Y$, such that:

$$\hat{\tau}^2 = \frac{\widehat{\mathrm{Var}(\hat{b}_X)}}{\hat{b}_X} + \frac{\widehat{\mathrm{Var}(\hat{p}_Y)}}{\hat{p}_Y} + \frac{\widehat{\mathrm{Var}(\hat{p}_X)}}{\hat{p}_X} + \frac{\widehat{\mathrm{Var}(\hat{b}_Y)}}{\hat{b}_Y} + 2\frac{\widehat{\mathrm{Cov}(\hat{b}_X, \hat{p}_X)}}{\hat{b}_X \hat{p}_X} + 2\frac{\widehat{\mathrm{Cov}(\hat{b}_Y, \hat{p}_Y)}}{\hat{b}_Y \hat{p}_Y}.$$

Proof. See Derivation C. □

If I fail to reject H_0 in Proposition 3, then I take it that $M_X \approx M_Y$, and $\gamma \approx 1$, and hence we have that $k^{(X)} = k^{(Y)} = (ml)^{\frac{1}{4}} \cdot \left(2\rho\widetilde{M}^2\right)^{\frac{1}{4}}$, i.e., there are the same number groupings of X-nodes and of Y-nodes. This assumption is implicitly made in widely-used previous solutions to the co-community detection problem, as in (Dhillon, 2001).

It is also worth noting that if the number of X and Y nodes, m and l respectively, are very different, then this does not preclude $\gamma \approx 1$: there are, in effect, two independent mappings, which take place in getting from the adjacency matrix to the equi-smooth graphon. The first of these linearly maps the adjacency matrix X-nodes $i \in \{1, ..., m\}$ and Y-nodes $j \in \{1, ..., l\}$, onto the anisotropic graphon bounded on $(0, 1)^2$. The second linearly stretches the anisotropic

graphon by scale factor γ with respect to the X nodes, and $1/\gamma$ with respect to the Y-nodes, giving the equi-smooth graphon.

A co-blockmodel approximates the graphon, which requires that $k^{(X)}$ and $k^{(Y)}$ grow with m and l, respectively. However, the co-blockmodel also allows us to model the scenario in which $k^{(X)}$ and $k^{(Y)}$ grow at different rates with respect to m and l, i.e., that the number of X and Y-node groupings are unrelated.

5.4 Identification and comparison of co-communities

Fitting the stochastic co-blockmodel by spectral clustering as described in Algorithm 1, involves using k-means to cluster the X and Y-nodes in the spaces of the left and right singular vectors of the co-Laplacian (equation 5.7). However, as will be subsequently illustrated, this leads to a problem of identifiability which does not arise when fitting the symmetric stochastic blockmodel to unipartite networks by spectral clustering. This problem of identifiability is precisely the question of estimating the set C (Definition 6) of pairings of X-node groupings $g^{(X)} \in G^{(X)}$ with Y-node groupings $g^{(Y)} \in G^{(Y)}$.

Fitting the symmetric blockmodel in the unipartite community detection setting, there are exactly $k = k^{(X)} = k^{(Y)}$ communities (because of symmetry). Each row grouping matches up with exactly one column grouping, because the row and column groupings are the same thing. On the other hand, fitting the asymmetric co-blockmodel by spectral clustering as in Algorithm 1 leads to $k^{(X)}$ and $k^{(Y)}$ row and column clusters. Hence, these $k^{(X)}$ and $k^{(Y)}$ row and column clusters provide $k^{(X)} \times k^{(Y)}$ potential co-communities. Which of these are significant? The best-known solution to this problem (Dhillon, 2001), instead of clustering the X and Y-nodes separately, instead normalises and concatenates the left and right singular vectors, and then clusters all the nodes at once. However, this approach has serious limitations: it again requires $k^{(X)} = k^{(Y)}$. Also, if the two types of nodes represent very different types of observations, then in practice I have found that method to perform less well. For this comparison, I define performance in terms of overlap with previously-defined groupings of nodes or variables.

So how should we assess and compare the $k^{(X)} \times k^{(Y)}$ potential co-communities, each of which is a different pairing of an estimated X-node grouping $\hat{g}^{(X)} \in \hat{G}^{(X)}$, with an estimated Y-node grouping $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$, to provide an assignment of the X-nodes and Y-nodes to co-communities, which is in some sense optimal? In practice, I expect the number of co-communities, $T = |C|$ (where $|\cdot|$ represents cardinality), to be significantly less than $k^{(X)} \times k^{(Y)}$. In the unipartite community detection setting, $k^{(X)} = k^{(Y)} = k$, and hence in effect there we have $T = k = \sqrt{k^{(X)} \times k^{(Y)}}$.

To estimate the set of co-communities, $c_t \in C$, $t = \{1, ..., T\}$, in this bipartite network setting, I calculate the 'local co-modularity' for each pairing $\hat{g}^{(X)}$ with $\hat{g}^{(Y)}$, by considering a relevant sub-part of the co-modularity matrix **B** (equation 5.15):

Definition 8 (Local co-modularity)**.** *With* $\mathbf{A}$ *given by Definition 3, with* $\mathbf{d}^{(X)}$, $\mathbf{d}^{(Y)}$ *and* d^{++} *given by Definition 6, with*

$$B_{ij} = A_{ij} - \frac{d_i^{(X)} d_j^{(Y)}}{d^{++}}, \qquad \mathbf{B} = \mathbf{A} - \frac{1}{d^{++}} \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)}\right)^\top, \tag{5.15}$$

and with the set of X*-node groupings and the set of* Y*-node groupings estimated according to Algorithm 1 as* $\hat{G}^{(X)}$ *and* $\hat{G}^{(Y)}$ *respectively, where* $\left|\hat{G}^{(X)}\right| = k^{(X)}$ *and* $\left|\hat{G}^{(Y)}\right| = k^{(Y)}$*, where* $|\cdot|$ *represents cardinality, for a particular pairing of estimated* X*-node grouping* $\hat{g}^{(X)} \in \hat{G}^{(X)}$ *with estimated* Y*-node grouping* $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$*, the local co-modularity* $Q_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right)$ *is defined as:*

$$Q_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right) = \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij}. \tag{5.16}$$

Each of the $k^{(X)} \times k^{(Y)}$ possible pairings of $\hat{g}^{(X)}$ with $\hat{g}^{(Y)}$ can be defined, or not, as a co-community; doing so means that they are included in, or excluded from, the estimated set of co-communities $\hat{C}$ (Definition 6). To consider all permutations, $2^{k^{(X)} \times k^{(Y)}}$ such assignments would need to be considered, which would be computationally very demanding. However, this problem can be avoided by defining summary statistics targeted for particular purposes. The three such purposes which I consider here are described in the following subsections: 5.4.1 Comparing potential co-communities and assessing their strength; 5.4.2 Arranging the co-communities for visualisation; 5.4.3 Defining an algorithmic objective function to be optimised, when determining co-community partitions.

5.4.1 Comparing and assessing significance of co-communities

Under a null model of no co-community structure, $\theta_{z^{(X)}(i), z^{(Y)}(j)} =$ constant, for all i, j. Therefore, referring to the log-linear model (Perry & Wolfe, 2012), equation 5.1 becomes:

$$A_{ij} \sim \text{Bernouilli}\left(\frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}}\right), \tag{5.17}$$

where I have defined:

$$\theta_{z^{(X)}(i), z^{(Y)}(j)} = 1/\pi^{++}. \tag{5.18}$$

Hence under this null,

$$\mathbb{E}\left(A_{ij}\right) = \frac{\pi_i^{(X)}\pi_j^{(Y)}}{\pi^{++}} \qquad \Longrightarrow \qquad \mathbb{E}\left(B_{ij}\right) = 0.$$

I define the informal idealised quantities $\widetilde{\mathbf{B}}$ and $\widetilde{Q}_{XY}$ in comparison with equations 5.15 and 5.16:

$$\widetilde{\mathbf{B}} = \mathbf{A} - \frac{1}{\pi^{++}}\boldsymbol{\pi}^{(X)}\left(\boldsymbol{\pi}^{(Y)}\right)^{\top}, \tag{5.19}$$

and

$$\widetilde{Q}_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right) = \frac{1}{\pi^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} \widetilde{B}_{ij}, \tag{5.20}$$

where the empirical degree distributions $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ have been replaced by the theoretical node connectivity parameters $\boldsymbol{\pi}^{(X)}$ and $\boldsymbol{\pi}^{(Y)}$, and the empirical normalisation factor d^{++} is also replaced by the theoretical normalisation factor π^{++}.

If the pairing of X and Y-node groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ exhibit some co-community structure, then equation 5.18 no longer holds, and so the null model does not hold either. The stronger this co-community structure is, the further we move from the null model, and the greater θ becomes relative to $1/\pi^{++}$. This corresponds to $\mathbb{E}\left(A_{ij}\right)$ becoming larger than $\pi_i^{(X)}\pi_j^{(Y)}/\pi^{++}$, which is equivalent to the observed number of edges in the co-community becoming greater than the expected, under the null of no co-community structure. This in turn means that $\widetilde{Q}_{XY}$ also becomes more positive. In other words, the further we move from the null model, the greater tendency of the X-nodes and Y-nodes of these groups to form connections with one another (compared with their expected propensity to make connections with any nodes, of the opposite type), and therefore constitute a strong co-community. Hence, a parsimonious method of comparing potential co-communities is simply to compare their local co-modularity, $Q_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right)$. This naturally leads to a ranking of potential co-communities according to their strength.

An estimate of statistical significance of a potential co-community can also be made, as follows. Noting that, with adjacency matrix $\mathbf{A}$ defined according to the Bernoulli distribution of Definition 3, with fixed $\theta_{z^{(X)}(i),z^{(Y)}(j)} = 1/\pi^{++}$,

$$\mathrm{Var}\left(\widetilde{B}_{ij}\right) = \mathrm{Var}\left(A_{ij}\right) = \left(\frac{\pi^{(X)}\pi^{(Y)}}{\pi^{++}}\right)\left(1 - \frac{\pi^{(X)}\pi^{(Y)}}{\pi^{++}}\right),$$

and assuming probabilities of observing links between different pairs of nodes are independent,

the variance of $\widetilde{Q}_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right)$ can be approximated as:

$$\mathrm{Var}\left(\widetilde{Q}_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right)\right) = \frac{1}{\left(\pi^{++}\right)^2} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} \left(\frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}}\right)\left(1 - \frac{\pi_i^{(X)} \pi_j^{(Y)}}{\pi^{++}}\right), \quad (5.21)$$

where the factor $1/\left(\pi^{++}\right)^2$ is due to the factor $1/\left(\pi^{++}\right)$ in equation 5.20. Hence, assuming $\mathbf{d}^{(X)} \xrightarrow{p} \boldsymbol{\pi}^{(X)}$, $\mathbf{d}^{(Y)} \xrightarrow{p} \boldsymbol{\pi}^{(Y)}$ and $d^{++} \xrightarrow{p} \pi^{++}$, and assuming the potential co-community defined by $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ is comprised of sufficiently many nodes for a Gaussian approximation to hold, we can test the significance of $Q_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right)$ with a z-test, with zero mean and with $\mathrm{Var}\left(Q_{XY}\right)$ estimated as $\mathrm{Var}\left(\widetilde{Q}_{XY}\right)$ in equation 5.21, also replacing $\pi_i^{(X)}$ with $d_i^{(X)}$, $\pi_j^{(Y)}$ with $d_j^{(Y)}$ and π^{++} with d^{++}. A pairing $\hat{g}_p^{(X)}$ and $\hat{g}_q^{(Y)}$ is then defined as a co-community $\hat{c}$ and included in $\hat{C}$ (Definition 6), i.e., $\{p, q\} = \hat{c} \in \hat{C}$, if and only if this pairing $\hat{g}_p^{(X)}$ with $\hat{g}_q^{(Y)}$ is significant according to this z-test, at some significance level. I note that, in practice, this is only a rough approximation of significance, also because by specifying in advance the co-community node-groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$, we have introduced dependencies between the X and Y-nodes of this co-community.

5.4.2 Arranging the co-communities for visualisation

A standard task in exploratory data analysis using variants of the stochastic block model, is arranging the detected communities so they can be visualised in a helpful way. This visualisation is usually carried out by way of a heatmap representation of the adjacency matrix with the nodes grouped into communities. In the symmetric/unipartite community detection scenario, the communities occur along the leading diagonal of this ordered adjacency matrix. The communities themselves are often ordered along the leading diagonal according to their edge densities. In the bipartite co-community detection setting, co-communities may be present away from the leading diagonal, and there is no longer a restriction on how many co-communities a node may be part of - although I do not consider here the possibility of overlapping co-communities.

I propose then, that once the X-node groupings and Y-node groupings have been determined by spectral clustering as described above, a natural way to order these groups with respect to one another, is via row and column co-modularities, which I define as follows.

Definition 9. *With d^{++} given by Definition 6, and with $\mathbf{B}$ given by Definition 8, with with the set of X-node groupings and the set of Y-node groupings estimated according to Algorithm 1 as $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$ respectively, the row and column modularities $Q_{row}\left(\hat{g}^{(X)}\right)$ and $Q_{column}\left(\hat{g}^{(Y)}\right)$*

are defined, for $\hat{g}^{(X)} \in \hat{G}^{(X)}$ *and* $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$*, as:*

$$Q_{row}\left(\hat{g}^{(X)}\right) = \sum_{\hat{g}^{(Y)} \in \hat{G}^{(Y)}} \left| \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij} \right| \tag{5.22}$$

and

$$Q_{column}\left(\hat{g}^{(Y)}\right) = \sum_{\hat{g}^{(X)} \in \hat{G}^{(X)}} \left| \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij} \right|. \tag{5.23}$$

Considering the absolute values of the local co-modularities in these sums serves to prioritise the most extreme choices of divisions of nodes into co-communities, according to their local co-modularities. On the other hand if absolute values were not considered here, the row and column modularities would always be zero, because the rows and columns of $\mathbf{B}$ must always sum to zero. The row and column co-modularities are the sums, respectively, of the absolute values of the local co-modularities along the rows and columns respectively, of the ordered adjacency matrix. Hence, they represent a measure of how extreme the co-community divisions are, in each row and column, according to the groupings defined by $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$. By ordering the X-node and Y-node groupings by decreasing $Q_{row}\left(\hat{g}^{(X)}\right)$ and $Q_{column}\left(\hat{g}^{(Y)}\right)$ respectively, co-communities with the largest local co-modularities will tend to congregate towards the top-left of the ordered adjacency matrix. This is a natural arrangement for visualisation as a heatmap, because it tends to place the strongest co-communities together in this corner, and so the attention is intuitively drawn to this region.

I note that there may be other equally effective ways of arranging the adjacency matrix for visualisation as a heatmap. However, this method is effective, and it is a parsimonious solution in the context of co-modularity, because row and column modularities are very simply and intuitively related to local co-modularity. In the case that there is no co-community structure present, such as under the null model of equation 5.17, then Q_{row} and Q_{column} as defined in Definition 9 would also tend to be close to zero, and the ordering would cease to be meaningful. However, if there are even a few significant co-communities present, their corresponding X and Y-node groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ would stand out, as assessed by Q_{row} and Q_{column}. Therefore these $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$ would be placed at the top of the respective orderings, with the co-community pairings tending towards in the top-left corner. The other rows and columns, which do not contain significant co-communities, would have corresponding Q_{row} and Q_{column} close to zero. Hence, these rows and columns would be naturally ordered according to their irrelevance. They would accordingly be placed further away from the top-left of the heatmap, giving the intuition that they are unimportant.

5.4.3 Defining an objective function for optimising the co-community partitions

Defining an objective function over the whole network, in terms of the assignments of the nodes to X-node and Y-node groupings $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$, allows optimisation of these node assignments. It also provides a means of comparison of algorithmic parameters and other design choices in the practical implementation of the methods. It would be most ideal, for a trial assignment of nodes to $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$, to estimate the set of co-communities $\hat{C}$ using the method of Section 5.4.1, and then to calculate the co-modularity according to Definition 6. However, for a large number of repetitions within an algorithm, or for an iterative search and optimisation, this would be computationally inefficient. Instead, I define the global co-modularity to be used as an objective function for such purposes, as follows:

Definition 10. *With d^{++} given by Definition 6, and with $\mathbf{B}$ given by Definition 8, with with the set of X-node groupings and the set of Y-node groupings estimated according to Algorithm 1 as $\hat{G}^{(X)}$ and $\hat{G}^{(Y)}$ respectively, the global co-modularity is defined, for $\hat{g}^{(X)} \in \hat{G}^{(X)}$ and $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$, as:*

$$Q_{global} = \sum_{\hat{g}^{(Y)} \in \hat{G}^{(Y)}} \sum_{\hat{g}^{(X)} \in \hat{G}^{(X)}} \left| \frac{1}{d^{++}} \sum_{i \in \hat{g}^{(X)}} \sum_{j \in \hat{g}^{(Y)}} B_{ij} \right|. \tag{5.24}$$

For a pairing $\hat{g}^{(X)}$ and $\hat{g}^{(Y)}$, the local co-modularity $Q_{XY}\left(\hat{g}^{(X)}, \hat{g}^{(Y)}\right)$ represents the strength of the co-community structure in that grouping of X-nodes and Y-nodes. If the absolute value was not considered in the sum, Q_{global} would always be zero. Hence, by prioritising a sum of the absolute values of the local co-modularity of all pairings $\hat{g}^{(X)}$ with $\hat{g}^{(Y)}$, I prioritise an extreme division of the X-nodes and Y-nodes into co-communities, as measured by the local co-modularity. This therefore corresponds to an extreme partition in terms of co-community structure, as assessed by co-modularity.

Spectral clustering usually requires the nodes to be grouped in the spaces of the top singular vectors of the co-Laplacian, and this grouping is often carried out by k-means, as described in Algorithm 1. Because k-means optimisation is not convex, the converged result may be a local optimum. Hence, implementations of k-means often begin at a random start-point, with the optimisation run several times from random start-points, choosing the result which is in some sense optimal. In the community-detection setting, a natural statistic to maximise in this optimisation is the Newman-Girvan modularity. An equivalent statistic here to maximise in this co-community detection setting is hence the global co-modularity, which is intuitively linked to the local co-modularity measure of co-community structure. In the community-detection

setting, assignments to communities can also be optimised by carrying out node-swapping between communities, in order to maximise the Newman-Girvan modularity (Blondel *et al.* , 2008). The global co-modularity is a statistic which could be equivalently maximised, in this co-community detection setting.

5.5 Examples

I now present results of applying the above methodology to simulated data, and to linked DNA methylation and gene expression data. I fit the degree-corrected stochastic co-blockmodel by spectral clustering, as detailed above, with the following additional practical details.

In the context of community detection, fitting the degree-corrected stochastic blockmodel using spectral clustering, when calculating the Laplacian it is advantageous to slightly inflate the degree distribution (regularisation) (Qin & Rohe, 2013), a trick which made Google's original page-rank algorithm (Page *et al.* , 1999) so effective in web-searching. Here in the co-community detection setting, correspondingly when calculating the co-Laplacian (equation 5.7), I inflate the diagonals of $\mathbf{D}^{(X)}$ and $\mathbf{D}^{(Y)}$ by the medians of $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$, respectively. Further, when fitting variants of the stochatic blockmodel by spectral clustering with k-means, nodes with small leverage score (which are usually low-degree nodes) can be excluded from the k-means step (Qin & Rohe, 2013); this practice is also followed here. I note that these regularisation steps have not previously been carried out in this co-community detection / co-clustering setting.

I also note that while spectral clustering is in general computationally intensive, binary adjacency matrices such as those dealt with in this setting tend to be very sparse. Further, we only require $k = \mathrm{Max}\left(k^{(X)}, k^{(Y)}\right)$ components of the singular value decomposition, a number which tends to be two or more orders of magnitude smaller than the maximum dimension of the adjacency matrix. Efficient computational methods exist to find the top few components in the singular value decomposition of large sparse matrices (Sørensen, 1992; Lehoucq & Sørensen, 1996), with implementations in *Matlab* and *R*, meaning that these methods are easy

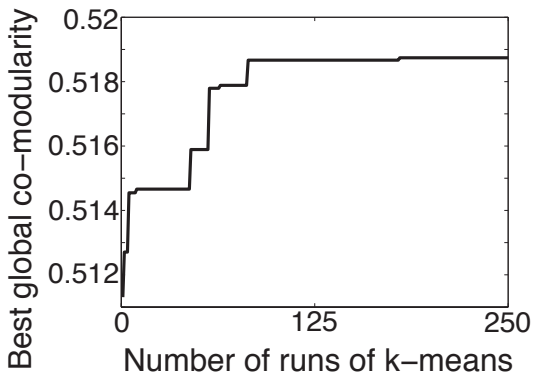


Figure 5.1: *Convergence of the co-modularity. The co-modularity converges well to a maximum, within 250 runs of k-means, in the linked DNA methylation and gene expression data. For reference, the co-modularity is consistently found to be 0 when calculated based on randomly assigned co-community partitions of similar size.*

to implement and practical for large networks.

The k-means clustering algorithm begins with a random start-point, and hence it can provide a different result each time it is run. I therefore run the k-means step in the spectral clustering several times, choosing the result which maximises the global co-modularity (equation 5.24). I run k-means repeatedly until the output is visually assessed to have stabilised, at which point it can be seen from the convergence plot that there is very little, if any, improvement in co-modularity achieved by further runs of k-means. An example of such convergence in the linked DNA methylation and gene expression data is presented in Section 5.5.2 is shown in Figure 5.1.

5.5.1 Simulation study

I carried out a simulation study, to evaluate the effectiveness of this co-community detection methodology against generated networks with known ground-truth co-communities. A classic generative model for exchangeable random networks with heterogenous degrees is the logistic-linear model (Perry & Wolfe, 2012). I use a version here for bipartite networks, with additional co-community structure. This additional co-community structure takes the form of 'blocks'. This block structure is very general: as noted in (Olhede & Wolfe, 2014) it can be used as a model for community structure in relation to many real data-sets for which the true generative mechanism of the community structure is not exactly such block structure. The generative model for this simulation study is defined as:

$$\text{Logit}\left(p_{ij}\right) = \alpha_i^{(X)} + \alpha_j^{(Y)} + \theta_{ij},$$

where p_{ij} defines the probability of an edge being observed between nodes i and j. I choose to use this model, because the parameters can take any real values, and the edge probabilities p_{ij} will still be between 0 and 1. This model only deviates from the equivalent log model when the parameter values become very large, which is what prevents p_{ij} from reaching (and exceeding) 1 (Perry & Wolfe, 2012). Further, the blockmodel approximates any smooth function, and hence the model can be used purely in the sense of approximation (Olhede & Wolfe, 2014; Choi *et al.* , 2014). The node-specific parameters $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ are elements of parameter vectors $\boldsymbol{\alpha}^{(X)}$ and $\boldsymbol{\alpha}^{(Y)}$ which define power law degree distributions for the X and Y-nodes. We would like power-law degree distributions for the nodes; this is a characteristic of scale-free networks (Barabási & Oltvai, 2004), which are found to be physically realistic in a wide range of scenarios, including biological networks (Wagner, 2002), and social networks (Barabási & Albert, 1999). The parameters $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ are each generated as the logarithms of samples

taken from a bounded Pareto distribution as in (Olhede & Wolfe, 2012). I note that because $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ are chosen to be random, the generated networks are exchangeable (Kallenberg, 2005), whereas if $\alpha_i^{(X)}$ and $\alpha_j^{(Y)}$ were defined deterministically, these networks would instead be generated under the inhomogenous random graph model (Bollobás *et al.* , 2007). The co-community parameter θ_{ij} is allowed to take two values: $\theta_{ij} = \theta_{\text{in}}$ if i and j are in the same co-community, and $\theta_{ij} = \theta_{\text{out}}$ otherwise, which is equivalent to the modelling constraint I applied in equation 5.2. After generating the p_{ij}, the network is generated by sampling each A_{ij},

$$A_{ij} \sim \text{Bernouilli}\left(p_{ij}\right).$$

I note that this is the correctly specified model, for the co-community detection which I describe, and carry out in this chapter.

The co-communities themselves are planted in the network as randomly chosen groups of 150 of each type of node, with the maximum number of co-communities equal to $k^{(X)} \times k^{(Y)}$. By analogy with the unipartite/symmetric community detection setting, I choose to set the number of co-communities T as the square-root of this theoretical maximum, $T = \sqrt{k^{(X)} \times k^{(Y)}}$. As discussed in Section 5.4, in the unipartite community detection setting there is a constraint on the number of communities, $k = k^{(X)} = k^{(Y)}$, because the X-node and Y-node groupings are the same thing. This constraint does not exist in the bipartite co-community detection setting, and so the theoretical maximum number of co-communities is $k^{(X)} \times k^{(Y)}$, i.e., the square of the number of communities in the equivalent symmetric community detection setting. However, I expect the number of co-communities to be significantly less than this in practice, and so by default, I choose $T = \sqrt{k^{(X)} \times k^{(Y)}}$ as the number of co-communities, although I note that many other choices would also be valid here.

I test the methods on networks generated with $k^{(X)}$ and $k^{(Y)}$ ranging from 8 and 6 respectively up to 80 and 60 respectively (corresponding to values of numbers of nodes, m and l, ranging from 1200 and 800 up to 12000 and 8000, respectively). I also test the methods on networks generated with values of θ_{in} from 10 to 50, which corresponds to within co-community edge density $\rho_{\text{in}} \in \{0.039, 0.15, 0.34, 0.6\}$, and I set $\theta_{\text{out}} = 1$, corresponding to outside or between co-community edge density $\rho_{\text{in}} = 0.0013$. For each combination of parameters, I carry out 50 repetitions of network generation and co-community detection, to enable assessment of the variability of the accuracy of the co-community detection (with more repetitions, the computational cost becomes prohibitive).

After generating the networks, I detect co-communities according to the methods described

above, based on the same values of $k^{(X)}$ and $k^{(X)}$ that I used to generate the networks. I keep these values the same, to understand specifically how the co-community detection methodology is working. This means there are $k^{(X)} \times k^{(Y)}$ potential co-communities, and I assess each in terms of strength and significance, as discussed in Section 5.4.1. Hence, I define the estimated set of co-communities $\hat{C}$, as all combinations of detected X and Y-node groupings $\hat{g}^{(X)} \in \hat{G}^{(X)}$ with $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$ which are significant according to a z-test with zero mean and variance calculated as in Equation 5.21. I define significance according to FDR (false discovery rate) corrected (Benjamini & Hochberg, 1995) p-value < 0.05. This tends to result in more co-communities being detected than were originally planted (primarily due to some being split), however I note that the main aim of this methodology is to find a good representation of the underlying co-community structure (as assessed by co-modularity), rather than to reproduce it exactly.

To compare detected co-communities with the ground-truth planted co-communities, I use the normalised mutual information (NMI) (Danon *et al.* , 2005). The NMI compares the numbers of nodes which appear together in the found co-communities, compared with whether they appeared together in the planted co-communities (adjusted for group sizes). It has been used previously in the co-community detection context (Larremore *et al.* , 2014), as well as the unipartite community-detection context (Zhao *et al.* , 2012). The NMI takes the value 1 if the co-communities are perfectly reproduced in the co-community detection, and 0 if they are not reproduced at all, and somewhere in between if they are partially reproduced. The results, together with examples of randomly generated adjacency matrices, are shown in Figure 5.2, which shows that the method performs well as long as there is sufficient within-co-community edge density, and performs well as the number of co-communities increases.

5.5.2 Application to linked DNA methylation and gene expression data

I present an example of a practical application of these methods to a challenging problem analysing linked DNA methylation data and gene expression data. Much is still unknown about the interaction between DNA methylation patterns and gene expression patterns (Jones, 2012). It is of interest to uncover groups of genes with methylation patterns which are linked to the expression patterns of other groups of genes, to allow biological hypotheses to be formed, which can then be investigated further, experimentally and computationally. Hence, this is a natural scenario to be approached with co-community detection, as the method offers the potential to uncover latent structure not easily identifiable otherwise.

As a measure of the DNA methylation (DNAm) pattern of each gene, I choose to consider here intra-gene DNA methylation variability (IGV), as presented in chapter 2, as it is a per-gene

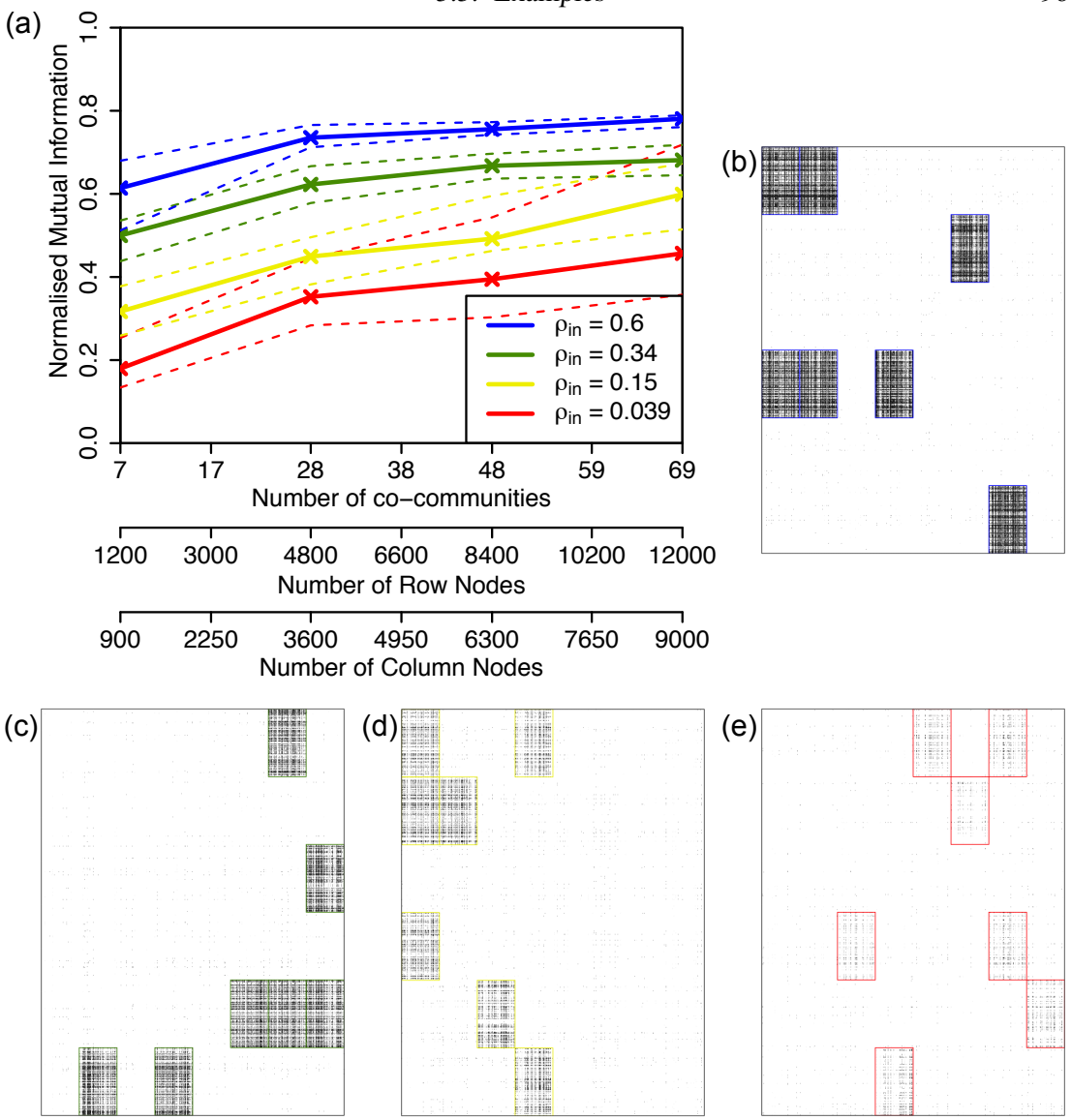


Figure 5.2: *Simulation study.*
(a) Normalised mutual information (NMI) compares detected co-communities with ground-truth planted co-communities. (b)-(e) Examples of generated networks all with $nR = 1200$, $nC = 900$, $kR = 8$, $kC = 6$, *and 7 planted co-communities; entries in the adjacency matrix equal to 1 (representing a network edge) are marked in black; planted co-communities are outlined in colour. (b)* $\theta_{\text{in}} = 40$, *within-community edge density* $\rho_{\text{in}} = 0.6$*; (b)* $\theta_{\text{in}} = 30$, $\rho_{\text{in}} = 0.34$*; (c)* $\theta_{\text{in}} = 20$, $\rho_{\text{in}} = 0.15$*; (d)* $\theta_{\text{in}} = 10$, $\rho_{\text{in}} = 0.039$. *For all networks,* $\theta_{\text{out}} = 1$, *outside/between co-community edge density* $\rho_{\text{out}} = 0.0013$

measure of DNA methylation variance which is strongly associated with disease (as shown in chapter 2). I denote the gene expression variables $X(i)$, $i = 1, ..., m$ and the DNAm variables $Y(j)$, $j = 1, ..., l$; i.e., $X(i)$ and $Y(j)$ refer to the measurements for particular genes of gene expression and DNA methylation IGV respectively. I use Spearman correlation as a measure of association between the DNAm and gene expression variables, such that there is one correlation statistic for each pairing of $X(i)$ with $Y(j)$, for $i = 1, ..., m$ and $j = 1, ..., l$, leading to an $m \times l$ correlation matrix. Using the methodology presented in chapter 4, I then infer an $m \times l$ binary adjacency matrix from this correlation matrix such that $A_{ij} \in \{0, 1\}$ for all i, j, which

is equivalent to setting $A_{ij} = 1$ if variables $X(i)$ and $Y(i)$ are significantly correlated, and $A_{ij} = 0$ otherwise. When $A_{ij} = 1$, I also record the directionality information, i.e., whether the association between $X(i)$ and $Y(j)$ corresponds to a positive correlation (activation) or a negative correlation (inhibition).

I carried out co-community detection on this data set according to the methods described above (data source: The Cancer Genome Atlas (Hampton, 2006), breast cancer invasive carcinoma data set, basal tumour samples only). Figure 5.3(a) shows the adjacency matrix after carrying out co-community detection, ordering the X and Y-node groupings by row and column co-modularity (equations 5.22 and 5.23). Figure 5.3(b) (inlay) shows the same adjacency matrix ordered along its margins alphabetically by gene name, i.e., without ordering the margins using co-community detection. Hence, Figure 5.3(b) shows a baseline in which the nodes are essentially randomly ordered, against which to compare the adjacency matrix after co-community detection, and ordering based upon it. The co-community structure is clearly revealed in Figure 5.3(a), whereas no co-community structure is visible in Figure 5.3(b). I define a co-community $\hat{c} \in \hat{C}$ as a combination of X-node grouping $\hat{g}^{(X)} \in \hat{G}^{(X)}$ with Y-node grouping $\hat{g}^{(Y)} \in \hat{G}^{(Y)}$ which is significant according to a z-test with zero mean and variance calculated as in equation 5.21, with significance defined by FDR-corrected p-value < 0.05. The numbers of X and Y-node groupings, $k^{(X)}$ and $k^{(Y)}$, are estimated according to equations 5.12 and 5.13 as 89 and 67 respectively, leading to 5963 potential co-communities, of which $\hat{T} = 2018$ are found to be significant. I tested these 2018 significant co-communities for domain relevance, by testing the overlap of the genes (nodes) of each co-community, separately with each of 10295 known gene-groups (data source: *http://www.broadinstitute.org/gsea/msigdb/*). This type of analysis is often called 'gene set enrichment analysis' (GSEA) (Subramanian *et al.* , 2005). I found that 1340 (66%) overlap significantly (Fisher's exact test, FDR-adjusted $p < 0.05$) with these known gene-sets (including many gene-sets related to cancer biology, stem-cell biology and cellular proliferation), confirming the domain relevance of this result, as well as indicating novel findings which could be investigated further by experimental biologists.

5.6 Conclusion

I have introduced the notion of co-modularity. I have shown how it can be used to perform co-community detection in bipartite networks, and how it fits with the notion of the stochastic co-blockmodel. I have shown how co-modularity can be used to compare co-communities, to calculate their strength and significance, to arrange them for visualisation, and to calculate an algorithmic objective function for optimisation. I have introduced the anisotropic graphon

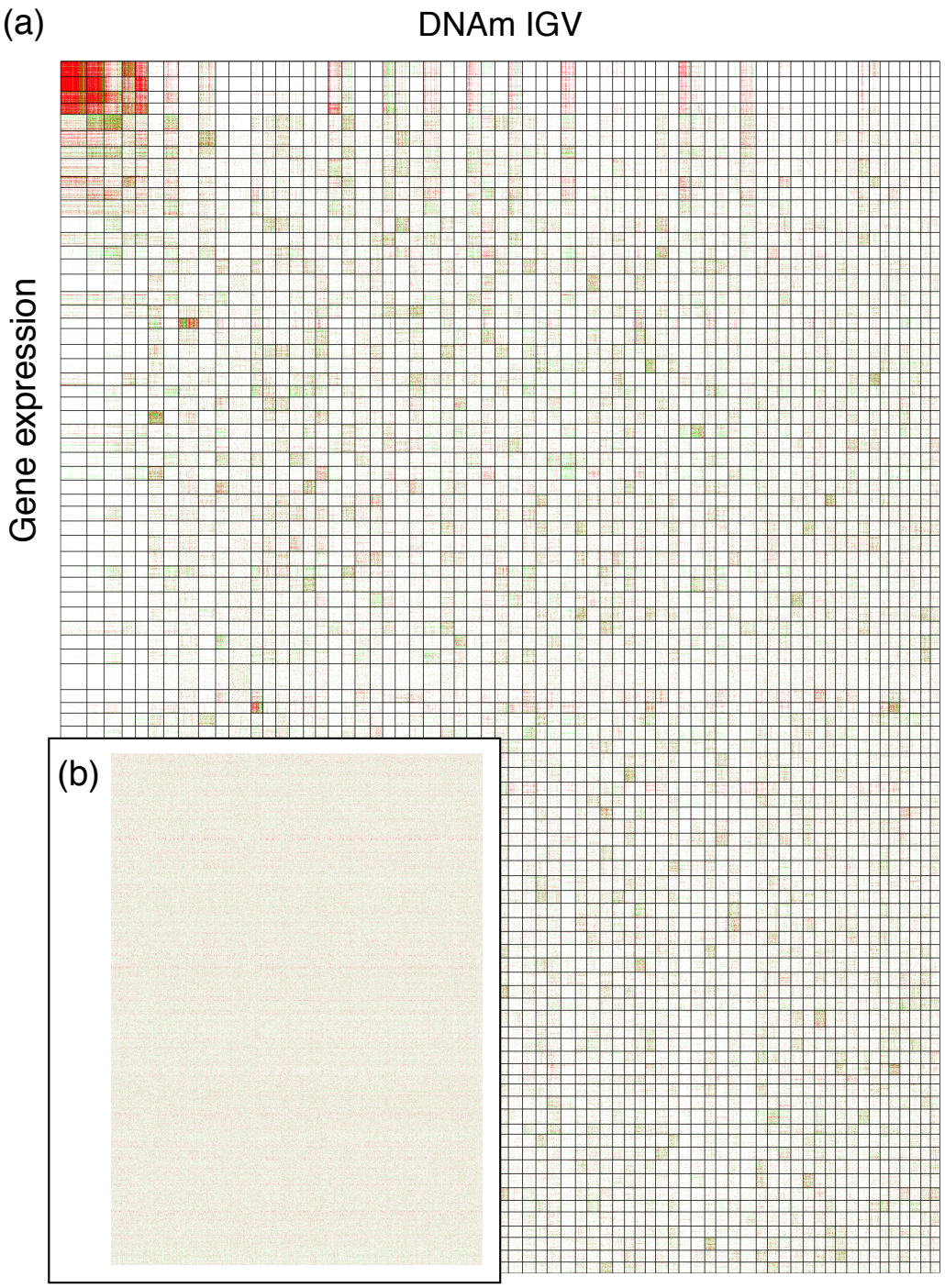


Figure 5.3: *Co-communities in the linked DNA methylation and gene expression data.*
(a) Genes are ordered along the margins of the adjacency matrix, according to co-communities detected by the methods presented here. Partitions between detected co-communities are shown with black lines. (b) The same adjacency matrix ordered along its margins alphabetically by gene name, i.e., without ordering the margins using co-community detection. Entries in the adjacency matrix equal to 1 (representing a network edge) are coloured, with green and red indicating positive and negative associations, respectively.

class, and have shown how to use it to estimate the optimum number of groups into which to divide the two types of nodes. I have also shown how this estimation can be simplified in certain circumstances for a more parsimonious scheme, and how to test whether this simplification is

justified. I have addressed practical points about the implementation of the methodology, and have demonstrated its utility with a simulation study and application to linked DNA methylation and gene expression data.

An interesting extension to this methodology would be to consider overlapping blocks in the stochastic co-blockmodel, a problem which has already been successfully addressed in the context of the stochastic blockmodel for unipartite networks (Latouche *et al.* , 2011), and in co-clustering without fitting the stochastic blockmodel (Madeira & Oliveira, 2004). Another interesting application would be to develop an online version of the method as a computationally efficient approach to large and growing data-sets (Zanghi *et al.* , 2010). This methodology could also be expected to work in even more general settings of bi-clustering or co-clustering, in which the variables being clustered together are simply correlated, rather than having any tangible interactive behaviour in the real world. These methods are based on commonly available computationally efficient methods for large sparse matrices, and perform well on large datasets, with large numbers of co-communities, often performing better than methods based on model likelihoods.

5.7 Derivations

Derivation A: Derivation relating to Algorithm 1

Define m, l, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{d}^{(X)}$, $\mathbf{d}^{(Y)}$, d^{++}, $g^{(X)}$, $g^{(Y)}$, $k^{(X)}$, $k^{(Y)}$, Ψ and Q_{XY} according to Definitions 3 - 8. Specify that $k^{(X)} = k^{(Y)} = 2$, that $T = 2$, that $c_1 = \{1, 1\}$, and that $c_2 = \{2, 2\}$; i.e., that there are two co-communities, the first of which consists of $g_1^{(X)}$ paired with $g_1^{(Y)}$, and the second of which consists of $g_2^{(X)}$ paired with $g_2^{(Y)}$. Define co-community label vectors $\mathbf{s}$ and $\mathbf{r}$ for the X and Y-nodes respectively, such that:

$$s_i = \begin{cases} 1, & \text{if } X\text{-node } i \text{ is in co-community 1,} \\ -1, & \text{if } X\text{-node } i \text{ is in co-community 2,} \end{cases} \tag{5.25}$$

and

$$r_j = \begin{cases} 1, & \text{if } Y\text{-node } j \text{ is in co-community 1,} \\ -1, & \text{if } Y\text{-node } j \text{ is in co-community 2.} \end{cases} \tag{5.26}$$

Hence:

$$\Psi\left(C; G^{(X)}, G^{(Y)}; i, j\right) = \frac{1}{2}\left(s_i r_j + 1\right),$$

and

$$Q_{XY} = \frac{1}{2d^{++}} \sum_{i=1}^{m} \sum_{j=1}^{l} B_{ij} \left(s_i r_j + 1 \right).$$

Note that the rows of $\mathbf{B}$ sum to zero:

$$\sum_{j=1}^{l} B_{ij} = \sum_{j=1}^{l} A_{ij} - \frac{d_i^{(X)}}{d^{++}} \sum_{j=1}^{l} d_j^{(Y)} = d_i^{(X)} - \frac{d_i^{(X)}}{d^{++}} \cdot d^{++} = 0.$$

Also, the columns of $\mathbf{B}$ also sum to zero, by a similar argument. Hence:

$$Q_{XY} = \frac{1}{2d^{++}} \sum_{i=1}^{m} \sum_{j=1}^{l} B_{ij} s_i r_j. \tag{5.27}$$

When (Newman, 2013) derives the properties of unipartite network community detection he relaxes the constraint that the co-community labels take the values of ± 1, to be able to arrive at an algorithmic solution. Nodes are then assigned to one community or the other, according to their sign (in the two-community scenario). A similar relaxation is made here, allowing $s_i \in \mathbb{R}$ and $r_j \in \mathbb{R}$, subject also to the following elliptical constraints, which allow for degree heterogeneity as in the degree corrected stochastic blockmodel:

$$\sum_{i=1}^{m} d_i^{(X)} s_i^2 = d^{++}, \tag{5.28}$$

$$\sum_{j=1}^{l} d_j^{(Y)} r_j^2 = d^{++}. \tag{5.29}$$

In the extreme scenario, in which $s_i \in \{-1, 1\}$ and $r_j \in \{-1, 1\}$, these constraints are equivalent to $d^{++} = \sum_{i=1}^{m} d_i^{(X)} = \sum_{j=1}^{l} d_j^{(Y)}$ (i.e., as per definition 6). This relaxation is equivalent to saying that nodes may be partly in one group, and partly in another group. N.B., ultimately each node will be assigned entirely to only the group it is most strongly associated with (according to s_i or r_j), and hence mixed membership does not occur in the final assignment of nodes to groups. For homogenous degree distributions, the constraints of equation 5.28 and 5.29 prevent the co-modularity from becoming arbitrarily large, as nodes are assigned many times over to many groups. For heterogenous degree distributions, the effect of the constraint is equivalent, except that the constraint is weighted to give importance to high-degree nodes. This is achieved by the constraints of equation 5.28 and 5.29 restricting the weighted sum of the degrees (weighted by the assignment of nodes to groups) to be the equal to the total number of edges.

We wish to find the community assignment vectors $\mathbf{r}$ and $\mathbf{s}$ which maximise the co-

modularity, i.e., we want to maximise Q_{XY} with respect to both $\mathbf{r}$ and $\mathbf{s}$. To do this, I employ the Lagrange multipliers λ and μ, and equate the derivatives to zero, N.B., the partial derivatives with respect to $s_{i'}$ and $r_{j'}$ are used as the derivatives are taken with respect to these individual $i' \in \{1, ..., l\}$, and $j' \in \{1, ..., m\}$.

$$\frac{\partial}{\partial s_{i'}}\left[\sum_{i=1}^{m}\sum_{j=1}^{l} B_{ij}s_i r_j - \lambda\sum_{i=1}^{m} d_i^{(X)}s_i^2 - \mu\sum_{j=1}^{l} d_j^{(Y)}r_j^2\right] = 0,$$

$$\text{and} \quad \frac{\partial}{\partial r_{j'}}\left[\sum_{i=1}^{m}\sum_{j=1}^{l} B_{ij}s_i r_j - \lambda\sum_{i=1}^{m} d_i^{(X)}s_i^2 - \mu\sum_{j=1}^{l} d_j^{(Y)}r_j^2\right] = 0,$$

$$\implies \quad \sum_{j=1}^{l} B_{ij}r_j - 2\lambda d_i^{(X)}s_i = 0, \tag{5.30}$$

$$\text{and} \quad \sum_{i=1}^{m} B_{ij}s_i - 2\mu d_j^{(Y)}r_j = 0. \tag{5.31}$$

Hence, taking $\mathbf{D}^{(X)}$ and $\mathbf{D}^{(Y)}$ as the diagonal matrices with the degree vectors $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ respectively on their leading diagonals,

$$\mathbf{B}\mathbf{r} = 2\lambda\mathbf{D}^{(x)}\mathbf{s} \tag{5.32}$$

$$\text{and} \quad \mathbf{B}^\top\mathbf{s} = 2\mu\mathbf{D}^{(y)}\mathbf{r}. \tag{5.33}$$

Substituting for $\mathbf{s}$, equation 5.33 in 5.32, gives:

$$\left(\mathbf{D}^{(y)}\right)^{-1}\mathbf{B}^\top\left(\mathbf{D}^{(x)}\right)^{-1}\mathbf{B}\mathbf{r} = 4\lambda\mu\mathbf{r}, \tag{5.34}$$

$$\implies \left(\mathbf{D}^{(y)}\right)^{-1/2}\mathbf{B}^\top\left(\mathbf{D}^{(x)}\right)^{-1/2}\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}\mathbf{r} = 4\lambda\mu\mathbf{r},$$

$$\implies \left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)^\top\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)\mathbf{r} = 4\lambda\mu\mathbf{r}, \tag{5.35}$$

$$\implies \mathbf{M}^\top\mathbf{M}\mathbf{r} = 4\lambda\mu\mathbf{r}, \tag{5.36}$$

where

$$\mathbf{M} = \left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{B}\left(\mathbf{D}^{(y)}\right)^{-1/2}.$$

By an identical argument, substituting 5.32 in 5.33 and re-arranging equivalently,

$$\mathbf{M}\mathbf{M}^\top\mathbf{s} = 4\lambda\mu\mathbf{s}. \tag{5.37}$$

Hence, $\mathbf{s}$ and $\mathbf{r}$ are eigenvectors of $\mathbf{M}\mathbf{M}^\top$ and $\mathbf{M}^\top\mathbf{M}$ respectively, with $4\lambda\mu$ the corresponding eigenvalue in both cases. Therefore, $\mathbf{s}$ and $\mathbf{r}$ are left and right singular vectors respectively of:

$$\mathbf{M} = \left(\mathbf{D}^{(x)}\right)^{-1/2} \mathbf{B} \left(\mathbf{D}^{(y)}\right)^{-1/2},$$

with corresponding singular value $2\sqrt{\lambda\mu}$.

Multiplying equation 5.30 by $s_i/2d^{++}$, summing over i and referring to equation 5.28 gives:

$$\frac{1}{2d^{++}} \sum_{i=1}^{m}\sum_{j=1}^{l} B_{ij}s_i r_j = \frac{2\lambda}{2d^{++}} \sum_{i=1}^{m} d_i^{(X)} s_i^2 = \frac{2\lambda \cdot d^{++}}{2d^{++}} = \lambda,$$

hence referring to equation 5.27, we get:

$$Q_{XY} = \lambda. \tag{5.38}$$

Then equivalently multiplying equation 5.31 by $r_j/2d^{++}$, summing over j and referring to equation 5.29, and then referring to equation 5.27 gives:

$$Q_{XY} = \mu. \tag{5.39}$$

Therefore, referring again to equations 5.36 and 5.37, the maximum modularity solution is for the left and right singular vectors of M which correspond to the greatest singular value 2λ.

Now substituting equation 5.15 in equation 5.33, we get:

$$\mathbf{s}^\top \left(\mathbf{A} - \frac{1}{d^{++}}\mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)}\right)^\top\right) = 2\mu\mathbf{r}^\top\mathbf{D}^{(y)},$$

$$\implies \mathbf{s}^\top\mathbf{A} = \frac{1}{d^{++}}\mathbf{s}^\top\mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)}\right)^\top + 2\mu\mathbf{r}^\top\mathbf{D}^{(y)}. \tag{5.40}$$

Post-multiplying equation 5.40 by $\mathbf{1} = (1, 1, 1...)$ leads to:

$$\mathbf{s}^\top\mathbf{d}^{(X)} = \frac{1}{d^{++}}\mathbf{s}^\top\mathbf{d}^{(X)} \cdot d^{++} + 2\mu\mathbf{r}^\top\mathbf{d}^{(Y)}$$

$$\therefore \quad \mu\mathbf{r}^\top\mathbf{d}^{(Y)} = 0.$$

Assuming that there is co-community structure present in $\mathbf{A}$, there must be positive co-modularity, i.e., $Q_{XY} > 0 \implies \mu > 0$ (referring back to equation 5.39), and therefore $\mathbf{r}^\top\mathbf{d}^{(Y)} = 0$. By an identical argument, also $\mathbf{s}^\top\mathbf{d}^{(X)} = 0$. Therefore, for eigenvectors $\mathbf{r}$

corresponding to $Q_{XY} > 0$,

$$\mathbf{Br} = \left(\mathbf{A} - \frac{1}{d^{++}}\mathbf{d}^{(X)}\left(\mathbf{d}^{(Y)}\right)^{\top}\right)\mathbf{r} = \mathbf{Ar}$$

and so to find these eigenvectors with Q_{XY} maximised, instead of equation 5.35 we can consider

$$\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{A}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)^{\top}\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{A}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)\mathbf{r} = (2\lambda)^2\,\mathbf{r} \quad (5.41)$$

which, referring back to equation 5.7, can be written in terms of the co-Laplacian $\mathbf{L}_{XY}$ as:

$$\mathbf{L}_{XY}^{\top}\mathbf{L}_{XY}\mathbf{r} = (2\lambda)^2\,\mathbf{r}.$$

By identical argument, we can also write:

$$\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{A}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{A}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)^{\top}\mathbf{s} = (2\lambda)^2\,\mathbf{s} \quad (5.42)$$

and

$$\mathbf{L}_{XY}\mathbf{L}_{XY}^{\top}\mathbf{s} = (2\lambda)^2\,\mathbf{s}.$$

Hence, the co-Laplacian $\mathbf{L}_{XY}$ has left and right singular vectors $\mathbf{s}$ and $\mathbf{r}$ respectively, with corresponding singular values 2λ. It can be seen that equation 5.41 has the eigenvector $\mathbf{1} = (1, 1, 1, ...)$, as follows:

$$\begin{aligned}
\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{A}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)^{\top}\left(\left(\mathbf{D}^{(x)}\right)^{-1/2}\mathbf{A}\left(\mathbf{D}^{(y)}\right)^{-1/2}\right)\mathbf{1} &= (2\lambda)^2\,\mathbf{1} \\
\implies \left(\mathbf{D}^{(y)}\right)^{-1}\mathbf{A}^{\top}\left(\mathbf{D}^{(x)}\right)^{-1}\mathbf{A}\mathbf{1} &= (2\lambda)^2\,\mathbf{1} \\
\implies \left(\mathbf{D}^{(y)}\right)^{-1}\mathbf{A}^{\top}\left(\mathbf{D}^{(x)}\right)^{-1}\mathbf{d}^{(X)} &= (2\lambda)^2\,\mathbf{1} \\
\implies \left(\mathbf{D}^{(y)}\right)^{-1}\mathbf{A}^{\top}\mathbf{1} &= (2\lambda)^2\,\mathbf{1} \\
\implies \left(\mathbf{D}^{(y)}\right)^{-1}\mathbf{d}^{(Y)} &= (2\lambda)^2\,\mathbf{1} \\
\mathbf{1} &= (2\lambda)^2\,\mathbf{1}
\end{aligned}$$

and hence the corresponding eigenvalue is $(2\lambda)^2 = 1$, which by the Perron-Frobenius theorem, must be the greatest eigenvalue (Hom & Johnson, 1991; Newman, 2013). An identical argument can also be applied to $\mathbf{s}$ in equation 5.42. This means that the greatest singular value $2\lambda = 1$ corresponds to these left and right singular vectors which are both $\mathbf{1}$ (of lengths m and l respectively), however such singular vectors do not satisfy $\mathbf{r}^\top \mathbf{d}^{(Y)} = 0$ and $\mathbf{s}^\top \mathbf{d}^{(X)} = 0$. Therefore, to maximise the co-modularity in the case of two co-communities, we should divide the X and Y-nodes according to the left and right singular vectors respectively which correspond to the second greatest singular value.

The above explains how Algorithm 1 works for the case of two co-communities. An equivalent extension to k communities has been made In the unipartite community detection setting (Riolo & Newman, 2012). To do so, the community labels are identified with the vertices of $k-1$ simplicies, i.e., for detection of 3 communities, the co-community labels would be the vertices of a triangle. Relaxing constraints equivalent to equations 5.28 and 5.29 means allowing the nodes to move away from the vertices of the simplex. This amounts to clustering the nodes in the space of the eigenvectors corresponding to the 2nd to k^{th} greatest eigenvalues of the Laplacian $\mathbf{L}$. This clustering is conventionally done using k-means. The reader is referred to (Riolo & Newman, 2012) for the detailed technical derivations relating to this. A similar extension can naturally be made in this co-community detection setting. To detect $k^{(X)}$ X-node groupings, and $k^{(Y)}$ Y-node groupings, the X and Y-nodes can be separately clustered (using k-means independently for the X and Y-nodes) in the spaces of the left and right singular vectors (respectively) corresponding to the 2nd to $k^{(X)\text{th}}$ and 2nd to $k^{(Y)\text{th}}$ greatest singular values, respectively, of the singular value decomposition of the co-Laplacian $\mathbf{L}_{XY}$.

Derivation B: Proof of Proposition 1

For the case of two co-communities, with θ_{in} and θ_{out} defined according to equation 5.2, with the co-community labels r_i and s_j defined as in Derivation A / section 5.7 (equations 5.25 and 5.26), and with $G^{(X)}$ and $G^{(Y)}$ defined according to Definition 3, I note (equivalently to (Newman, 2013)) that:

$$\theta_{z^{(X)}(i),z^{(Y)}(j)} = \frac{1}{2}\left(\theta_{\text{in}} + \theta_{\text{out}} + r_i s_j \left(\theta_{\text{in}} - \theta_{\text{out}}\right)\right), \tag{5.43}$$

$$\text{and} \quad \ln\left(\theta_{z^{(X)}(i),z^{(Y)}(j)}\right) = \frac{1}{2}\left(\ln\left(\theta_{\text{in}}\theta_{\text{out}}\right) + r_i s_j \ln\left(\frac{\theta_{\text{in}}}{\theta_{\text{out}}}\right)\right), \tag{5.44}$$

N.B., equations 5.43 and 5.44 only hold because $s_i \in \{-1, 1\}$ and $r_j \in \{-1, 1\}$. Substituting equations 5.43 and 5.44 into equation 5.3, and estimating the node-specific connectivity param-

eters $\boldsymbol{\pi}^{(X)}$ and $\boldsymbol{\pi}^{(Y)}$ by the degree distributions $\mathbf{d}^{(X)}$ and $\mathbf{d}^{(Y)}$ leads to the profile likelihood:

$$\ell\left(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}\right) = \sum_{i=1}^{m}\sum_{j=1}^{l}\left[\frac{A_{ij}}{2}\left(\ln\left(\theta_{\text{in}}\theta_{\text{out}}\right) + r_i s_j \ln\left(\frac{\theta_{\text{in}}}{\theta_{\text{out}}}\right)\right)\right.$$
$$\left. - \frac{d_i^{(X)} d_j^{(Y)}}{2}\left(\theta_{\text{in}} + \theta_{\text{out}} + r_i s_j\left(\theta_{\text{in}} - \theta_{\text{out}}\right)\right)\right]$$

$$\implies \quad \ell\left(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}\right) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{l}\left[A_{ij}\ln\left(\theta_{\text{in}}\theta_{\text{out}}\right) - d_i^{(X)} d_j^{(Y)}\left(\theta_{\text{in}} + \theta_{\text{out}}\right)\right.$$
$$\left. + \ln\left(\frac{\theta_{\text{in}}}{\theta_{\text{out}}}\right)\left(A_{ij} - d_i^{(X)} d_j^{(Y)} \cdot \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\ln\theta_{\text{in}} - \ln\theta_{\text{out}}}\right) s_i r_j\right].$$

We seek to maximise $\ell\left(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}\right)$ with respect to $G^{(X)}$ and $G^{(Y)}$ by choosing the co-community labels s_i and r_j. Therefore, we can drop the terms constant in s_i and r_j to give:

$$\widetilde{\ell}\left(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}\right) = \sum_{i=1}^{m}\sum_{j=1}^{l}\left(A_{ij} - d_i^{(X)} d_j^{(Y)} \cdot \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\ln\theta_{\text{in}} - \ln\theta_{\text{out}}}\right) s_i r_j,$$

and defining:

$$\eta = \frac{\theta_{\text{in}} - \theta_{\text{out}}}{\ln\theta_{\text{in}} - \ln\theta_{\text{out}}},$$

we therefore have:

$$\widetilde{\ell}\left(\boldsymbol{\theta}; \mathbf{d}^{(X)}, \mathbf{d}^{(Y)}; G^{(X)}, G^{(Y)}\right) = \sum_{i=1}^{m}\sum_{j=1}^{l}\left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)}\right) s_i r_j, \tag{5.45}$$

which I note as equivalent to equation 22 in (Newman, 2013). Proceeding similarly to that work, by applying to equation 5.45 the constraints of equations 5.28 and 5.29 with Lagrange multipliers λ and μ and differentiating and equating to zero, we get:

$$\frac{\partial}{\partial s_{i'}}\left[\sum_{i=1}^{m}\sum_{j=1}^{l}\left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)}\right) s_i r_j - \lambda\sum_{i=1}^{m} d_i^{(X)} s_i^2 - \mu\sum_{j=1}^{l} d_j^{(Y)} r_j^2\right] = 0,$$
$$\frac{\partial}{\partial r_{j'}}\left[\sum_{i=1}^{m}\sum_{j=1}^{l}\left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)}\right) s_i r_j - \lambda\sum_{i=1}^{m} d_i^{(X)} s_i^2 - \mu\sum_{j=1}^{l} d_j^{(Y)} r_j^2\right] = 0,$$

$$\implies \quad \sum_{j=1}^{l} \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) r_j - 2\lambda d_i^{(X)} s_i = 0,$$

$$\text{and} \quad \sum_{i=1}^{m} \left(A_{ij} - \eta d_i^{(X)} d_j^{(Y)} \right) s_i - 2\mu d_j^{(Y)} r_j = 0,$$

$$\therefore \quad \left(\mathbf{A} - \eta \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top \right) \mathbf{r} = 2\lambda \mathbf{D}^{(X)} \mathbf{s}, \tag{5.46}$$

$$\text{and} \quad \left(\mathbf{A}^\top - \eta \mathbf{d}^{(Y)} \left(\mathbf{d}^{(X)} \right)^\top \right) \mathbf{s} = 2\mu \mathbf{D}^{(Y)} \mathbf{r}. \tag{5.47}$$

Combining equations 5.46 and 5.47 by substituting for s and r, and following simplification identical to equations 5.34 to 5.35, gives:

$$\mathbf{W}^\top \mathbf{W} \mathbf{r} = 4\lambda\mu \mathbf{r},$$

$$\text{and} \quad \mathbf{W}\mathbf{W}^\top \mathbf{s} = 4\lambda\mu \mathbf{s},$$

where

$$\mathbf{W} = \left(\mathbf{D}^{(X)} \right)^{-1/2} \left(\mathbf{A} - \eta \mathbf{d}^{(X)} \left(\mathbf{d}^{(Y)} \right)^\top \right) \left(\mathbf{D}^{(Y)} \right)^{-1/2}.$$

Hence $\mathbf{s}$ and $\mathbf{r}$ are left and right singular vectors of the singular value decomposition of $\mathbf{W}$, again with corresponding singular values $4\lambda\mu$. Pre-multiplying 5.46 and 5.47 by $\mathbf{1} = (1, 1, 1...)$ leads to:

$$\mathbf{r}^\top \mathbf{d}^{(Y)} \left(1 - d^{++}\eta \right) = 2\lambda \mathbf{s}^\top \mathbf{d}^{(X)}, \tag{5.48}$$

$$\text{and} \quad \mathbf{s}^\top \mathbf{d}^{(X)} \left(1 - d^{++}\eta \right) = 2\lambda \mathbf{r}^\top \mathbf{d}^{(Y)}. \tag{5.49}$$

Substituting for $\mathbf{s}^\top \mathbf{d}^{(X)}$ and $\mathbf{r}^\top \mathbf{d}^{(Y)}$ in equations 5.49 and 5.48 gives:

$$\mathbf{s}^\top \mathbf{d}^{(X)} \left[\left(1 - d^{++}\eta \right)^2 - 4\mu\lambda \right] = 0,$$

$$\text{and} \quad \mathbf{r}^\top \mathbf{d}^{(Y)} \left[\left(1 - d^{++}\eta \right)^2 - 4\mu\lambda \right] = 0,$$

and therefore because $\left(1 - d^{++}\eta \right)^2 - 4\mu\lambda$ is not guaranteed to be zero,

$$\mathbf{s}^\top \mathbf{d}^{(X)} = 0,$$

$$\text{and} \quad \mathbf{r}^\top \mathbf{d}^{(Y)} = 0.$$

Therefore, equations 5.46 and 5.47 reduce to:

$$\mathbf{A}\mathbf{r} = 2\lambda \mathbf{D}^{(x)}\mathbf{s}$$
$$\text{and} \quad \mathbf{A}^{\top}\mathbf{s} = 2\mu \mathbf{D}^{(y)}\mathbf{r},$$

and again combining these equations by substituting for $\mathbf{s}$ and $\mathbf{r}$ and following equivalent simplification to equations 5.34 to 5.35, we hence find that $\mathbf{s}$ and $\mathbf{r}$ are left and right singular vectors of the co-Laplacian (equation 5.7). Therefore, the choice of the co-community labels $\mathbf{s}$ and $\mathbf{r}$ which maximises the model likelihood specified in equation 5.3, subject also to the constraint of equation 5.2, is equivalent to the maximum co-modularity assignment obtained via Algorithm 1.

Derivation C: Proof of Lemma 1

Define $\mathbf{A}$, $k^{(X)}$, $k^{(Y)}$ according to Definition 3, define $\xi^{(X)}$ and $\xi^{(Y)}$ according to Definition 4, define f, $\widetilde{f}$ and γ according to Definition 7, and define ρ and $\widetilde{M}$ according to Lemma 1. Define bandwidths $h_p^{(X)} = \left|g_p^{(X)}\right|$ and $h_q^{(Y)} = \left|g_q^{(Y)}\right|$, where $|\cdot|$ represents cardinality, define $\omega\left(p,q\right)$ as the domain of integration over the block corresponding to the pairing of X-node grouping $g_p^{(X)}$ with Y-node grouping $g_q^{(Y)}$, and define $\bar{A}_{p,q}$ as the corresponding block average,

$$\bar{A}_{p,q} = \frac{\sum_{j \in g_q^{(Y)}} \sum_{i \in g_p^{(X)}} A_{ij}}{h_p^{(X)} \cdot h_q^{(Y)}}.$$

Then, the bias-variance decomposition of the MISE of the blockmodel approximation of the graphon function $\hat{f}$ can be written as (Olhede & Wolfe, 2014):

$$\begin{aligned}\text{MISE}\left(\hat{\mathrm{f}}\right) \leq \mathbb{E}\iint_{(0,1)^2}\left|f(x,y) - \hat{f}(x,y)\right|^2 dx\, dy = \\ \sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}}\iint_{\omega(p,q)}\left\{\left|f(x,y) - \frac{\mathbb{E}\left(\bar{A}_{p,q}\right)}{\rho}\right|^2 + \frac{\text{Var}\left(\bar{A}_{p,q}\right)}{\rho^2}\right\} dx\, dy. \quad (5.50)\end{aligned}$$

As well as specifying groupings of X and Y-nodes, $g_p^{(X)}$ and $g_q^{(Y)}$ imply mappings between the adjacency matrix margins and the graphon margins, and the domain of integration $\omega\left(p,q\right)$ is hence a contiguous region of the graphon, which maps to entries of the adjacency matrix which are not necessarily contiguous.

Modelling the equi-smooth graphon $\widetilde{f}$ as a linear stretch transformation of the anisotropic

graphon f, by anisotropy factor γ, means that we can write:

$$f(x,y) = \widetilde{f}\left(\gamma x, y/\gamma\right).$$

I define the graphon oracle (Wolfe & Olhede, 2013; Olhede & Wolfe, 2014) ordering of the X and Y-nodes according to $\xi^{(X)}$ and $\xi^{(Y)}$ respectively. These are unobservable latent random vectors, which map the locations of the X and Y nodes from the margins of the graphon to the margins of the adjacency matrix. I.e., $\xi_i^{(X)}$ and $\xi_j^{(Y)}$ provide the locations on the graphon margins which correspond to the X and Y-nodes i and j respectively, where i and j are the adjacency matrix indices of these nodes. I define $(i)^{-1}$ as a function which gives the rank of $\xi_i^{(X)}$, $1 \leq i \leq m$, and similarly $(j)^{-1}$ as a function which gives the rank of $\xi_j^{(Y)}$, $1 \leq j \leq l$. Therefore, $(i)^{-1}$ and $(j)^{-1}$ are functions which take the ordering along the adjacency matrix margins, and return the ordering along the graphon margins. Hence, the inverses of these functions, (i) and (j), take the ordering along the graphon margins, and return the corresponding ordering along the adjacency matrix margins. Adapting the proof of Lemma 3 from (Olhede & Wolfe, 2014) to the anisotropic graphon, by defining $i_m = i/(m+1)$ and $j_l = j/(l+1)$, and assuming that $\widetilde{f}$ is Lipschitz-continuous, gives:

$$\begin{aligned}\left|f\left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}\right) - f\left(i_m, j_l\right)\right| &= \left|\widetilde{f}\left(\gamma\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}/\gamma\right) - \widetilde{f}\left(\gamma i_m, j_l/\gamma\right)\right| \\ &\leq \widetilde{M}\left|\left(\gamma\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}/\gamma\right) - \left(\gamma i_m, j_l/\gamma\right)\right|.\end{aligned}$$

Writing the variances and applying Jensen's inequality as in (Olhede & Wolfe, 2014) we get,

$$\begin{aligned}&\mathrm{Var}\left(\xi_{(i)}^{(X)}\right) = \frac{i_m(1-i_m)}{m+2} \leq \frac{1/4}{m+2}, \\ &\mathrm{Var}\left(\xi_{(j)}^{(Y)}\right) = \frac{j_l(1-j_l)}{l+2} \leq \frac{1/4}{l+2}, \\ \Longrightarrow\ &\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left\{\gamma^2\left(\xi_{(i)}^{(X)} - i_m\right)^2 + \frac{1}{\gamma^2}\left(\xi_{(j)}^{(Y)} - j_l\right)^2\right\}^{\frac{1}{2}} \\ &\quad\leq \left(\gamma^2\mathrm{Var}\left(\xi_{(i)}^{(X)}\right) + \frac{1}{\gamma^2}\mathrm{Var}\left(\xi_{(j)}^{(Y)}\right)\right)^{\frac{1}{2}} \\ &\quad\leq \left\{\gamma^2\cdot\frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}, \\ \therefore\ &\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left|f\left(\xi_{(i)}^{(X)}, \xi_{(j)}^{(Y)}\right) - f\left(i_m, j_l\right)\right| \leq \widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}.\end{aligned} \tag{5.51}$$

Now adapting Lemma 2 from (Olhede & Wolfe, 2014), I apply the law of iterated expectations to $A_{(i)(j)}$, to obtain:

$$\mathbb{E}\left(A_{(i)(j)}\right) = \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\mathbb{E}_{A|\xi^{(X)},\xi^{(Y)}}\left(A_{(i)(j)}\middle|\xi^{(X)},\xi^{(Y)}\right)\right] = \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right], \tag{5.52}$$

then using Jensen's inequality we get:

$$\left|\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right] - \rho f\left(i_m,j_l\right)\right| \leq \rho\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\left|f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right) - f\left(i_m,j_l\right)\right|\right], \tag{5.53}$$

and hence combining equations 5.51-5.53, we have:

$$\left|\mathbb{E}\left(A_{(i)(j)}\right) - \rho f\left(i_m,j_l\right)\right| \leq \rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}. \tag{5.54}$$

Now applying the law of total variance to $A_{(i)(j)}$, as in Lemma 2 from (Olhede & Wolfe, 2014), we obtain:

$$\begin{aligned}
\text{Var}\left(A_{(i)(j)}\right) =& \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\text{Var}_{A|\xi^{(X)},\xi^{(Y)}}\left(A_{(i)(j)}\middle|\xi^{(X)},\xi^{(Y)}\right)\right] \\
&+ \text{Var}_{\xi^{(X)},\xi^{(Y)}}\left[\mathbb{E}_{A|\xi^{(X)},\xi^{(Y)}}\left(A_{(i)(j)}\middle|\xi^{(X)},\xi^{(Y)}\right)\right] \\
=& \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\left(1-\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right)\right] \\
&+ \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho^2\left(f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right)^2\right] - \left(\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right]\right)^2 \\
=& \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right] - \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho^2\left(f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right)^2\right] \\
&+ \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho^2\left(f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right)^2\right] - \left(\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right]\right)^2 \\
=& \mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right]\left\{\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[1-\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right]\right\}.
\end{aligned} \tag{5.55}$$

From equation 5.51, we get:

$$\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right] \leq \rho f\left(i_m,j_l\right) + \rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}} \tag{5.56}$$

and

$$-\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[\rho f\left(\xi^{(X)}_{(i)},\xi^{(Y)}_{(j)}\right)\right] \leq -\rho f\left(i_m,j_l\right) + \rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}, \tag{5.57}$$

and hence also

$$\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[1-\rho f\left(\xi_{(i)}^{(X)},\xi_{(j)}^{(Y)}\right)\right] \geq 1-\rho f\left(i_m,j_l\right)-\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}} \tag{5.58}$$

and

$$-\mathbb{E}_{\xi^{(X)},\xi^{(Y)}}\left[1-\rho f\left(\xi_{(i)}^{(X)},\xi_{(j)}^{(Y)}\right)\right] \geq -1+\rho f\left(i_m,j_l\right)-\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}. \tag{5.59}$$

Now combining equation 5.56 with the negative of equation 5.59 and applying equation equation 5.55 we get:

$$\begin{aligned}\text{Var}\left(A_{(i)(j)}\right) \leq &\left[\rho f\left(i_m,j_l\right)+\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right]\\ &\cdot\left[1-\rho f\left(i_m,j_l\right)+\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right]\end{aligned}$$

and hence:

$$\begin{aligned}\text{Var}\left(A_{(i)(j)}\right) \leq &\rho f\left(i_m,j_l\right)\left[1-\rho f\left(i_m,j_l\right)\right]\\ &+\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\left[1+\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right].\end{aligned} \tag{5.60}$$

Similarly combining the negative of equation 5.57 with equation 5.58 and applying equation equation 5.55 we get:

$$\begin{aligned}\text{Var}\left(A_{(i)(j)}\right) \geq &\left[\rho f\left(i_m,j_l\right)-\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right]\\ &\cdot\left[1-\rho f\left(i_m,j_l\right)-\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right],\end{aligned}$$

and hence:

$$\begin{aligned}\text{Var}\left(A_{(i)(j)}\right) \geq &\rho f\left(i_m,j_l\right)\left[1-\rho f\left(i_m,j_l\right)\right]\\ &-\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\left[1-\rho\widetilde{M}\left\{\gamma^2\cdot\frac{1}{4(m+2)}+\frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right],\end{aligned}$$

and therefore:

$$
\begin{aligned}
-\text{Var}\left(A_{(i)(j)}\right) &\leq -\rho f\left(i_m, j_l\right)\left[1-\rho f\left(i_m, j_l\right)\right] \\
&\quad + \rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\left[1-\rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right] \\
&\leq -\rho f\left(i_m, j_l\right)\left[1-\rho f\left(i_m, j_l\right)\right] \qquad (5.61) \\
&\quad + \rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\left[1+\rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right],
\end{aligned}
$$

and hence combining equations 5.60 and 5.61 we get:

$$
\begin{aligned}
&\left|\text{Var}\left(A_{(i)(j)}\right) - \rho f\left(i_m, j_l\right)\left[1-\rho f\left(i_m, j_l\right)\right]\right| \\
&\leq \rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}} \cdot \left[1+\rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}^{\frac{1}{2}}\right].
\end{aligned} \qquad (5.62)
$$

Now referring to equation 5.54 and comparing it to equation 6 of Supporting Information Section A in (Olhede & Wolfe, 2014), allows us to re-write the covariance expression in Lemma 2 of (Olhede & Wolfe, 2014) giving:

$$
\text{Cov}\left(A_{(i)(j)}, A_{(i')(j')}\right) \qquad \leq \qquad \rho^2\widetilde{M}^2\left\{\gamma^2 \cdot \frac{1}{4(m+2)} + \frac{1}{\gamma^2}\cdot\frac{1}{4(l+2)}\right\}, \quad (5.63)
$$

$i \neq i'$, $j \neq j'$. We can then use equations 5.54, 5.62 and 5.63 to adapt Proposition 1 from (Olhede & Wolfe, 2014), denoting the average values of f and f^2 over the block corresponding to the pairing of $g_p^{(X)}$ with $g_q^{(Y)}$ as $\bar{f}_{p,q}$ and $\bar{f^2}_{p,q}$ respectively,

$$
\bar{f}_{p,q} = \frac{1}{\left|\omega\left(p,q\right)\right|}\iint_{\omega(p,q)} f(x,y)dx\,dy \qquad (5.64)
$$

and

$$
\bar{f^2}_{p,q} = \frac{1}{\left|\omega\left(p,q\right)\right|}\iint_{\omega(p,q)} f^2(x,y)dx\,dy, \qquad (5.65)
$$

where

$$
\left|\omega\left(p,q\right)\right| = \frac{h_p^{(X)}}{m}\cdot\frac{h_q^{(Y)}}{l},
$$

to give:

$$
\left|\mathbb{E}\left(\bar{A}_{p,q}\right) - \rho\bar{f}_{p,q}\right| \leq \rho\widetilde{M}\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2}\cdot\frac{1}{4l}\right\}^{\frac{1}{2}}\left\{1+o\left(1\right)\right\} \qquad (5.66)
$$

and

$$\left|\mathrm{Var}\left(\bar{A}_{p,q}\right) - \frac{\rho\bar{f}_{p,q} - \rho^2\bar{f^2}_{p,q}}{h_p^{(X)} \cdot h_q^{(Y)}}\right|$$

$$\leq \frac{\rho\widetilde{M}}{h_p^{(X)} \cdot h_q^{(Y)}}\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}^{\frac{1}{2}}\{1 + o\left(1\right)\} + \rho^2\widetilde{M}^2\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}, \quad (5.67)$$

which is a conservative upper bound. Now substituting equation 5.67 back into equation 5.50, we get:

$$\begin{aligned}\mathrm{MISE}\left(\hat{\mathrm{f}}\right) \leq \sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}}\iint_{\omega(p,q)}\Big[\Big|\left\{f(x,y) - \bar{f}_{p,q}\right\} \\ + \left\{\bar{f}_{p,q} - \mathbb{E}\left(\bar{A}_{p,q}\right)/\rho\right\}\Big|^2 + \frac{\bar{f}_{p,q} - \rho\bar{f^2}_{p,q}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \\ + \frac{\widetilde{M}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}}\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}^{\frac{1}{2}}\{1 + o\left(1\right)\} + \widetilde{M}^2\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}\Big]\,dx\,dy,\end{aligned}$$

then substituting equation 5.66, integrating and rearranging, leads to:

$$\begin{aligned}\mathrm{MISE}\left(\hat{\mathrm{f}}\right) \leq \sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}}\Bigg[\iint_{\omega(p,q)}\left|f(x,y) - \bar{f}_{p,q}\right|^2 dx\,dy \\ + \Bigg(2\widetilde{M}^2\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}\{1 + o\left(1\right)\} + \frac{\bar{f}_{p,q} - \rho\bar{f^2}_{p,q}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \\ + \frac{\widetilde{M}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}}\left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}^{\frac{1}{2}}\{1 + o\left(1\right)\}\Bigg) \cdot \frac{h_p^{(X)}}{m} \cdot \frac{h_q^{(Y)}}{l}\Bigg]. \quad (5.68)\end{aligned}$$

Then, adapting the proof of Lemma 1 from (Olhede & Wolfe, 2014), we can write:

$$\begin{aligned}\left|\bar{f}_{p,q} - f(x,y)\right| &= \left|\frac{1}{\left|\omega\left(p,q\right)\right|}\iint_{\omega(p,q)} f(x',y')dx'\,dy' - f(x,y)\right| \\ &\leq \frac{1}{\left|\omega\left(p,q\right)\right|}\iint_{\omega(p,q)}\left|\widetilde{f}(\gamma x', y'/\gamma) - \widetilde{f}(\gamma x, y/\gamma)\right|dx'\,dy'.\end{aligned}$$

Assuming $\widetilde{f}$ is Lipschitz continuous, it therefore follows that:

$$\begin{aligned}\left|\bar{f}_{p,q} - f(x,y)\right| &\leq \frac{1}{\left|\omega\left(p,q\right)\right|}\iint_{\omega(p,q)}\widetilde{M}\left|(\gamma x', y'/\gamma) - (\gamma x, y/\gamma)\right|dx'\,dy' \\ &\leq \frac{1}{\left|\omega\left(p,q\right)\right|}\iint_{\omega(p,q)}\widetilde{M}\sqrt{\gamma^2 \cdot \frac{\left(h_p^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{\left(h_q^{(Y)}\right)^2}{l^2}}dx'\,dy'\end{aligned}$$

$$\implies \left|\bar{f}_{p,q} - f(x,y)\right| \leq \widetilde{M}\sqrt{\gamma^2 \cdot \frac{\left(h_p^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{\left(h_q^{(Y)}\right)^2}{l^2}}$$

and therefore

$$\frac{1}{|\omega\,(p,q)|}\iint_{\omega(p,q)} \left|\bar{f}_{p,q} - f(x,y)\right|^2 \leq \widetilde{M}^2 \left\{\gamma^2 \cdot \frac{\left(h_p^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{\left(h_q^{(Y)}\right)^2}{l^2}\right\},$$

and hence summing over all the blocks corresponding to all pairings of X-node groupings $g^{(X)} \in G^{(X)}$ with Y-node groupings $g^{(Y)} \in G^{(Y)}$, and assuming $h^{(X)}$ and $h^{(Y)}$ are both constants, we get:

$$\sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}} \iint_{\omega(p,q)} \left|\bar{f}_{p,q} - f(x,y)\right|^2 \leq \widetilde{M}^2 \left\{\gamma^2 \cdot \frac{\left(h^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{\left(h^{(Y)}\right)^2}{l^2}\right\}. \tag{5.69}$$

Recalling equation 5.64 and equation 5.10, i.e.,

$$\iint_{(0,1)^2} f(x,y)dx\,dy = 1,$$

and noting that:

$$\sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}} \frac{\bar{f}_{p,q} - \rho\bar{f}^2{}_{p,q}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} \leq \sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}} \frac{\bar{f}_{p,q}}{\rho \cdot h^{(X)} \cdot h^{(Y)}},$$

we can see that:

$$\begin{aligned}
\sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}} \frac{\bar{f}_{p,q} - \rho\bar{f}^2{}_{p,q}}{\rho \cdot h_p^{(X)} \cdot h_q^{(Y)}} &\leq \sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}} \frac{m \cdot l}{\rho \cdot \left(h^{(X)}\right)^2 \cdot \left(h^{(Y)}\right)^2} \cdot \frac{h^{(X)}}{m} \cdot \frac{h^{(Y)}}{l} \cdot \bar{f}_{p,q} \qquad (5.70)\\
&= \frac{m \cdot l}{\rho \cdot \left(h^{(X)}\right)^2 \cdot \left(h^{(Y)}\right)^2} \sum_{q=1}^{k^{(Y)}}\sum_{p=1}^{k^{(X)}} \iint_{\omega(p,q)} f(x,y)dx\,dy\\
&= \frac{m \cdot l}{\rho \cdot \left(h^{(X)}\right)^2 \cdot \left(h^{(Y)}\right)^2} \iint_{(0,1)^2} f(x,y)dx\,dy\\
&= \frac{m \cdot l}{\rho \cdot \left(h^{(X)}\right)^2 \cdot \left(h^{(Y)}\right)^2}.
\end{aligned}$$

Now substituting 5.69 and 5.70 into 5.68, and rearranging, we get:

$$\begin{aligned}\text{MISE}\left(\hat{\text{f}}\right) \leq \widetilde{M}^2 &\left\{\gamma^2 \cdot \frac{\left(h^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{\left(h^{(Y)}\right)^2}{l^2}\right\} \\ &+ 2\widetilde{M}^2 \left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\} \{1 + o(1)\} \\ &+ \frac{1}{\rho \cdot h^{(X)} \cdot\ h^{(Y)}} + \frac{\widetilde{M}}{\rho \cdot h^{(X)} \cdot h^{(Y)}} \left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\}^{\frac{1}{2}} \{1 + o(1)\}\end{aligned}$$

and hence:

$$\begin{aligned}\text{MISE}\left(\hat{\text{f}}\right) \leq \widetilde{M}^2 &\left\{\gamma^2 \cdot \frac{\left(h^{(X)}\right)^2}{m^2} + \frac{1}{\gamma^2} \cdot \frac{\left(h^{(Y)}\right)^2}{l^2}\right\} \\ &+ 2\widetilde{M}^2 \left\{\gamma^2 \cdot \frac{1}{4m} + \frac{1}{\gamma^2} \cdot \frac{1}{4l}\right\} \{1 + o(1)\} + \frac{1}{\rho \cdot h^{(X)} \cdot\ h^{(Y)}} \{1 + o(1)\}.\end{aligned}$$

Derivation D: Proof of Proposition 3

We wish to compare the null hypothesis:

$$H_0: \quad \gamma = 1$$

against the alternative hypothesis:

$$H_1: \quad \gamma \neq 1$$

where $\gamma^2 = M_Y/M_X$. In practice, we can only compare the ratio of the estimates of the gradients, which can be written as:

$$\frac{\hat{M}_Y}{\hat{M}_X} = \frac{\hat{b}_X \hat{p}_Y l}{\hat{p}_X \hat{b}_Y m} = \frac{(p_x + \epsilon_p^{(x)})(b_y + \epsilon_b^{(y)})}{(b_x + \epsilon_b^{(x)})(p_y + \epsilon_p^{(y)})} \cdot \frac{l}{m} = \frac{p_x b_y}{b_x p_y} \cdot \frac{(1 + \frac{\epsilon_p^{(x)}}{p_x})(1 + \frac{\epsilon_b^{(y)}}{b_y})}{(1 + \frac{\epsilon_b^{(x)}}{b_x})(1 + \frac{\epsilon_p^{(y)}}{p_y})} \cdot \frac{l}{m}.$$

Applying a first-order Maclaurin expansion, and neglecting products of errors, gives:

$$\hat{\gamma}^2 = \frac{\hat{M}_Y}{\hat{M}_X} = \frac{\hat{b}_X \hat{p}_Y}{\hat{p}_X \hat{b}_Y} \cdot \frac{m}{l} \approx \frac{p_x b_y}{b_x p_y} \cdot \left(1 + \frac{\epsilon_p^{(x)}}{p_x} + \frac{\epsilon_b^{(y)}}{b_y} + \frac{\epsilon_b^{(x)}}{b_x} + \frac{\epsilon_p^{(y)}}{p_y}\right) \cdot \frac{l}{m}$$

and hence, assuming b_x and b_y, p_x and p_y, b_x and p_y, p_x and b_y are independent, we can test against the null distribution

$$\hat{\gamma}^2 = \frac{\hat{b}_X \hat{p}_Y l}{\hat{p}_X \hat{b}_Y m} \sim \mathcal{N}\left(1, \hat{\tau}^2\right),$$

where

$$\hat{\tau}^2 = \frac{\widehat{\mathrm{Var}(\hat{b}_X)}}{\hat{b}_X} + \frac{\widehat{\mathrm{Var}(\hat{p}_Y)}}{\hat{p}_Y} + \frac{\widehat{\mathrm{Var}(\hat{p}_X)}}{\hat{p}_X} + \frac{\widehat{\mathrm{Var}(\hat{b}_Y)}}{\hat{b}_Y} + 2\frac{\widehat{\mathrm{Cov}(\hat{b}_X, \hat{p}_X)}}{\hat{b}_X \hat{p}_X} + 2\frac{\widehat{\mathrm{Cov}(\hat{b}_Y, \hat{p}_Y)}}{\hat{b}_Y \hat{p}_Y}, \quad (5.71)$$

where $\hat{b}_X$, $\hat{p}_X$, $\hat{b}_Y$, $\hat{p}_Y$ and their variances and covariances are estimated from the linear model fits as described.

Chapter 6

Intra-gene DNA Methylation Variability is a Technically and Clinically Independent Prognostic Marker in Women's Cancers

6.1 Introduction

In this chapter, I investigate further IGV, or intra-gene DNA methylation variability (Bartlett *et al.* , 2013), which was introduced in chapter 2, finding that it is independently prognostic, and does not require data normalisation. Using IGV, based on raw data, I derive a robust gene-panel prognostic signature for ovarian cancer (OC, $n = 221$), which validates in two independent data sets from Mayo Clinic ($n = 198$) and TCGA ($n = 358$), with significance of $p = 0.004$ in both sets. The OC prognostic signature gene-panel is comprised of four gene groups, which represent distinct biological processes. I show that the IGV of these gene groups is likely a surrogate readout for transcription factor (TF) binding/activity. Using the methodology of chapter 5 to analyse linked DNA methylation and gene expression data, I also find co-clusters which represent groups of genes with highly associated expression and IGV patterns, and provide a starting-point for further investigation into the mechanistic roles of the observed IGV patterns in disease. IGV is a self-calibrating measure of methylation variability which can be used to predict clinical outcome in patients individually, providing a surrogate read-out of hard-to-measure disease processes.

Ovarian cancer (OC) and endometrial cancer (EC) are the most common gynaecological cancers (Jemal *et al.* , 2011). Only one in three patients with advanced stage OC survive for five years after their initial diagnosis (Greenlee *et al.* , 2001). Very little is known about OC biology and how to manipulate this disease therapeutically. DNAm changes are important in cancer (Widschwendter *et al.* , 2006); the epigenome is an interface between the genome and the environment (Jirtle & Skinner, 2007; Feil & Fraga, 2012), and hence DNAm changes can

measure exposure to environmental risk factors of cancer. DNAm biomarkers which represent a surrogate for patterns of gene interaction have previously been associated with clinical outcome in a wide variety of cancers (Bartlett *et al.* , 2014), as well as specifically in women's cancers (Zhuang *et al.* , 2012).

Robust clinical multi-gene signature markers, which are not dependent on the technical variability of the system, are urgently required. IGV is more resistant to additive changes in methylation levels than conventionally used measures of DNAm. This is because IGV is calculated relative to the mean methylation level. Any overall additive changes in methylation level will have a similar influence on the methylation level at individual CpGs loci (which are aggregated together when calculating IGV) and the mean methylation level (Figure 6.1b, comparing the different coloured lines). Therefore, while individual methylation levels, and the mean methylation level, would record such an additive shift in methylation level, IGV would not. Hence, I hypothesise that IGV can be considered a 'self-calibrating' measure, which is much less dependent on DNAm data being properly normalised and batch corrected.

6.2 Results

6.2.1 Comparison of predictive robustness of per-gene methylation measures, in raw and normalised data

To assess the effectiveness and robustness of IGV compared to mean methylation levels, in raw and normalised data, I compared four per-gene methylation measures, based on mean methylation level and IGV (Figure 6.1). For each gene, I calculated mean methylation level and IGV, separately for the promoter (TSS200) and gene body regions, by using the Illumina Infinium HumanMethylation450 platform specifications of the CpGs in these regions for each gene. I considered different genomic regions separately, because methylation patterns vary greatly from one genomic region to another, and the effect of methylation level on gene regulation varies according to genomic region. The four measures I compared, are as follows:

- TSS200 mean methylation
- TSS200 IGV
- Gene body mean methylation
- Gene body IGV

In chapter 2, I found that the mean z-score was the most effective per-gene methylation measure, for discriminating cancerous from healthy tissue. However, the main ovarian cancer dataset analysed in this chapter consists only of samples of cancerous tissue, and calculation of

the mean z-score measure requires healthy samples on which to base the reference mean and standard-deviation methylation profiles. Therefore, it is not possible to consider the mean z-score measure in this chapter.

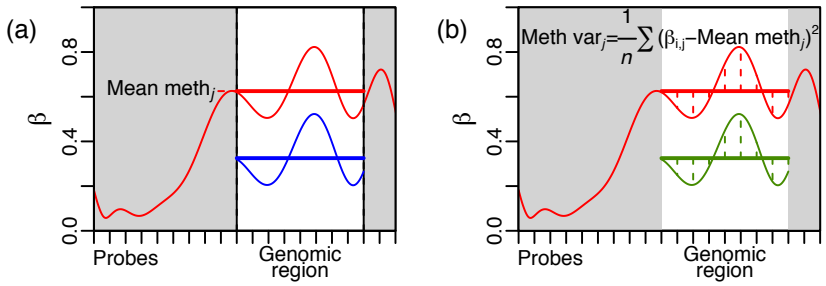
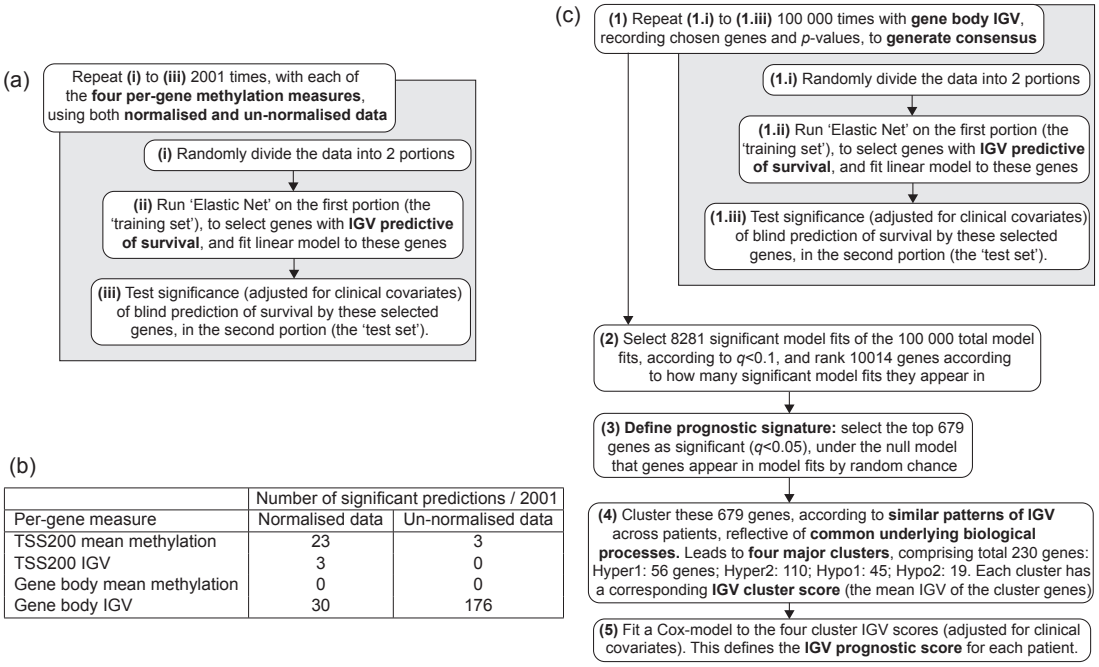


Figure 6.1: *Per-gene methylation measures.*
(a) The mean methylation level over a specific genomic region is calculated separately for the TSS200 (promoter) and gene body genomic regions. The blue curve indicates the new position of the red curve after an additive global shift in methylation level, which might be due to technological or other experimental factors, and the difference between the horizontal red and blue lines (mean levels) illustrates the effect of this shift on the mean methylation level. (b) The intra-gene methylation variability (IGV) is calculated from the variation around the mean methylation level, i.e., from the dashed vertical lines, and is similarly calculated separately for the TSS200 and gene body genomic regions. The vertical green lines are changed very little compared to the vertical red lines, illustrating that such a global additive shift in mean methylation level has much less effect on IGV, which is therefore referred to as a 'self-calibrating measure'.

I obtained genome-wide DNAm profiles, via the Illumina Infinium HumanMethylation450 platform, from 218 primary OC samples. For each of the four measures described, for both SWAN-normalised (Makismovic *et al.* , 2012) and raw (un-normalised) data, I used 'Elastic Net' (Zou & Hastie, 2005; Simon *et al.* , 2011) to find a prognostic selection of genes. Elastic net has been found to be an optimal linear modelling method to identify groups of genes which act together as part of a common biological process (Jojic *et al.* , 2013). It is a regression method which 'chooses' the set of genes which model the data best, trying to include as few genes in the model as possible, whilst ensuring that the model predicts the outcome of interest as accurately as possible. In doing so, it discards genes which do not provide useful information, or which provide repeated information.

I assessed the effectiveness of the per-gene methylation measures as prognostic measures, both using normalised and raw data, by randomly dividing the data into two portions: a 'training set', and a 'test set'. Elastic Net was used to select genes and fit a model to the training set, and the ability of this gene selection and model to blindly predict patient survival outcome (adjusted for clinical covariates) was assessed using the test-set. This was repeated 2001 times, and significantly predictive selected groups of genes were defined according to false discovery rate (FDR) adjusted (Benjamini & Hochberg, 1995) p-value (i.e., FDR q-value) < 0.1 (Figure 6.2a). As shown in Figure 6.2b, in normalised data, both promoter (TSS200) mean methylation

level and gene body IGV have some predictive ability, however for raw data, only gene body IGV predicts well - and far more so than any of the measures using normalised data. This indicates that IGV is more resilient to technical and systematic variation in recorded methylation levels (which usually necessitate normalisation). It also suggests that this normalisation actually diminishes the capacity of IGV to derive true biological meaning (in the form of ability to predict patient survival outcome).



	Number of significant predictions / 2001	
Per-gene measure	Normalised data	Un-normalised data
TSS200 mean methylation	23	3
TSS200 IGV	3	0
Gene body mean methylation	0	0
Gene body IGV	30	176

Figure 6.2: *Overview of methods.*
(a) Methodology overview for comparison of the four per-gene methylation measures, with normalised and raw data. (b) Results of this comparison. (c) Methodology overview for calculation of ovarian cancer IGV prognostic score.

6.2.2 Derivation of an ovarian cancer prognostic signature, and IGV prognostic score

I used IGV to derive an OC DNAm prognostic signature (Figure 6.2c), based on gene-body IGV (from here on simply referred to as 'IGV'), using only raw data. I did this by determining a consensus on a set of genes predictive of survival, by following the same procedure of splitting data into test and training sets, and then assessing the gene selection and fitted model for their ability to blindly predict patient survival outcome (adjusted for clinical covariates) in the test set. In order to ensure convergence to a stable result, I made 10^5 such partitions of the data, each resulting in a predictive selection of genes. Of these, 8281 were found as significant (FDR $q < 0.1$), and significance for each gene was then calculated based on the number of significant models in which that gene appeared. 679 genes were selected like this for inclusion in the OC prognostic signature at a significance level of FDR $q < 0.05$, with the least significant gene

present in 1057 out of 8281 model fits. The top 85 most significant of these genes appear in table 6.2.

Genes often act together as part of biological pathways, and processes. Hence, we can expect that these 679 OC prognostic signature genes can be represented by a smaller number of underlying biological processes, which are important to disease progression. Grouping genes with similar experimental measurements by using clustering methodology is well established as an effective approach for determining clinically relevant prognostic markers (Golub *et al.* , 1999; Valk *et al.* , 2004). Hence, to uncover such groupings in the 679 genes of the OC prognostic signature, I carried out consensus clustering (Monti *et al.* , 2003), to identify groups of genes with similar patterns of IGV across patients. Each cluster identified in this way reveals a different IGV trend, and therefore may correspond to a different underlying biological process, which gives rise to the pattern of IGV observed in that cluster. The clustering was carried out separately for genes which were individually associated with worse patient survival outcome for increased IGV ('hyper' genes) and for decreased IGV ('hypo' genes). The result was four clusters: two from the hyper genes, called clusters 'hyper 1' and 'hyper 2', and two from the hypo genes, called clusters 'hypo 1' and 'hypo 2'; they appear in 6.3 - 6.6. The mean IGV of the genes of each of the four clusters gives an IGV 'cluster score', for each cluster and for each patient, which are taken to be representative of the different IGV trends, and corresponding underlying biological processes, within the OC prognostic signature.

I then calculated an IGV prognostic score, by fitting a multivariate Cox proportional hazards model (accounting also for clinical covariates) to the four IGV cluster scores. It was not possible to fit such a model to the full set of 10014 genes, because there are many more predictor variables (genes) than samples (Vittinghoff & McCulloch, 2007), and doing so would have resulted in over-fitting. This was the reason for using the Elastic Net penalised Cox regression method, which does not have this limitation on the number of predictor variables. However, reducing the prognostic signature to 4 cluster scores, i.e., 4 predictors, allows the Cox proportional hazards model to be fitted. This results in a model coefficient for each cluster score/predictor; these are used to calculate the IGV prognostic score. The IGV prognostic score is a one-number prognostic indicator for a single sample/patient, and I note that it must be calculated based on all four cluster scores, to be significantly prognostic.

The median of this IGV prognostic score was used to divide the patients of the main OC data set into better and worse prognostic groups, shown in Figure 6.3a and 6.3b. The IGV prognostic score was validated in two independent sets of cancers derived from the Mullerian tract. A new OC set from the Mayo Clinic ($n = 198$) confirmed the prognostic capacity of the

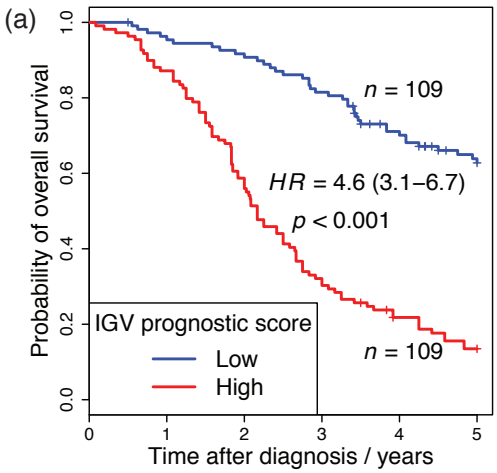
IGV prognostic score in both univariate (Figure 6.3c) and multivariate (Figure 6.3d) analyses. In order to test whether the IGV prognostic score is only limited to OC, or whether it is also predictive in other cancers which arise from the same embryological structure (i.e., the Mullerian duct), I applied the prognostic score to a publically available uterine corpus endometrioid carcinoma (UCEC) set from *The Cancer Genome Atlas* (TCGA) (Collins & Barker, 2007) ($n = 358$). Again, in both univariate (Figure 6.3e) and multivariate (Figure 6.3f) analyses, I was able validate the IGV prognostic score.

6.2.3 Functional role of transcription-factor activity in IGV

In order to investigate the question, 'what is IGV', I examined TF binding to the gene body regions of the OC prognostic signature genes, and tested the correlation of TF expression with the IGV of the genes they bind to (in a TCGA set of basal breast cancers). I found that each prognostic signature cluster shows its own distinctive pattern of TF binding (Figure 6.4a), which we can hypothesise is associated with the biological processes responsible for the characteristic pattern of IGV observed in that cluster.

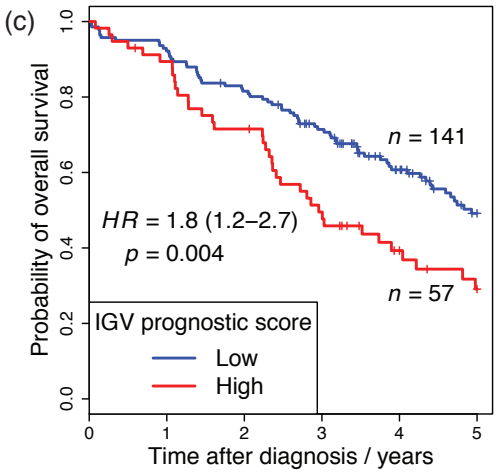
Transcription factor binding site information, obtained from the ENCODE (Encyclopedia of DNA Elements) project (Consortium *et al.* , 2004), was available for the gene body regions of all the genes represented on the Illumina HumanMethylation 450K array, for 55 transcription factors. I tested each of these 55 TFs, for significantly increased or decreased binding to the genes of each prognostic signature cluster. Cluster hypo 2 only consists of 19 genes, and hence we would not expect to see many significant correlations, due to small sample size. For cluster hyper 2, we see that 20% (11/55) of the TFs tested show significantly more binding to these genes than expected, whereas 16% show significantly less binding than expected. On the other hand, for clusters hyper 1 and hypo 1, not a single TF showed higher than expected binding, whereas 27% and 38% of TFs show lower than expected binding to the genes comprising cluster hyper 1 and hypo 1, respectively. This indicates that, also referring to the 'hyper' and 'hypo' directionality information, the clusters each represent TF binding patterns which are distinct from one another, and hence may be associated with different biological processes.

I also wanted to test the actual correlation of expression of the TFs with IGV of the genes they bind to, and genes they do not bind to, genome-wide. To do this, I used a TCGA set of basal breast cancers, for which 450k methylation data as well as expression data exist. It has been comprehensively demonstrated by the TCGA consortium that high-grade serous ovarian and uterine and BRCA basal cancers are extremely molecularly similar, and may share the same molecular origin (Network *et al.* , 2012). Figures 6.4b and 6.4c show TFs with significantly more positive, and more negative, correlation with IGV of the genes they bind to, compared



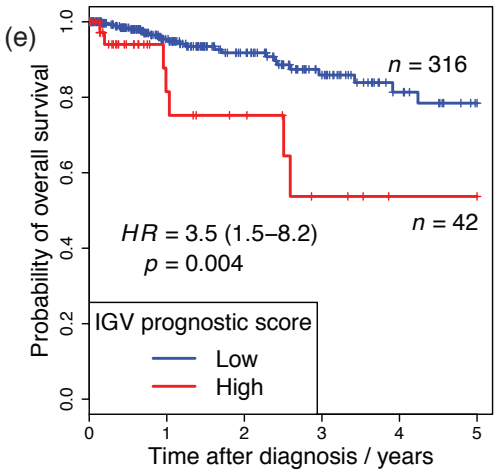
(b)

	HR (95%CI)	*p*	*n*
IGV prognostic score	4.1 (2.8-6.2)	<0.001	209
Age	1 (0.7-1.4)	1	209
Stage	5.2 (1.9-14.4)	0.002	209
Grade	0.9 (0.6-1.4)	0.7	209
Residual disease	1.4 (1-2)	0.05	209



(d)

	HR (95%CI)	*p*	*n*
IGV prognostic score	1.9 (1.1-3.1)	0.01	149
Age	0.8 (0.5-1.3)	0.4	149
Stage	12.3 (1.5-99.2)	0.02	149
Grade	1.5 (0.5-4.2)	0.5	149
Residual disease	2.5 (1.4-4.4)	0.001	149



(f)

	HR (95%CI)	*p*	*n*
IGV prognostic score	2.8 (1.1-7.2)	0.04	308
Age	1.2 (0.5-2.9)	0.7	308
Stage	2.1 (0.8-5.5)	0.1	308
Grade	0.9 (0.3-2.9)	0.9	308
Residual disease	8.9 (3.1-26)	<0.001	308

Figure 6.3: *IGV OC prognostic signature validation*

(a), (c) and (e): Comparison of survival curves of groups defined by the IGV prognostic score, in: (a) the main OC data set, (c) the Mayo Clinic OC validation set, (e) the uterine cancer TCGA validation set. The groups are divided by the median IGV prognostic score derived in the main OC DNAm data-set. The hazard ratio (HR) is displayed with 95% C.I. in brackets, with corresponding p-value calculated by univariate Cox regression. (d), (e) and (f): Multivariate Cox regression comparing the same groups defined by the IGV prognostic score.

to the genes they do not. It is interesting that the two most highly ranked transcription factors according to increased positive correlation of their expression with IGV in bound genes,

Rad21 and Brg1 (SMARCA4), are both parts of chromatin modifying complexes with relevance to stem cell identity (Nitzsche *et al.* , 2011; Attanasio *et al.* , 2014). In particular, Brg1 (SMARCA4) has been shown recently to have particular relevance to small-cell ovarian cancer (Witkowski *et al.* , 2014; Ramos *et al.* , 2014; Jelinic *et al.* , 2014). The overlap between the TFs which show significantly different binding patterns in relation to the OC prognostic signature genes, and TFs which display significantly altered correlation of their expression with IGV of genes they bind to, is shown in Figure 6.4d. Much relevant detail has already been reported about most of these TFs (references noted in the figure): either their binding is influenced by methylation (or *vice-versa*), or they are involved with chromatin remodelling in stem cells. The TFs shown in Figure 6.4d are important to the processes underlying disease progression, which are associated with the OC prognostic signature. Therefore I hypothesise that IGV, in the OC prognostic signature gene panel, represents a surrogate measure for their activity and role in disease transformation.

Genome-wide Analysis of Gene-Expression Correlation with IGV

I carried out a more general analysis of gene expression and IGV, again using the TCGA BRCA data-set, basal samples. As in chapter 5, section 5.5.2, I calculated the Spearman correlation of the IGV of each gene with the expression of every gene, genome wide, resulting in a matrix of gene expression and IGV correlations. In this correlation matrix, each column represents the correlation structure of the IGV of a particular gene with the gene expression patterns across all genes. Correspondingly, each row in the matrix represents the correlation structure of the expression of a particular gene with the IGV patterns across all genes. Also as in chapter 5, section 5.5.2, I used the methodology developed in chapter 4 to infer an adjacency matrix from this correlation matrix, in doing so identifying significant correlations (correponding to 1 in the adjacency matrix), and discarding insignificant correlations (corresponding to 0 in the adjacency matrix). Using the methodology developed in chapter 5, as in section 5.5.2, I co-clustered this correlation matrix, to find groups, or co-clusters, with a large number of significant correlations between IGV and gene expression. Co-clustering is most familiar in genomics in the scenario in which gene expression patterns are compared with arrays or subjects, grouping both genes and arrays/subjects simultaneously, and it can be used equivalently here to compare gene expression and IGV patterns. Here, by co-clustering, we seek groups of genes with similar expression patterns, while at the same time finding groups of genes with similar patterns of IGV. The only way the IGV of one gene can 'see' that of another, is via the expression patterns of other genes which it is correlated with, and similarly, the only way the expression pattern of one gene can 'see' that of another, is via the IGV patterns of other genes which it is correlated with.

(a)

	Hyper1	Hyper2	Hypo1	Hypo2
BAF155	q=0.00053, OR=0.37 (0.2-0.67)	q=5.3e-05, OR=2.5 (1.6-4)	q=1.5e-06, OR=0.18 (0.07-0.38)	q=0.65, OR=1.4 (0.49-4.1)
BAF170	q=0.053, OR=0.53 (0.27-0.98)	q=0.0089, OR=1.8 (1.2-2.6)	q=0.0075, OR=0.32 (0.13-0.7)	q=0.64, OR=1.3 (0.47-3.6)
BCL3	q=0.56, OR=0.8 (0.41-1.5)	q=0.01, OR=0.48 (0.28-0.79)	q=0.29, OR=0.56 (0.24-1.2)	q=0.56, OR=0.58 (0.14-1.8)
c-Fos	q=0.89, OR=1 (0.57-1.9)	q=0.0013, OR=0.42 (0.24-0.7)	q=0.0037, OR=0.26 (0.08-0.66)	q=0.29, OR=1.8 (0.66-5)
c-Myc	q=0.00018, OR=0.34 (0.18-0.62)	q=2.6e-07, OR=3.3 (2-5.4)	q=5.6e-05, OR=0.25 (0.11-0.52)	q=0.49, OR=1.5 (0.53-4.4)
CEBPB	q=0.16, OR=0.59 (0.32-1.1)	q=0.7, OR=0.91 (0.61-1.4)	q=0.0043, OR=0.3 (0.12-0.66)	q=0.7, OR=1.2 (0.44-3.4)
CTCF	q=0.34, OR=0.7 (0.38-1.2)	q=1.6e-05, OR=0.37 (0.23-0.58)	q=0.54, OR=0.79 (0.41-1.5)	q=0.34, OR=1.7 (0.64-5)
EBF	q=0.088, OR=0.57 (0.31-1)	q=0.44, OR=0.85 (0.57-1.3)	q=0.041, OR=0.44 (0.22-0.86)	q=0.14, OR=0.44 (0.14-1.2)
FOSL2	q=0.04, OR=0.5 (0.25-0.93)	q=0.12, OR=0.72 (0.47-1.1)	q=3e-05, OR=0.15 (0.039-0.41)	q=0.12, OR=0.4 (0.096-1.2)
FOXP2	q=0.12, OR=0.19 (0.0047-1.1)	q=0.023, OR=0.19 (0.023-0.72)	q=0.06, OR=0 (0-0.92)	q=1, OR=0.58 (0.014-3.7)
GABP	q=0.012, OR=0.26 (0.051-0.79)	q=0.0041, OR=2.1 (1.3-3.1)	q=0.0054, OR=0.11 (0.0026-0.62)	q=0.56, OR=0.53 (0.06-2.3)
GR	q=0.052, OR=0.49 (0.21-1)	q=0.029, OR=0.57 (0.33-0.93)	q=0.029, OR=0.33 (0.1-0.83)	q=0.029, OR=0.14 (0.0034-0.9)
HEY1	q=0.029, OR=0.52 (0.29-0.93)	q=2.9e-11, OR=4.6 (2.8-8)	q=2.5e-06, OR=0.18 (0.067-0.4)	q=0.65, OR=1.3 (0.48-3.7)
HNF4A	q=0.023, OR=0.44 (0.2-0.86)	q=0.15, OR=0.7 (0.44-1.1)	q=0.00039, OR=0.18 (0.046-0.49)	q=0.48, OR=0.64 (0.18-1.9)
Ini1	q=7.6e-05, OR=0.33 (0.19-0.58)	q=3.1e-08, OR=5.1 (2.6-11)	q=3.1e-08, OR=0.17 (0.078-0.33)	q=0.21, OR=2.4 (0.68-13)
JunD	q=0.3, OR=0.71 (0.4-1.2)	q=0.0058, OR=0.56 (0.37-0.84)	q=0.0054, OR=0.35 (0.17-0.71)	q=1, OR=1.1 (0.38-2.9)
Max	q=0.0022, OR=0.42 (0.23-0.74)	q=1e-04, OR=2.5 (1.6-4)	q=0.00018, OR=0.28 (0.13-0.56)	q=0.36, OR=1.6 (0.58-5.3)
NFKB	q=0.008, OR=0.46 (0.26-0.81)	q=1.4e-07, OR=5.6 (2.6-14)	q=0.007, OR=0.41 (0.21-0.77)	q=0.31, OR=2 (0.57-11)
NRSF	q=0.0091, OR=0.44 (0.22-0.81)	q=0.25, OR=0.8 (0.53-1.2)	q=8e-04, OR=0.26 (0.11-0.58)	q=0.15, OR=0.43 (0.12-1.3)
Pbx3	q=1, OR=0.88 (0.46-1.6)	q=0.038, OR=0.55 (0.33-0.88)	q=0.29, OR=0.57 (0.24-1.2)	q=1, OR=1 (0.32-2.9)
POU2F2	q=0.0035, OR=0.39 (0.2-0.73)	q=4e-04, OR=2.2 (1.4-3.3)	q=0.3, OR=0.67 (0.34-1.3)	q=0.65, OR=1.3 (0.48-3.7)
PU.1	q=0.46, OR=1.3 (0.76-2.5)	q=0.083, OR=0.67 (0.45-0.99)	q=0.023, OR=0.42 (0.21-0.81)	q=0.82, OR=0.83 (0.3-2.3)
Rad21	q=0.84, OR=0.79 (0.44-1.4)	q=1.7e-05, OR=0.37 (0.23-0.59)	q=0.88, OR=0.93 (0.48-1.8)	q=0.86, OR=1.4 (0.5-3.8)
Sin3Ak-20	q=0.034, OR=0.39 (0.14-0.92)	q=0.0023, OR=2 (1.4-3.1)	q=0.013, OR=0.24 (0.047-0.75)	q=1, OR=0.88 (0.21-2.8)
STAT1	q=0.0064, OR=0.37 (0.18-0.72)	q=0.38, OR=1.2 (0.8-1.8)	q=0.044, OR=0.46 (0.21-0.93)	q=0.23, OR=1.9 (0.7-5.4)
TAF1	q=0.34, OR=0.76 (0.43-1.3)	q=3.8e-12, OR=5.8 (3.2-11)	q=0.042, OR=0.48 (0.24-0.91)	q=0.084, OR=2.9 (0.91-12)
TCF12	q=0.0021, OR=0.32 (0.14-0.65)	q=0.92, OR=1 (0.69-1.5)	q=0.33, OR=0.62 (0.3-1.2)	q=0.92, OR=0.87 (0.29-2.4)
USF-1	q=0.024, OR=0.47 (0.23-0.91)	q=0.0078, OR=0.51 (0.31-0.8)	q=0.0084, OR=0.33 (0.12-0.74)	q=1, OR=1 (0.34-2.8)

(b)

	Median bound	Median unbound cor	q-val
Rad21	0.09	0.035	1.1e-21
Brg1	0.12	0.076	1.6e-12
GABP	0.072	0.041	3.9e-10
c-Myc	0.043	0.025	4.6e-06
Nrf1	0.11	0.086	4.7e-06
BCL11A	0.069	0.049	1.7e-05
FOXP2	0.083	0.04	1e-04
Pbx3	0.016	0.0027	0.00038
CTCF	0.02	0.01	0.001
SRF	0.037	0.0075	0.005
SIX5	0.023	-0.0039	0.0076
HNF4A	0.015	0.0059	0.014
Sin3Ak-20	0.11	0.096	0.026

(c)

	Median bound	Median unbound cor	q-val
GR	-0.22	-0.11	2.5e-51
BCL3	-0.19	-0.11	5.5e-28
PU.1	-0.18	-0.097	5.9e-27
NRSF	-0.11	-0.062	1.8e-21
c-Jun	-0.063	-0.037	3.7e-18
c-Fos	-0.095	-0.048	2e-17
BATF	-0.23	-0.14	4.1e-17
JunD	-0.11	-0.067	5.6e-12
TCF12	-0.029	-0.0056	4e-11
RXRA	-0.083	-0.049	1.2e-10
EBF	-0.12	-0.093	1.9e-08
p300	-0.06	-0.042	3e-08
FOSL2	-0.11	-0.073	7.5e-08
STAT1	-0.12	-0.092	4.1e-06
IRF4	-0.13	-0.12	0.00012
NFKB	-0.024	-0.013	0.0039

(d)

	Hyper1	Hyper2	Hypo1	Hypo2
Increased Binding, Positive Correlation with IGV		c-Myc (Gartel, 2006), GABP (Yokomori *et al.* , 1998), Sin3Ak-20 (Williams *et al.* , 2011)		
Increased Binding, Negative Correlation with IGV		NFKB (Kirillov *et al.* , 1996)		
Decreased Binding, Positive Correlation with IGV	c-Myc (Gartel, 2006), GABP (Yokomori *et al.* , 1998), HNF4A, Sin3Ak-20 (Williams *et al.* , 2011)	CTCF (Nitzsche *et al.* , 2011), FOXP2 (Zechner *et al.* , 2012), Pbx3, Rad21 (Nitzsche *et al.* , 2011)	c-Myc (Gartel, 2006), GABP (Yokomori *et al.* , 1998), HNF4A, Sin3Ak-20 (Williams *et al.* , 2011)	
Decreased Binding, Negative Correlation with IGV	FOSL2, NFKB (Kirillov *et al.* , 1996) , NRSF (Coulson, 2005), STAT1, TCF12	BCL3, c-Fos (Gustems *et al.* , 2014), GR, JunD (Ng *et al.* , 2013)	c-Fos (Gustems *et al.* , 2014), EBF (Malone *et al.* , 2001), FOSL2, GR, JunD (Ng *et al.* , 2013), NFKB (Kirillov *et al.* , 1996), NRSF (Coulson, 2005), PU.1 (Zhu *et al.* , 2003), STAT1	GR

Figure 6.4: *Transcription factor binding and expression correlation with IGV*

(a) False discovery rate adjusted p-values and odds-ratios (OR) show enrichment of binding of specific transcription factors (TFs), to the gene body regions of the genes of each cluster. TFs for which binding is significantly over or under enriched (Fisher's exact test, FDR $q < 0.05$) are coloured green and red, respectively. (b) TFs which show significantly more positive correlation with IGV of the genes they bind to, compared to the genes they do not bind to. (c) TFs which show significantly more negative correlation with IGV of the genes they bind to, compared to the genes they do not bind to. (d) TFs which are significant according to (a) and either (b) or (c); TFs with known relevance are indicated with a reference to the relevant study. The lack of enrichment of TF binding to the genes of cluster hypo2, is a reflection of the small number (19) of genes in this cluster.

This co-clustering was carried out without any reference to the prognostic clusters, and the result is shown in figure 6.5. Then, testing for association of these discovered co-clusters with the OC prognostic signature clusters found previously, five of the IGV-groupings (columns) are significantly enriched (Fisher's exact test, FDR $q < 0.05$) by genes of those prognostic clusters. The most significant of these prognostic co-clusters are are highlighted in blue; they represent the most relevant co-clusters to the genes of the OC prognostic signature. Gene-set enrichment analysis (Subramanian *et al.* , 2005) was carried out on the gene-expression groupings (rows) of these blue prognostic co-clusters; many significant gene sets are found for each of these groupings, including several which relate to known cancer, stem-cell, and immune function gene sets. These prognostic-signature associated gene-expression groups (rows in figure 6.5) of particular interest are:

- Groups 3, 10, 12, 14, 24, 37, 45, 49, 52, 54 and 67 contain many gene-sets relating to stem-cells and differentiation processes, as well as many known cancer gene-sets, including sets relating to breast and brain cancers.
- Groups 1 and 9 are related to many immune-function gene-sets.
- Group 74 is associated with three out of four of the OC prognostic signature IGV clusters, and it is highly enriched for gene-sets relating to histone proteins, providing further evidence for the link between IGV and chromatin modifying factors.

6.3 Discussion

I have found that IGV (a per-gene measure of intra-gene variability of DNAm) is a far more robust prognostic marker tool than mean methylation levels, which can show considerable technological and experimental variation (necessitating complicated normalisation and batch-correction). Figure 6.2b indicates that gene body IGV has the potential to become a very effective prognostic tool, perhaps more so than measures of mean methylation. While it is true that the Illumina HumanMethylation 450K array provides more measurements for the gene-body than any other genomic region, and hence gene-body derived measures can potentially provide more information than those derived from the promoter region when using this technology, this is unlikely to be the whole explanation for its effectiveness in this study. I note that it has previously been found that the most variably methylated CpGs occur more frequently in gene bodies than in promoters (Consortium *et al.* , 2012). However, while it is well established that promoter methylation in CpG-dense regions is associated with gene repression (Jones, 2012), the effects of gene-body methylation are less clear. Gene body methylation has recently been

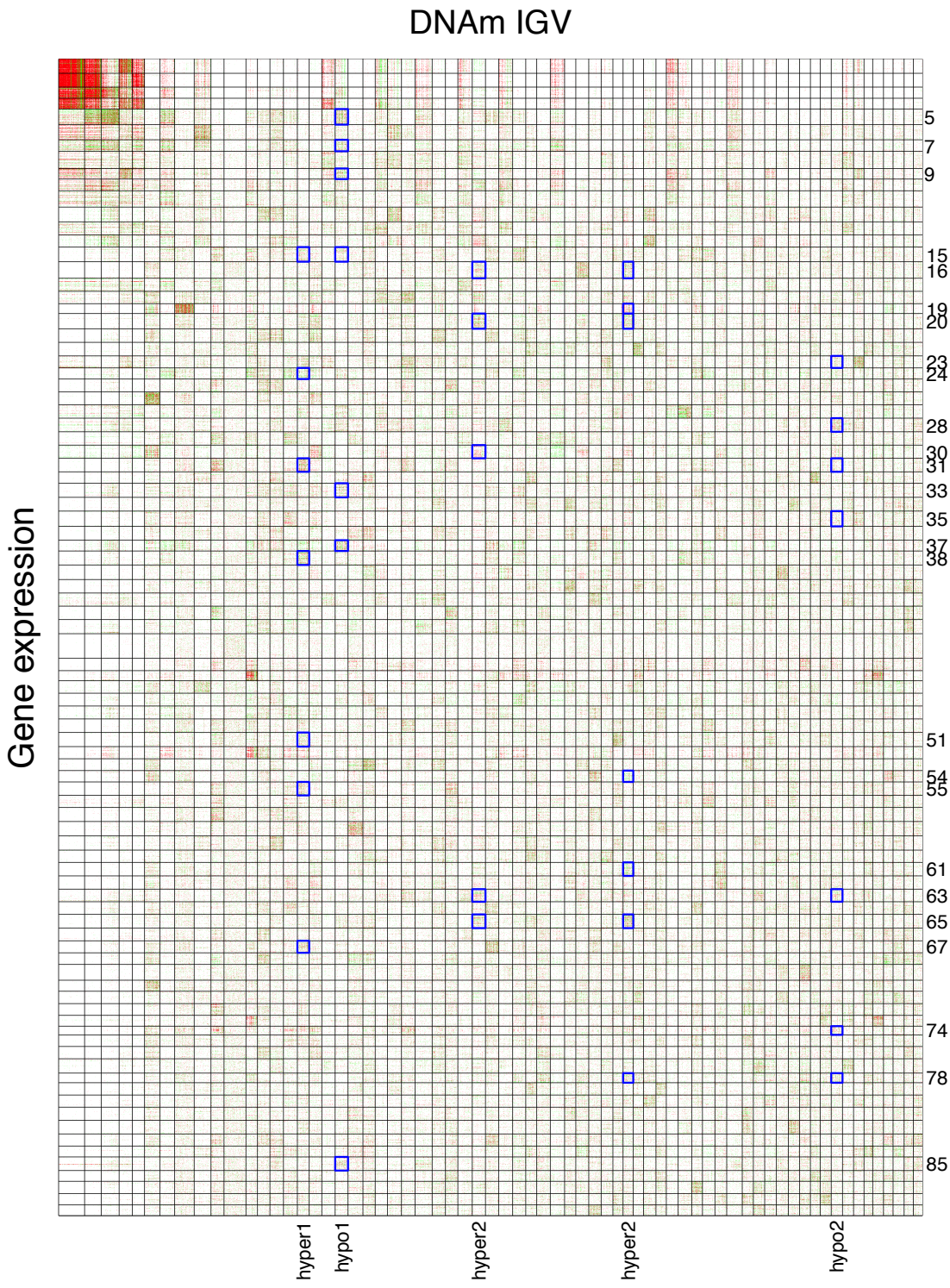


Figure 6.5: *Correlation of IGV patterns with genome-wide expression patterns. Co-clustering was carried out, in order to group genes which are highly correlated, in terms of their expression patterns, with the IGV of specific and different groups of genes, and vice-versa. Significant positive and negative correlations are shown in green and red, respectively. The most significant co-clusters which relate to the previously-defined OC prognostic signature clusters are outlined in blue.*

shown to have a direct effect on gene expression level (Yang *et al.* , 2014), however it may also be associated with other influences on transcription and translation, such as prevalence

of alternatively spliced gene products (Jones, 2012). Findings are also starting to emerge that gene-body methylation may be an effective therapeutic target in cancer (Yang *et al.* , 2014).

The OC prognostic signature which I have developed based on IGV is able to blindly predict patient prognostic outcome in two independent data sets from studies by the Mayo Clinic and TCGA ($n = 198$ and $n = 358$, respectively), with highly statistically significantly different clinical outcomes between these groups ($p = 0.004$ in both data sets). Asking the question, what is IGV, I examined binding of TFs and the correlation of their expression with IGV of genes they bind to. This revealed a distinctive pattern of TF binding to different groups of genes, and identified a panel of TFs which are highly associated with prognostic IGV. Using linked DNA methylation and gene expression data, I also identified groups of genes with highly associated expression and IGV patterns, providing a starting-point for further investigation into the mechanistic roles of the observed IGV patterns.

I have conclusively demonstrated that the OC prognostic signature is an effective and robust prognostic tool, and I also hypothesise that it is an easy to measure surrogate for disease processes mediated by specific transcription factors. IGV is an independent and robust prognostic marker, which does not require complex data normalisation to be effective.

6.4 Methods

6.4.1 Data and preprocessing

The main ovarian cancer (OC) data set, which was used to derive the OC prognostic signature, consists of 221 samples each of which was taken from a different patient, of whom 158 died from the disease before the end of the study. For each sample, a DNA methylation profile collected via the Illumina Infinium HumanMethylation450 platform was available, together with information on the clinical variables survival status (alive or not), survival time (i.e., time to last follow up or time to death), disease stage (I-IV), disease grade (1-3), and residual disease status (present or not). 3 samples were removed due to missing clinical data, leaving the the $n = 218$ samples used to derive the OC prognostic signature. A further 9 samples were excluded from the multivariate analysis of the IGV prognostic score, due to additional missing clinical data.

An independent data set from a study of OC carried out by the Mayo Clinic was used for validation of the OC prognostic signature. Data from this study similarly included a DNA methylation profile for each sample collected via the Illumina Infinium HumanMethylation450 platform; clinical data was also available for this data set for the same variables as the main OC data set. There were $n = 198$ samples in this data set, of whom 115 died from the disease

before the end of the study. 49 samples were excluded from the multivariate analysis of the IGV prognostic score, due to missing clinical data.

An additional independent data set from a study of uterine corpus endometrioid carcinoma (UCEC) for further validation of the OC prognostic signature was downloaded with the *The Cancer Genome Atlas* (TCGA) project (Collins & Barker, 2007). Data from this study similarly included a DNA methylation profile for each sample collected via the Illumina Infinium HumanMethylation450 platform, which was downloaded at level 3; clinical data was also downloaded if possible for each sample for the same variables as the OC data set. There were 358 samples in this data set, of whom 32 died from the disease before the end of the study. 50 samples were excluded from the multivariate analysis of the IGV prognostic score, due to missing clinical data.

For the gene expression analysis in BRCA basal samples, I downloaded DNAm data for breast cancer invasive carcinoma (BRCA) basal samples from TCGA (42 samples), again collected via the Illumina Infinium HumanMethylation450 platform, and downloaded at level 3. I also downloaded gene expression data for the same 42 samples from TCGA, at level 3.

Probes with non-unique mappings and which map to SNPs had already been removed from the UCEC and BRCA TCGA DNAm data before they were downloaded, and these same probes were also removed from the other DNAm data sets. Probes mapping to sex chromosomes were also removed; in total 98384 probes were removed from the DNAm data sets, of the 482421 probes originally present on the array. After removal of these probes, 270985 probes with known gene annotations remained. Individually for each data set, probes were then removed if they had less than 95% coverage across samples; probe values were also replaced if they had corresponding detection p-value greater than 5%, by KNN (k nearest neighbour) imputation ($k = 5$).

A summary of the data-sets analysed here appears in Table 6.1.

Data-set	Patients	Deaths	Removed
Main OC DNAm	221	158	12
Mayo OC DNAm	198	115	49
TCGA UCEC DNAm	358	32	50
TCGA BRCA basal DNAm	42	NA	NA
TCGA BRCA basal Expr	42	NA	NA

Table 6.1: *Data-sets analysed*
Abbreviations: DNAm, DNA methylation; OC, ovarian cancer; UCEC, uterine corpus endometrial carcinoma; BRCA, breast cancer invasive carcinoma.

6.4.2 Per-gene methylation measures

Four per-gene measures were tested, as follows:

- **TSS200 mean** The mean methylation level of the probes annotated to the TSS200 region, which is the region within 200bp upstream of the TSS (transcriptional start site); approximately the promoter region.
- **TSS200 IGV** The variance of the methylation level of the probes annotated to the TSS200 region.
- **Gene body mean** The mean methylation level of the probes annotated to the gene body.
- **Gene body IGV** The variance of the methylation level of the probes annotated to the gene body.

To calculate these measures, annotation information specifying which probes map to each gene and genomic region was used, as downloaded from Gene Expression Omnibus (GEO) (Edgar *et al.* , 2002), and as part of the *R* / *Bioconductor* software package *IlluminaHumanMethylation450k*. The mean methylation was calculated for genes with any number of probes annotated to the relevant genomic region (12970 and 15839 genes for TSS200 and gene body respectively). The methylation variance was calculated for genes with at least 3 probes annotated to the relevant genomic region (7557 and 10014 genes for TSS200 and gene body respectively).

6.4.3 Cross-validation to compare per-gene methylation measures and derive OC prognostic signature

The samples (patients) of the main OC data-set were randomly split in to a 'training set' (2/3 of the data, 145 samples) and a 'test set' (the remaining 1/3 of the data, 73 samples). The Elastic Net (Zou & Hastie, 2005; Simon *et al.* , 2011) was used to select a prognostic group of genes and fit a predictive model to these genes based on the training set; this model was then assessed using the test set. This was repeated 2001 times for each of the four per-gene methylation measures and for both SWAN-normalised (Makismovic *et al.* , 2012) data and raw (un-normalised) data.

As the aim here is to predict clinical outcome, the Elastic Net was used in its penalised Cox regression form, as implemented in the *R* package *GLMNET* (Simon *et al.* , 2011). Cox regression fits the model by setting the model coefficients so as to maximise the partial likelihood, as defined by equation (6.1),

$$L(\boldsymbol{\theta}) = \prod_{j \in S} \frac{e^{\boldsymbol{\theta}^\top \mathbf{x}_j}}{\sum_{j' \in R_j} e^{\boldsymbol{\theta}^\top \mathbf{x}_{j'}}}, \tag{6.1}$$

where $\boldsymbol{\theta}$ denotes the vector of model coefficients (of dimension equal to the number of genes considered, typically of the order of 10000), $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ are the vectors of predictor variable values for samples j and j' respectively (here, per-gene methylation measures), S is the set of patients who died during the study, and R_j is the set of samples 'at risk' during the time interval when patient j died, defined as $R_j = \left\{j' \middle| Y_{j'} \geq Y_j\right\}$, where Y_j and $Y_{j'}$ are the times of death of patients j and j' respectively. The Elastic Net penalises the log-likelihood corresponding to equation (6.1), constraining it according to the magnitude of the model fit coefficients, by subtracting this constraint from the likelihood; in doing so, it 'chooses' the best combination of predictor variables (per-gene methylation measures), by adjusting the corresponding model coefficients, and setting these coefficients to zero where the variables provide no useful information or redundant information. The constraint is a combination of some multiples of the L_1 and L_2 norms of the model fit coefficients; the severity and balance of the constraint is controlled by the parameters λ (a 'magnitude' parameter) and α (a 'blending' parameter). Hence, the Elastic Net Cox model is fitted by finding model coefficients $\hat{\boldsymbol{\theta}}$ which maximise the penalised log likelihood $\phi(\boldsymbol{\theta}, \lambda, \alpha)$ in equation (6.2),

$$\phi(\boldsymbol{\theta}, \lambda, \alpha) = \frac{2}{N} l(\boldsymbol{\theta}) - \lambda \left(\alpha \|\boldsymbol{\theta}\|_{L_1} + \frac{(1-\alpha)}{2} \|\boldsymbol{\theta}\|_{L_2}^2 \right), \tag{6.2}$$

where N is the number of samples, $\|\cdot\|_{L_1}$ and $\|\cdot\|_{L_2}$ are the L_1 and L_2 norms, and $l(\boldsymbol{\theta}) = \log\left(L(\boldsymbol{\theta})\right)$. The *R* package *GLMNET* used for these model fits sets the λ parameter internally using ten-fold cross validation, and requires the user to set the α parameter ($0 \leq \alpha \leq 1$), which was in this case set by choosing the value which minimises the model error after trialling values from 0 to 1 in evenly-spaced intervals of 0.1. Model fitting in this way leads to a set of model coefficients $\hat{\boldsymbol{\theta}}$ for a particular set of predictors (i.e., genome-wide per-gene methylation measures), with one coefficient per predictor, defining those predictors which are present in the model (i.e., predictors with corresponding non-zero coefficients), and their relative weightings.

The fitted model coefficients $\hat{\boldsymbol{\theta}}$ calculated according to equations (6.1) and (6.2) and the training set data were used to calculate a score $\hat{\boldsymbol{\theta}}^\top \mathbf{x}_j$ for each patient j, based on the corresponding per-gene methylation measures $\mathbf{x}_j$. These scores were then used to divide the training set into tertiles, defining high and low risk groups. The cutoffs defining the top and bottom tertiles in the training set were then used to divide the test set into three portions, and those most and least at risk (i.e., those test set patients with scores above the top cutoff, and below the bottom cutoff) were compared by Mantzel-Haenszel test, stratified for age, stage and residual disease, to assess the ability of this model fit to blindly predict patient survival, adjusted for

significant clinical covariates. Disease grade was not included in this stratification because it was not associated with survival for this data set, as assessed by multivariate Cox-regression. This is likely to be because disease grade does not offer any significant predictive ability in addition to other clinical covariates which are more strongly associated with survival outcome, such as disease stage. Upper and lower tertiles were compared here as previously by other authors (Zhuang *et al.* , 2012) for the OC prognostic signature generation, and the reasoning for doing so in this discovery stage, rather than comparing two groups separated by the median score, was in order to prioritise larger effect sizes. If the samples were split into two groups divided by the median score, relatively small differences in the per-gene methylation measures used to generate this score might result in patients being categorised as high or low risk, with corresponding significant test results from this small variation between patients. Comparison of upper and lower tertiles would be expected to be more robust / stable with respect to such small differences in per-gene methylation measures.

Each randomly-selected training set which the Elastic Net model was fitted to lead to a different set of genes being chosen. This is likely to be due to patient to patient heterogeneity in the main OC data set which was used to generate the OC prognostic signature. In order to infer a consistent OC prognostic signature from this data set, i.e., a consensus, the same process of randomly partitioning the data and fitting the model was repeated a total of 10^5 times for the gene-body IGV measure, with raw data. Of these, 8281 model fits were able to significantly predict survival in the respective test set (FDR $q < 0.1$). To generate the OC prognostic signature, genes were first ranked by how many of these 8281 significant model fits they appeared in. In the case of ties, genes were additionally ranked by, for each model fit, calculating the proportion of the sum of the absolute coefficient values for that model, which each gene selected as part of that model accounted for, and then comparing, for each tied gene, the mean of these proportions for that gene, across all the models it was selected as part of. Genes were assigned significance according to how many models they were selected as being part of, y, out of the total $k = 8281$ models selected as significantly associated with survival, under the null hypothesis that they were present in these observed y significant model fits by chance. If there were the same number of genes selected as part of each of these 8281 model fits, then this significance under the null hypothesis might be modelled by a binomial distribution, with the probability p_b of any gene being selected by chance as part of one model fit approximated by $p_b = f/m$, where f is the number of genes selected as part of each and every model fit, and m is the total number of genes for which gene-body methylation variance information is available. The probability of seeing a gene purely by chance in at least y model fits, out of a possible total

of k, with constant probability p_b of appearing in each of these k models, would then be given by equation (6.3),

$$P(Y \geq y) = \sum_{r=y}^{k} \left[\binom{k}{r} p_b^r \left(1 - p_b\right)^{(k-r)} \right]. \tag{6.3}$$

However, the number of genes selected, f, as part of each model, varies considerably (from 7 to 1697), and consequently p_b cannot be assumed to be constant. Alternatively, p_b could be modelled as being variable and bounded on $[0, 1]$, with a corresponding probability distribution $\pi_b\left(p_b\right)$. The distribution $\pi_b\left(p_b\right)$ can be estimated as the observed distribution of f among the $k = 8281$ significant model fits, again using $p_b = f/m$. This leads to a modelled probability, equation (6.4), of seeing any gene at least y times out of k model fits purely by chance, with p_b variable and with its distribution $\pi_b\left(p_b\right)$ empirically estimated as $\hat{\pi}_b\left(p_b\right)$,

$$P(Y \geq y) = \sum_{r=y}^{k} \left\{ \int_0^1 \hat{\pi}_b\left(p_b\right) \left[\binom{k}{r} p_b^r \left(1 - p_b\right)^{(k-r)} \right] \mathrm{d}p_b \right\}, \tag{6.4}$$

with the square brackets included in equation (6.4) to highlight the comparison with equation (6.3). In practice, the integral in equation (6.4) is replaced with a sum over the observed values of p_b, as calculated from the observed values of f, which range between 7 and 1697. A kernel-smoothed plot of $\hat{\pi}_b\left(p_b\right)$, the empirical probability density distribution of f and corresponding p_b, appears in figure 6.6.

6.4.4 Calculation of the DNAm IGV ovarian cancer prognostic score

Clustering was performed to identify groups of genes in the OC prognostic signature with similar patterns of IGV across patients. The clustering was carried out separately for genes individually associated with worse patient survival outcome for increased IGV ('hyper' genes) and for decreased IGV ('hypo' genes). Consensus clustering (Monti *et al.* , 2003) was used for the clustering, with a hierarchical clustering inner loop, using $1 - \rho$ as the distance measure, where ρ is the Spearman rank correlation coefficient. The following additional settings were used: probability of selecting a sample $= 0.8$, probability of selecting a feature $= 1$, number of resamplings $= 10^5$, maximum number of clusters $= 20$.

The discovered clusters were then filtered (to remove noise, and uncertainty associated with trends inferred from small groups of genes in these genome-wide data), retaining only those clusters which contained at least 10 genes, and only those clusters with mean IGV significantly associated with patient survival outcome. After filtering, four clusters remained, for two of which an increase in the cluster mean IGV was associated with worse patient survival outcome (called 'hyper 1' and 'hyper 2'), and for two of which a decrease in the cluster mean IGV

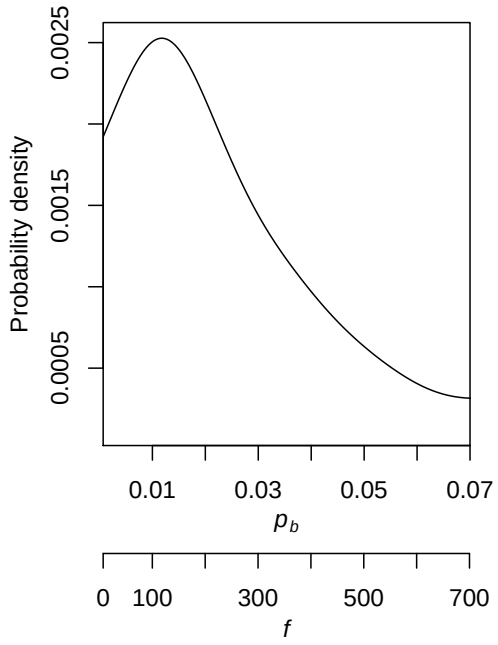


Figure 6.6: *Probability density distribution of the probabilities of a gene being included in a fitted model. The plot shows a kernel-smoothed empirical estimate of the probability density distribution of the number of genes included in each model, f, over the 8281 significant gene body methylation variance model fits, with corresponding probability of a gene being included in a model $p_b = f/m$, where m is the number of genes with gene body methylation variance information available.*

was associated with worse survival outcome (called 'hypo 1' and 'hypo 2'). The IGV cluster scores were then calculated, as the means of the IGV of the genes each of these four clusters.

In order to calculate the IGV prognostic score from these components, a Cox model (adjusted for clinical covariates) was fitted to these four IGV cluster scores. The coefficients for this model (standardised by the variance of the predictors) are fairly similar for each of the clusters (hyper 1: 0.22; hyper 2: 0.25; hypo 1: 0.23; hypo 2: 0.30), indicating that each cluster is important to the model, and to the prognostic predictions. The median of the IGV prognostic score calculated from this Cox model was used to divide the 218 patients in the main DNAm OC data-set used to derive the OC prognostic signature, into better and worse prognostic groups.

6.4.5 Validation of the ovarian cancer prognostic signature

The DNAm prognostic signature derived from the OC data set was validated in two independent DNAm data sets. The first of these data sets was taken from another study of OC ($n = 198$), and was supplied by the Mayo Clinic. The second of these data sets was taken from a study of uterine corpus endometrioid carcinoma (UCEC) ($n = 358$), and was downloaded from *The Cancer Genome Atlas* (TCGA) project (Collins & Barker, 2007).

The IGV prognostic score was similarly calculated by fitting a Cox model to the four IGV cluster scores in the main OC DNAm data set, adjusted for clinical covariates, then applying this

model to the equivalent IGV cluster scores in the Mayo Clinic OC and the TCGA UCEC validation sets. In order to make prognostic predictions in these independent data sets using only the DNAm data, the model was used to calculate the IGV prognostic score for the samples in the independent data sets from the fitted model coefficients corresponding to IGV cluster scores only, and not the clinical covariates. This IGV prognostic score was used to define better and worse prognostic groups in the independent data sets, separated by the median IGV prognostic score in the main OC data set. These prognostic groups were then compared, assessing statistical significance with univariate and multivariate Cox regression (i.e., respectively without and with adjustment for the clinical covariates).

6.4.6 Testing Transcription-factor binding correlation with IGV

I examined transcription factor binding to the OC prognostic signature genes, using the ENCODE (Encyclopedia of DNA Elements) chromatin immunoprecipitation (ChIP) data (Consortium *et al.* , 2004), with the ANNOVAR software (Wang *et al.* , 2010). Transcription factor binding site information was available, for the gene body regions defined, for 55 transcription factors. Each of these TFs was tested for significant over or under enrichment binding to the genes of each of the four prognostic signature clusters, with Fisher's exact test. I also tested the correlation of the expression level of each of these 55 TFs, with the IGV of genes the TF binds to, and the genes the TF does not bind to. I used a Kolmogorov-Smirnov test to assess whether, for each TF, there is significantly more positive, or more negative, correlation with IGV of the genes it binds to, compared to genes it does not. For this expression correlation analysis, I used the 42 TCGA BRCA basal samples with both expression and DNAm data available, because it was comprehensively demonstrated by the TCGA consortium that high-grade serous ovarian and uterine and BRCA basal cancers are extremely molecularly similar, and may share the same molecular origin (Network *et al.* , 2012).

6.5 Additional tables

Gene	No. models sig, out of 8281	*p*-val	*q*-val	Chr	Gene info
SEMA4A	8274	0	0	1	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4A
PYCARD	8196	0	0	16	PYD and CARD domain containing
RNF122	8129	0	0	8	ring finger protein 122
ALPL	8127	0	0	1	alkaline phosphatase, liver/bone/kidney
RIC3	8108	0	0	11	resistance to inhibitors of cholinesterase 3 homolog (C. elegans)
CXXC4	8075	0	0	4	CXXC finger protein 4
RSPH9	7945	0	0	6	radial spoke head 9 homolog (Chlamydomonas)
TMEM17	7877	0	0	2	transmembrane protein 17
PCTP	7869	0	0	17	phosphatidylcholine transfer protein
UBAP2L	7837	0	0	1	ubiquitin associated protein 2-like
PLDN	7652	0	0	15	biogenesis of lysosomal organelles complex-1, subunit 6, pallidin
ZNF727	7626	0	0	7	zinc finger protein 727
NRL	7583	0	0	14	neural retina leucine zipper
OBFC1	7488	0	0	10	oligonucleotide/oligosaccharide-binding fold containing 1
SERHL2	7245	0	0	22	serine hydrolase-like 2
FOXR1	7220	0	0	11	forkhead box R1
PROSC	7218	0	0	8	proline synthetase co-transcribed homolog (bacterial)
SLC50A1	7207	0	0	1	solute carrier family 50 (sugar transporter), member 1
PDE1B	7143	0	0	12	phosphodiesterase 1B, calmodulin-dependent
MEX3B	7109	0	0	15	mex-3 homolog B (C. elegans)
NOVA1	7003	0	0	14	neuro-oncological ventral antigen 1
EIF2C4	7002	0	0	1	argonaute RISC catalytic component 4
BK250D10	6945	0	0		
SPATA13	6857	0	0	13	spermatogenesis associated 13
C14orf64	6854	0	0	14	chromosome 14 open reading frame 64
ACRBP	6849	0	0	12	acrosin binding protein
CLRN3	6846	0	0	10	clarin 3
ARL10	6839	0	0	5	ADP-ribosylation factor-like 10
PITHD1	6828	0	0	1	PITH (C-terminal proteasome-interacting domain of thioredoxin-like) domain containing 1
CGN	6761	0	0	1	cingulin
SEC14L4	6746	0	0	22	SEC14-like 4 (S. cerevisiae)
HOXB9	6742	0	0	17	homeobox B9
SEC14L2	6700	0	0	22	SEC14-like 2 (S. cerevisiae)
ANKRD13B	6681	0	0	17	ankyrin repeat domain 13B
NINL	6672	0	0	20	ninein-like
COMMD6	6641	0	0	13	COMM domain containing 6
TBC1D20	6589	0	0	20	TBC1 domain family, member 20
IL17REL	6573	0	0	22	interleukin 17 receptor E-like
ENG	6572	0	0	9	endoglin
GPC5	6557	0	0	13	glypican 5
SAMD10	6483	0	0	20	sterile alpha motif domain containing 10
SRC	6482	0	0	20	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)
EGR3	6481	0	0	8	early growth response 3
FAM26F	6436	0	0	6	family with sequence similarity 26, member F
TMEM185B	6338	0	0	2	transmembrane protein 185B
SYBU	6326	0	0	8	syntabulin (syntaxin-interacting)
C14orf126	6316	0	0	14	D-tyrosyl-tRNA deacylase 2 (putative)
WDR65	6307	0	0	1	WD repeat domain 65
RPF2	6286	0	0	6	ribosome production factor 2 homolog (S. cerevisiae)
SNRNP27	6277	0	0	2	small nuclear ribonucleoprotein 27kDa (U4/U6.U5)
ATG4D	6213	0	0	19	autophagy related 4D, cysteine peptidase
ABCB6	6158	0	0	2	ATP-binding cassette, sub-family B (MDR/TAP), member 6
IFNGR1	6133	0	0	6	interferon gamma receptor 1
SLC27A6	6132	0	0	5	solute carrier family 27 (fatty acid transporter), member 6
TLE4	6111	0	0	9	transducin-like enhancer of split 4 (E(sp1) homolog, Drosophila)
TCL6	6078	0	0	14	T-cell leukemia/lymphoma 6 (non-protein coding)
ECEL1P2	6047	0	0	2	endothelin converting enzyme-like 1, pseudogene 2
LOC100134259	6006	0	0	2	uncharacterized LOC100134259
ZNF300P1	5956	0	0	5	zinc finger protein 300 pseudogene 1
SP5	5944	0	0	2	Sp5 transcription factor
ICOSLG	5928	0	0	21	inducible T-cell co-stimulator ligand
CENPW	5699	0	0	6	centromere protein W
GAMT	5692	0	0	19	guanidinoacetate N-methyltransferase
IQCG	5655	0	0	3	IQ motif containing G
PIK3R3	5641	0	0	1	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)
EGR2	5631	0	0	10	early growth response 2
GRASP	5628	0	0	12	GRP1 (general receptor for phosphoinositides 1)-associated scaffold protein
LOC391322	5608	0	0	22	D-dopachrome tautomerase-like
HRASLS5	5577	0	0	11	HRAS-like suppressor family, member 5
GPNMB	5576	0	0	7	glycoprotein (transmembrane) nmb
B4GALNT1	5542	0	0	12	beta-1,4-N-acetyl-galactosaminyl transferase 1
TCFL5	5518	0	0	20	transcription factor-like 5 (basic helix-loop-helix)
C1orf173	5501	0	0	1	chromosome 1 open reading frame 173
NOX4	5468	0	0	11	NADPH oxidase 4
ZBTB10	5461	0	0	8	zinc finger and BTB domain containing 10
VSTM5	5431	0	0	11	V-set and transmembrane domain containing 5
RPS7	5403	0	0	2	ribosomal protein S7
COL17A1	5360	0	0	10	collagen, type XVII, alpha 1
ZNF346	5324	0	0	5	zinc finger protein 346
CHSY3	5308	0	0	5	chondroitin sulfate synthase 3
ARMC7	5287	0	0	17	armadillo repeat containing 7
ARSA	5238	0	0	22	arylsulfatase A
PLA2G16	5232	0	0	11	phospholipase A2, group XVI
CD74	5226	0	0	5	CD74 molecule, major histocompatibility complex, class II invariant chain
GCM2	5202	0	0	6	glial cells missing homolog 2 (Drosophila)

Table 6.2: *Ovarian cancer prognostic signature - top 85 genes.*

Significance is assigned to genes according to the frequency with which they appear in model fits which are significantly predictive of patient outcome (survival, adjusted for clinical covariates). Gene body DNA methylation variance was used as a per-gene measure for the model fits.

	Symbol	Rank in prog. sig.	q-val	Chr	Info
1	*SEMA4A*	1	0	1	sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4A
2	*RSPH9*	7	0	6	radial spoke head 9 homolog (Chlamydomonas)
3	*C14orf64*	25	0	14	chromosome 14 open reading frame 64
4	*SRC*	42	0	20	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)
5	*C1orf173*	73	0	1	chromosome 1 open reading frame 173
6	*CHSY3*	80	0	5	chondroitin sulfate synthase 3
7	*MYO3A*	94	0	10	myosin IIIA
8	*PLK2*	97	0	5	polo-like kinase 2
9	*GPR137B*	127	0	1	G protein-coupled receptor 137B
10	*ABCA8*	137	0	17	ATP-binding cassette, sub-family A (ABC1), member 8
11	*HERC5*	141	0	4	HECT and RLD domain containing E3 ubiquitin protein ligase 5
12	*TMEM101*	154	0	17	transmembrane protein 101
13	*FAM162B*	158	0	6	family with sequence similarity 162, member B
14	*KANK1*	161	0	9	KN motif and ankyrin repeat domains 1
15	*TUBB2B*	176	0	6	tubulin, beta 2B class IIb
16	*PLK1*	191	0	16	polo-like kinase 1
17	*LHCGR*	194	0	2	luteinizing hormone/choriogonadotropin receptor
18	*C1QL3*	225	0	10	complement component 1, q subcomponent-like 3
19	*RHPN2*	240	0	19	rhophilin, Rho GTPase binding protein 2
20	*PARP15*	298	1.1e-166	3	poly (ADP-ribose) polymerase family, member 15
21	*IRF4*	307	3.9e-152	6	interferon regulatory factor 4
22	*COL6A5*	326	7.1e-124	3	collagen, type VI, alpha 5
23	*KBTBD8*	335	1.4e-115	3	kelch repeat and BTB (POZ) domain containing 8
24	*MLF1*	345	6.8e-103	3	myeloid leukemia factor 1
25	*PTH2R*	359	3.6e-81	2	parathyroid hormone 2 receptor
26	*ACHE*	364	2e-74	7	acetylcholinesterase
27	*C6orf97*	379	1.3e-58	6	coiled-coil domain containing 170
28	*GRB14*	390	1.3e-47	2	growth factor receptor-bound protein 14
29	*NPR3*	416	8.2e-34	5	natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)
30	*C20orf194*	438	2.5e-24	20	chromosome 20 open reading frame 194
31	*NID2*	446	2.4e-21	14	nidogen 2 (osteonidogen)
32	*KCNMB4*	471	5.1e-11	12	potassium large conductance calcium-activated channel, subfamily M, beta member 4
33	*B3GNT9*	476	1.7e-10	16	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 9
34	*NPTX1*	483	3.1e-09	17	neuronal pentraxin I
35	*SHISA9*	499	1.6e-06	16	shisa homolog 9 (Xenopus laevis)
36	*ASIP*	509	1.7e-05	20	agouti signaling protein
37	*CCND2*	519	0.00011	12	cyclin D2
38	*SNX18*	524	0.00021	5	sorting nexin 18
39	*CPA2*	529	4e-04	7	carboxypeptidase A2 (pancreatic)
40	*PRR25*	533	0.00047	16	proline rich 25
41	*DUSP27*	535	5e-04	1	dual specificity phosphatase 27 (putative)
42	*CD302*	546	0.00087	2	CD302 molecule
43	*SLC12A8*	549	0.001	3	solute carrier family 12 (potassium/chloride transporters), member 8
44	*PHACTR3*	553	0.0011	20	phosphatase and actin regulator 3
45	*OTX2*	569	0.0026	14	orthodenticle homeobox 2
46	*CAV2*	573	0.0035	7	caveolin 2
47	*PALM3*	577	0.004	19	paralemmin 3
48	*SEZ6L2*	578	0.0042	16	seizure related 6 homolog (mouse)-like 2
49	*SUN3*	601	0.01	7	Sad1 and UNC84 domain containing 3
50	*EPHA6*	612	0.013	3	EPH receptor A6
51	*HLA-F-AS1*	628	0.018	6	HLA-F antisense RNA 1
52	*ALDH1A2*	629	0.019	15	aldehyde dehydrogenase 1 family, member A2
53	*TRPC1*	637	0.022	3	transient receptor potential cation channel, subfamily C, member 1
54	*TSHZ3*	646	0.029	19	teashirt zinc finger homeobox 3
55	*ITGA4*	648	0.031	2	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)
56	*NEGR1*	653	0.033	1	neuronal growth regulator 1

Table 6.3: *Prognostic signature, cluster 'hyper 1'*

	Symbol	Rank in prog. sig.	q-val	Chr	Info
1	*RNF122*	3	0	8	ring finger protein 122
2	*PLDN*	11	0	15	biogenesis of lysosomal organelles complex-1, subunit 6, pallidin
3	*EIF2C4*	22	0	1	argonaute RISC catalytic component 4
4	*TBC1D20*	37	0	20	TBC1 domain family, member 20
5	*C14orf126*	47	0	14	D-tyrosyl-tRNA deacylase 2 (putative)
6	*RPF2*	49	0	6	ribosome production factor 2 homolog (S. cerevisiae)
7	*SNRNP27*	50	0	2	small nuclear ribonucleoprotein 27kDa (U4/U6.U5)
8	*TLE4*	55	0	9	transducin-like enhancer of split 4 (E(sp1) homolog, Drosophila)
9	*CENPW*	62	0	6	centromere protein W
10	*PIK3R3*	65	0	1	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)
11	*RPS7*	77	0	2	ribosomal protein S7
12	*PLA2G16*	83	0	11	phospholipase A2, group XVI
13	*HNRNPA3*	86	0	2	heterogeneous nuclear ribonucleoprotein A3
14	*XRCC2*	96	0	7	X-ray repair complementing defective repair in Chinese hamster cells 2
15	*GOLPH3*	100	0	5	golgi phosphoprotein 3 (coat-protein)
16	*CHORDC1*	110	0	11	cysteine and histidine-rich domain (CHORD) containing 1
17	*GFPT1*	120	0	2	glutamine–fructose-6-phosphate transaminase 1
18	*CISD1*	129	0	10	CDGSH iron sulfur domain 1
19	*TRA2A*	133	0	7	transformer 2 alpha homolog (Drosophila)
20	*MKRN2*	134	0	3	makorin ring finger protein 2
21	*CAD*	144	0	2	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase
22	*SPG11*	157	0	15	spastic paraplegia 11 (autosomal recessive)
23	*RNF34*	177	0	12	ring finger protein 34, E3 ubiquitin protein ligase
24	*SERINC3*	188	0	20	serine incorporator 3
25	*C3orf38*	192	0	3	chromosome 3 open reading frame 38
26	*MYC*	203	0	8	v-myc myelocytomatosis viral oncogene homolog (avian)
27	*RPE*	204	0	2	ribulose-5-phosphate-3-epimerase
28	*SNRPG*	210	0	2	small nuclear ribonucleoprotein polypeptide G
29	*HSF2*	216	0	6	heat shock transcription factor 2
30	*DSCR3*	219	0	21	Down syndrome critical region gene 3
31	*DEPTOR*	226	0	8	DEP domain containing MTOR-interacting protein
32	*SLC30A6*	238	0	2	solute carrier family 30 (zinc transporter), member 6
33	*SRP19*	239	0	5	signal recognition particle 19kDa
34	*YEATS4*	252	3.4e-313	12	YEATS domain containing 4
35	*PYGL*	259	4.6e-296	14	phosphorylase, glycogen, liver
36	*POLR3H*	264	1.6e-280	22	polymerase (RNA) III (DNA directed) polypeptide H (22.9kD)
37	*FAM65B*	274	2.1e-220	6	family with sequence similarity 65, member B
38	*PVALB*	277	5.5e-217	22	parvalbumin
39	*FAM98A*	293	1.5e-180	2	family with sequence similarity 98, member A
40	*RAB8A*	294	1.7e-176	19	RAB8A, member RAS oncogene family
41	*TTC17*	301	2.3e-163	11	tetratricopeptide repeat domain 17
42	*CWC25*	308	4.1e-150	17	CWC25 spliceosome-associated protein homolog (S. cerevisiae)
43	*EIF4EBP2*	309	4.2e-146	10	eukaryotic translation initiation factor 4E binding protein 2
44	*TXNL1*	311	4.2e-146	18	thioredoxin-like 1
45	*SPAG9*	319	3.6e-132	17	sperm associated antigen 9
46	*STXBP3*	324	3.5e-125	1	syntaxin binding protein 3
47	*CLIC4*	331	1.1e-119	1	chloride intracellular channel 4
48	*PSMC6*	334	4.3e-117	14	proteasome (prosome, macropain) 26S subunit, ATPase, 6
49	*ITM2B*	336	2.7e-114	13	integral membrane protein 2B
50	*SFT2D2*	343	2.3e-104	1	SFT2 domain containing 2
51	*CWF19L1*	348	2.6e-95	10	CWF19-like 1, cell cycle control (S. pombe)
52	*HMGN1*	350	3.2e-93	21	high mobility group nucleosome binding domain 1
53	*CYB5R1*	351	1.6e-92	1	cytochrome b5 reductase 1
54	*LOC153684*	355	1.6e-87	5	uncharacterized LOC153684
55	*NKX2-2*	361	1.2e-75	20	NK2 homeobox 2
56	*NDUFA11*	363	1.3e-74	19	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 11, 14.7kDa
57	*CDKN2C*	366	3.7e-72	1	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
58	*LOC387647*	370	9.4e-67	10	patched domain containing 3 pseudogene 1
59	*MSH6*	372	1.3e-64	2	mutS homolog 6 (E. coli)
60	*NEMF*	388	9.1e-48	14	nuclear export mediator factor
61	*PAPOLA*	398	1.1e-44	14	poly(A) polymerase alpha
62	*C8orf38*	399	2.3e-44	8	NADH dehydrogenase (ubiquinone) complex I, assembly factor 6
63	*CCDC138*	406	1.2e-40	2	coiled-coil domain containing 138
64	*KIAA1429*	407	1.8e-40	8	KIAA1429
65	*HDAC2*	408	2e-39	6	histone deacetylase 2
66	*TNFSF9*	411	5.7e-39	19	tumor necrosis factor (ligand) superfamily, member 9
67	*CLTA*	421	1.9e-32	9	clathrin, light chain A
68	*MAPK12*	424	1.2e-31	22	mitogen-activated protein kinase 12
69	*SCFD1*	428	4.8e-29	14	sec1 family domain containing 1
70	*PARP11*	440	7.1e-24	12	poly (ADP-ribose) polymerase family, member 11
71	*UGGT1*	442	5.7e-23	2	UDP-glucose glycoprotein glucosyltransferase 1
72	*MICALCL*	443	1.5e-22	11	MICAL C-terminal like
73	*MPHOSPH8*	457	1.5e-15	13	M-phase phosphoprotein 8
74	*SUZ12P*	466	1.1e-13	17	suppressor of zeste 12 homolog pseudogene 1
75	*EPHX2*	468	2.2e-13	8	epoxide hydrolase 2, cytoplasmic
76	*MAT2B*	490	1.1e-08	5	methionine adenosyltransferase II, beta
77	*C6orf223*	491	2.7e-08	6	chromosome 6 open reading frame 223
78	*RARS*	492	2e-07	5	arginyl-tRNA synthetase
79	*BCAP29*	496	6.4e-07	7	B-cell receptor-associated protein 29
80	*BBS5*	502	3.8e-06	2	Bardet-Biedl syndrome 5
81	*DONSON*	510	2.1e-05	21	downstream neighbor of SON
82	*TEX14*	543	0.00083	17	testis expressed 14
83	*FAM21C*	547	0.00097	10	family with sequence similarity 21, member C
84	*L3MBTL2*	560	0.0014	22	l(3)mbt-like 2 (Drosophila)
85	*CLCN3*	562	0.0017	4	chloride channel, voltage-sensitive 3
86	*HFE*	575	0.004	6	hemochromatosis
87	*SRSF1*	586	0.0072	17	serine/arginine-rich splicing factor 1
88	*CCNB1*	587	0.0073	5	cyclin B1
89	*SLC30A1*	588	0.0074	1	solute carrier family 30 (zinc transporter), member 1
90	*POLR3B*	591	0.0085	12	polymerase (RNA) III (DNA directed) polypeptide B
91	*C9orf40*	594	0.0089	9	chromosome 9 open reading frame 40
92	*NFU1*	595	0.0093	2	NFU1 iron-sulfur cluster scaffold homolog (S. cerevisiae)
93	*KPNB1*	597	0.0094	17	karyopherin (importin) beta 1
94	*BLOC1S1*	607	0.012	12	biogenesis of lysosomal organelles complex-1, subunit 1
95	*LOC100132215*	614	0.014	2	uncharacterized LOC100132215

Table 6.4: *Prognostic signature, cluster 'hyper 2' (top 95 genes)*

	Symbol	Rank in prog. sig.	q-val	Chr	Info
1	*ZNF727*	12	0	7	zinc finger protein 727
2	*PDE1B*	19	0	12	phosphodiesterase 1B, calmodulin-dependent
3	*BK250D10*	23	0		
4	*ANKRD13B*	34	0	17	ankyrin repeat domain 13B
5	*IL17REL*	38	0	22	interleukin 17 receptor E-like
6	*GPC5*	40	0	13	glypican 5
7	*ZNF300P1*	59	0	5	zinc finger protein 300 pseudogene 1
8	*EGR2*	66	0	10	early growth response 2
9	*GPNMB*	70	0	7	glycoprotein (transmembrane) nmb
10	*B4GALNT1*	71	0	12	beta-1,4-N-acetyl-galactosaminyl transferase 1
11	*KCNJ9*	88	0	1	potassium inwardly-rectifying channel, subfamily J, member 9
12	*LTF*	102	0	3	lactotransferrin
13	*FSCN2*	105	0	17	fascin homolog 2, actin-bundling protein, retinal (Strongylocentrotus purpuratus)
14	*GPRIN1*	112	0	5	G protein regulated inducer of neurite outgrowth 1
15	*ZDHHC22*	116	0	14	zinc finger, DHHC-type containing 22
16	*RGR*	128	0	10	retinal G protein coupled receptor
17	*UFD1L*	130	0	22	ubiquitin fusion degradation 1 like (yeast)
18	*AQP2*	147	0	12	aquaporin 2 (collecting duct)
19	*LPAR5*	156	0	12	lysophosphatidic acid receptor 5
20	*ECEL1*	163	0	2	endothelin converting enzyme-like 1
21	*CSF1R*	202	0	5	colony stimulating factor 1 receptor
22	*RALGAPA2*	211	0	20	Ral GTPase activating protein, alpha subunit 2 (catalytic)
23	*KHDRBS2*	245	0	6	KH domain containing, RNA binding, signal transduction associated 2
24	*TMEM26*	258	2.2e-304	10	transmembrane protein 26
25	*RQCD1*	263	4.3e-287	2	RCD1 required for cell differentiation1 homolog (S. pombe)
26	*ADAMTSL3*	269	5.3e-239	15	ADAMTS-like 3
27	*KCNS1*	290	1.7e-188	20	potassium voltage-gated channel, delayed-rectifier, subfamily S, member 1
28	*HOXA9*	297	1.4e-167	7	homeobox A9
29	*LECT1*	317	4.7e-139	13	leukocyte cell derived chemotaxin 1
30	*EPHX3*	330	5.5e-121	19	epoxide hydrolase 3
31	*C12orf56*	360	1.5e-77	12	chromosome 12 open reading frame 56
32	*CDO1*	368	2.5e-69	5	cysteine dioxygenase, type I
33	*LOC644172*	382	8.9e-57	17	mitogen-activated protein kinase 8 interacting protein 1 pseudogene
34	*FAM198B*	397	1.1e-44	4	family with sequence similarity 198, member B
35	*UBE2QL1*	427	8.3e-30	5	ubiquitin-conjugating enzyme E2Q family-like 1
36	*NPY*	441	1.5e-23	7	neuropeptide Y
37	*KLB*	503	5e-06	4	klotho beta
38	*PIF1*	516	8e-05	15	PIF1 5'-to-3' DNA helicase homolog (S. cerevisiae)
39	*SLC10A4*	527	0.00038	4	solute carrier family 10 (sodium/bile acid cotransporter family), member 4
40	*ANXA6*	558	0.0013	5	annexin A6
41	*LINC00271*	565	0.0022	6	long intergenic non-protein coding RNA 271
42	*TSPAN32*	625	0.017	11	tetraspanin 32
43	*TRPM8*	647	0.031	2	transient receptor potential cation channel, subfamily M, member 8
44	*SPHKAP*	656	0.034	2	SPHK1 interactor, AKAP domain containing
45	*RCAN2*	676	0.048	6	regulator of calcineurin 2

Table 6.5: *Prognostic signature, cluster 'hypo 1'*

	Symbol	Rank in prog. sig.	q-val	Chr	Info
1	*NOX4*	74	0	11	NADPH oxidase 4
2	*CD74*	84	0	5	CD74 molecule, major histocompatibility complex, class II invariant chain
3	*DYRK1A*	167	0	21	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A
4	*CPE*	180	0	4	carboxypeptidase E
5	*MT1F*	267	1.7e-241	16	metallothionein 1F
6	*SPEF2*	276	4.9e-219	5	sperm flagellar 2
7	*NFKBIZ*	287	1e-191	3	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta
8	*CDR2*	295	3.1e-171	16	cerebellar degeneration-related protein 2, 62kDa
9	*GINS1*	302	3.8e-162	20	GINS complex subunit 1 (Psf1 homolog)
10	*COPZ1*	316	9.9e-141	12	coatomer protein complex, subunit zeta 1
11	*CCDC110*	338	5e-112	4	coiled-coil domain containing 110
12	*NEK9*	393	1.2e-45	14	NIMA-related kinase 9
13	*TMEM170B*	417	1.1e-33	6	transmembrane protein 170B
14	*NRIP3*	448	8e-21	11	nuclear receptor interacting protein 3
15	*NAAA*	464	3.4e-14	4	N-acylethanolamine acid amidase
16	*PPAT*	501	2e-06	4	phosphoribosyl pyrophosphate amidotransferase
17	*VWDE*	579	0.005	7	von Willebrand factor D and EGF domains
18	*SLC16A11*	642	0.027	17	solute carrier family 16, member 11 (monocarboxylic acid transporter 11)
19	*JHDM1D*	652	0.033	7	jumonji C domain containing histone demethylase 1 homolog D (S. cerevisiae)

Table 6.6: *Prognostic signature, cluster 'hypo 2'*

Chapter 7

Detection of Epigenomic Network Community Oncomarkers

7.1 Introduction

In this chapter, I present a DNA methylation-based measure of genomic interaction and association, and I show how to use it to infer prognostic genomic networks. I then show how to identify prognostic biomarkers from such networks, which I term 'network community oncomarkers'. As a cancer progresses, its signalling and control networks are re-arranged ('re-wired'), and this drives adaptive alterations in phenotype, which are advantageous for the cancer (Barabási & Oltvai, 2004). Previous research by other authors (Taylor *et al.* , 2009) found that patient survival outcome in breast cancer could be predicted well by network models of this re-wiring, based on gene expression data. It has been previously shown that DNA methylation can serve as a surrogate for activity at genomic-regulatory regions (Brocks *et al.* , 2014). Hence, DNA methylation measurements are a natural basis from which to construct genomic regulatory and associated networks, and such networks inferred from DNA methylation data are a promising basis for prognostic biomarkers.

The DNA methylation-based measure of interaction or association between pairs of genes which I present in this chapter is called the 'DNA methylation network interaction measure'. In the genomic networks which it is used to infer, an edge between a pair of genes/nodes indicates that the interaction or association between those genes is associated with disease progression. I show how to identify prognostic biomarkers from such networks using community detection to identify subnetwork modules within the network. These communities are groups of nodes/genes amongst which there is a high density of prognostic interactive or associative behaviour, and I term them 'network community oncomarkers'. I show that within these communities, the DNA methylation network interaction measure is highly associated with co-regulatory behaviour linked to gene expression (at the mRNA level), giving functional relevance

to the findings. Each network community oncomarker can be used to calculate a one-number prognostic score for each patient, based on DNA methylation data alone.

7.2 Methods and models

An overview of the methods presented here appears in Figure 7.1, following which the component parts of this methodology are presented in detail.

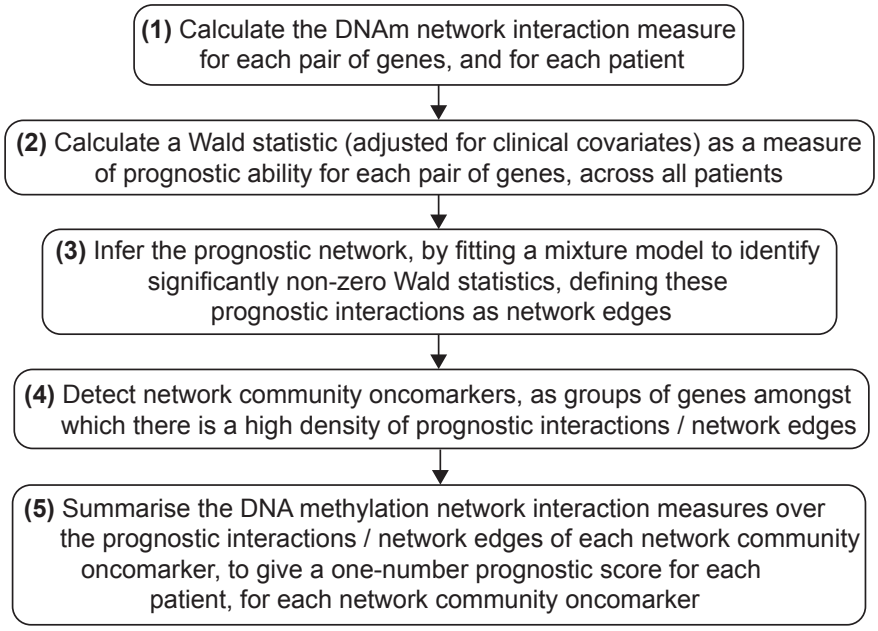


Figure 7.1: *Overview of methods.*

7.2.1 DNA methylation network interaction measure

DNA methylation is a chemical modification to DNA, which may occur at numerous locations within a gene, typically at CpG di-nucleotides. Hence, the pattern of these modifications within a gene forms a 'DNA methylation profile'. Using canonical correlation analysis (CCA) (Hotelling, 1936), I have developed a novel statistical measure (Bartlett *et al.* , 2014), of the level of interaction or association between a pair of genes (network nodes) in a single sample/patient, based on DNA methylation profiles (Figure 7.2). This DNA methylation network interaction measure quantifies the extent to which the DNA methylation profiles of a pair of genes explain each other. It is based only on measurements of the DNA methylation profiles of that pair of genes, and it acts as a surrogate for a measure of the extent to which this pair of genes behave interactively or associatively. Such behaviour may include transcriptional regulation or co-regulation, or other types of biochemical interaction, influencing gene expression levels, isoforms and the presence of alternatively spliced gene products, amongst other phenomena (Jones, 2012).

The DNA methylation network interaction measure is defined by analogy to CCA. CCA

aims to discover linear combinations of variables of one type, and linear combinations of variables of another type, so that these combinations best explain each other. In this context, a particular way of combining (by scaling and adding) the deviations from the mean methylation profile at a number of locations within one gene might be particularly effective at explaining a particular combination (again, by scaling and adding) of the deviations from the mean methylation profile at a number of locations in another gene, and *vice-versa*. There will probably be fewer ways in which the methylation levels of these genes covary across the samples, than there are locations at which methylation is measured along the genes; this is because the methylation level is highly correlated at many locations along a particular gene. CCA finds the most important components of this covariation across samples.

CCA seeks to find the vectors a and b, in the p and q dimensional spaces of variables $\mathbf{X} = (x_1, x_2, ..., x_p)'$ and $\mathbf{Y} = (y_1, y_2, ..., y_q)'$ respectively, which maximise the correlation $\rho = \text{cor}\,(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y})$, defined according to equation 7.1:

$$\rho = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{XX}\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b}}}, \tag{7.1}$$

where $\boldsymbol{\Sigma}_{XX} = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)'\right]$ and $\boldsymbol{\Sigma}_{YY} = \mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\mu}_Y)(\mathbf{Y} - \boldsymbol{\mu}_Y)'\right]$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and $\boldsymbol{\Sigma}_{XY} = \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{Y} - \boldsymbol{\mu}_Y)'\right]$ is the cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$.

Two genes X and Y have corresponding methylation profiles which are measured for sample / patient k at p and q CpGs (loci) respectively along these genes. Denoting these measurements by the variables $x_1, ...x_p$ and $y_1, ..., y_q$ for genes X and Y respectively, the DNA methylation profiles for these genes, for patient k, can be represented by the vectors $\mathbf{x}(k)$ and $\mathbf{y}(k)$, which have p and q entries respectively. A measure of DNA methylation network interaction $\rho_{XY}(k)$, of the methylation profiles of genes X and Y for sample k, can then be defined by analogy with equation 7.1, according to equation 7.2:

$$\rho_{XY}(k) = \frac{\mathbf{x}(k)^T\hat{\boldsymbol{\Sigma}}^{(h)}_{XY}\mathbf{y}(k)}{\sqrt{\mathbf{x}(k)^T\hat{\boldsymbol{\Sigma}}^{(h)}_{XX}\mathbf{x}(k)}\sqrt{\mathbf{y}(k)^T\hat{\boldsymbol{\Sigma}}^{(h)}_{YY}\mathbf{y}(k)}}, \tag{7.2}$$

where $\hat{\boldsymbol{\Sigma}}^{(h)}_{XX}$, $\hat{\boldsymbol{\Sigma}}^{(h)}_{YY}$ and $\hat{\boldsymbol{\Sigma}}^{(h)}_{XY}$ are estimated from healthy rather than cancer samples in the methylation data set, according to equations 7.3 - 7.5,

$$\hat{\boldsymbol{\Sigma}}^{(h)}_{XX} = \frac{1}{n_h}\sum_{k\in\text{healthy}}\left(\mathbf{x}(k) - \hat{\boldsymbol{\mu}}^{(h)}_X\right)\left(\mathbf{x}(k) - \hat{\boldsymbol{\mu}}^{(h)}_X\right)^T, \tag{7.3}$$

$$\hat{\mathbf{\Sigma}}_{YY}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \left(\mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)} \right) \left(\mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)} \right)^T, \tag{7.4}$$

$$\hat{\mathbf{\Sigma}}_{XY}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \left(\mathbf{x}(k) - \hat{\boldsymbol{\mu}}_X^{(h)} \right) \left(\mathbf{y}(k) - \hat{\boldsymbol{\mu}}_Y^{(h)} \right)^T, \tag{7.5}$$

where

$$\hat{\boldsymbol{\mu}}_X^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \mathbf{x}(k),$$

and

$$\hat{\boldsymbol{\mu}}_Y^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} \mathbf{y}(k),$$

and n_h is the number of healthy samples in the data set. When the DNA methylation network interaction measure $\rho_{XY}(k)$ is large (i.e., close to 1), the corresponding pair of genes explain each other's transcriptional or translational behaviour (as reflected in their methylation profiles) well, or have otherwise well-correlated interactive or associative behaviour, for sample/patient k. Hence, $\rho_{XY}(k)$ measures (according to their DNA methylation profiles) the level of interaction or association between genes X and Y in tumour sample k, compared to typical interactions between these genes in healthy tissue.

7.2.2 Prognostic network construction

To identify network oncomarkers, I consider a prognostic interaction network for m genes. This network is represented by the $m \times m$ adjacency matrix $\mathbf{A}$, in which an edge is defined to to be present (i.e., $A_{ij} = 1$) if and only if the corresponding pair of genes (nodes) are prognostic according to the DNA methylation network interaction measure. Otherwise, I set $A_{ij} = 0$; N.B., i and j are now redefined compared to the last section, so that they index genes rather than methylation loci; this will not be problematic, because all subsequent analysis is carried out at the level of genes rather than methylation loci. To identify these prognostic edges, for each of the $\binom{m}{2}$ pairs of genes in the network, I use the Cox proportional hazards model (Cox, 1972) to calculate a Wald-statistic z_{ij}. This quantifies the association of the DNA methylation network interaction measure ρ_{ij} for the pair of genes i and j ($i = 1, ..., m$ and $j = 1, ..., m$) with patient survival outcome across patients k ($k = 1, ...n$). I use a multivariate Cox model, and hence these Wald statistics are adjusted for clinical covariates, in order to detect novel DNA methylation biomarkers which are independent of known prognostic clinical features.

The Wald statistic is asymptotically normally distributed with unit variance (Harrell, 2001). We can therefore model the distribution of our observed Wald statistics, z_{ij}, as a mixture

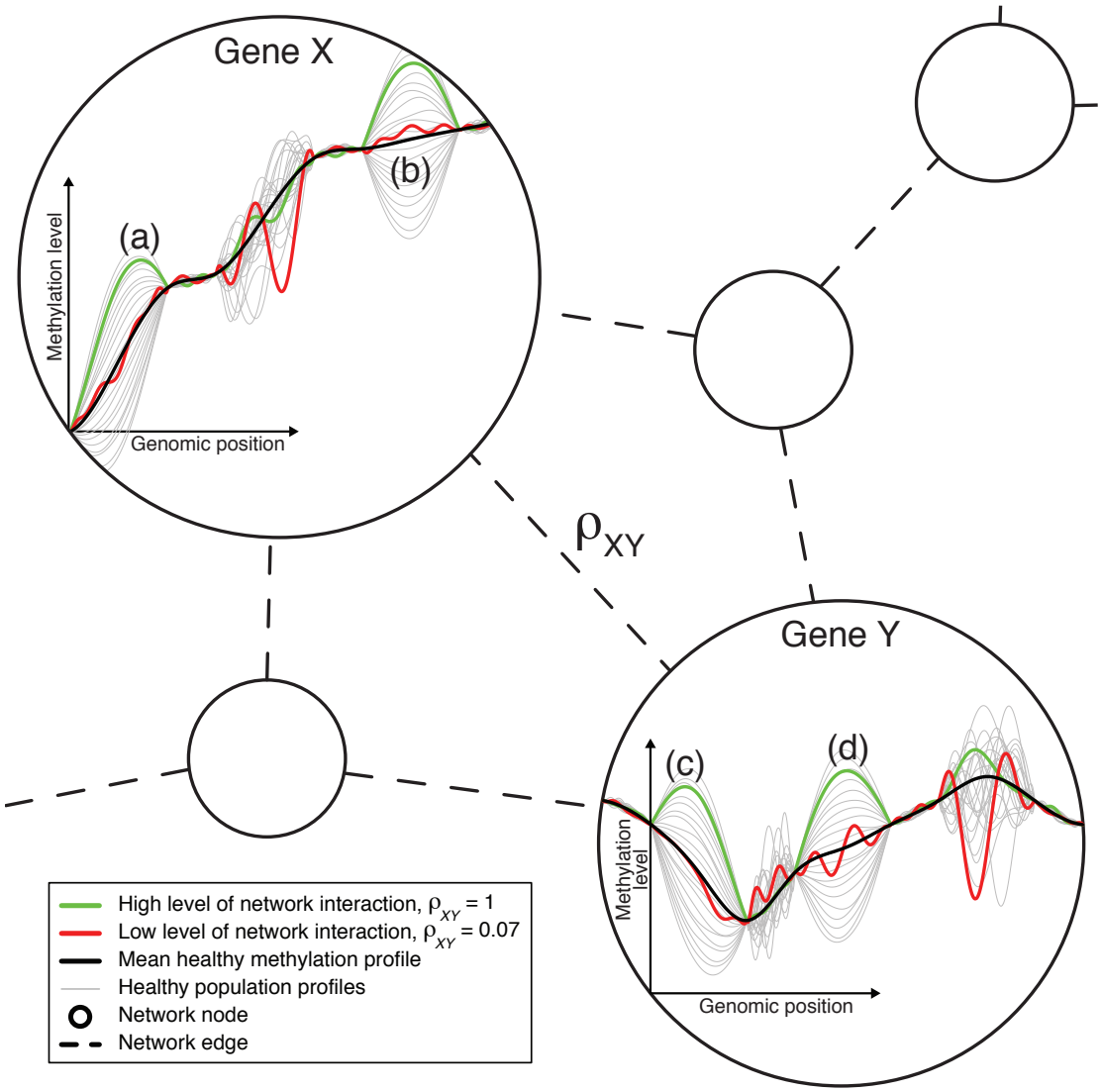


Figure 7.2: *The DNA methylation network interaction measure.*
A combination of the variation of the healthy methylation profiles in regions (a) and (b) of gene X explains well / is well-explained by a combination of the variation of the healthy methylation profiles in regions (c) and (d) of gene Y. The green cancer sample varies by a large amount about the mean methylation profile and in a typical way in these regions in both genes. Hence, the green sample corresponds to a high level of network interaction for this sample, $\rho_{XY} = 1$. The equivalent variations in the other regions of these genes do not explain each other well, and so the red sample, which varies by a large amount in these other regions and varies less and in an atypical way in regions (a) - (d), corresponds to a low level of network interaction, $\rho_{XY} = 0.07$. Genes X and Y are likely to have different numbers of methylation measurement locations (i.e., variables X and Y are of different dimension). The ordering of the measurement locations has no influence on the calculation of ρ, as long as the ordering is consistent across samples.

of Gaussians, as shown in chapter 4:

$$z_{ij} \sim \begin{cases} \mathcal{N}\left(\mu_{ij}, \sigma^2\right), & \text{if} \quad A_{ij} = 1, \\ \mathcal{N}\left(0, \sigma^2\right), & \text{if} \quad A_{ij} = 0, \end{cases} \tag{7.6}$$

where $\mathcal{N}\left(\mu_{ij}, \sigma^2\right)$ is the normal distribution, and $\sigma^2 = 1$. Hence, I fit a mixture model to

each observed statistic z_{ij}, and then infer whether, given z_{ij}, it is more likely that $\mu_{ij} = 0$, or $\mu_{ij} \neq 0$, leading to the estimates $\hat{A}_{ij} = 0$ or $\hat{A}_{ij} = 1$ respectively. I fit this model using the empirical Bayes procedure of (Johnstone & Silverman, 2004), defining a mixture prior distribution $f_{\text{prior}}(\mu_{ij})$ over the μ_{ij} of equation 7.6:

$$f_{\text{prior}}(\mu_{ij}) = (1-w)\,\delta(\mu_{ij}) + w\gamma(\mu_{ij}), \tag{7.7}$$

where w is the mixing parameter between the two components, which can also be interpreted as $w = \mathbb{E}[p(A_{ij} = 1)]$, and $\gamma(\cdot|a)$ is the Laplace probability density function,

$$\gamma(\mu_{ij}|a) = \frac{a}{2}\exp(-a|\mu_{ij}|),$$

where I use $a = 0.5$, as in (Johnstone & Silverman, 2004). Taking the mixture components to have Gaussian likelihoods, $f_{\mathcal{N}}\left(\cdot\middle|\mu_{ij}, \sigma^2\right)$, as in equation 7.6, it follows from equation 7.7 that the posterior density over the observed prognostic Wald statistic z_{ij} is:

$$f_{\text{posterior}}(\mu_{ij}|z_{ij}) = \frac{(1-w)\,\delta(\mu_{ij})\,f_{\mathcal{N}}\left(z_{ij}\middle|0,\sigma^2\right) + w\gamma(\mu_{ij})\,f_{\mathcal{N}}\left(z_{ij}\middle|\mu_{ij},\sigma^2\right)}{f_{\text{marginal}}(z_{ij})}, \tag{7.8}$$

where the marginal density is:

$$f_{\text{marginal}}(z_{ij}) = (1-w) f_{\mathcal{N}}\left(z_{ij}\middle|0,\sigma^2\right) + w g(z_{ij}), \tag{7.9}$$

where $g(\mu_{ij})$ is the convolution of the Laplace density with the standard normal density. If the Laplace distribution in the prior, equation 7.7, were replaced with a Gaussian, then the marginal distribution, equation 7.9, would be a mixture of Gaussians. However, as noted in (Johnstone & Silverman, 2004), this empirical Bayes procedure requires a prior with tails that are exponential or heavier. Hence, I similarly use the Laplace rather than Gaussian prior, in this practical implementation, which is a slight model mis-specification.

Although a separate model is fitted to each observed Wald statistic z_{ij}, a common weight w_i is used for each gene/node i. This estimate of w_i is found as the value which maximises the marginal likelihood (equation 7.10) of the observed statistics z_{ij} over all the pairwise comparisons of i with j, $j \neq i$. This allows the model for each such pairwise comparison (i, j) to 'borrow strength' from all the other comparisons (i, j'), $j' \neq i$, $j' \neq j$:

$$\hat{w}_i = \arg\max_{w} \sum_{j \neq i} \log\left\{(1-w)\phi(z_{ij}) + w g(z_{ij})\right\}. \tag{7.10}$$

For a particular gene i, if the z_{ij} are mostly close to zero, then w_i will be set low, which means that fewer edges ($A_{ij} = 1$) will be detected; this hence corresponds to i being a low-degree node. If for a different gene i the z_{ij} are generally further from zero, then $\hat{w}_i$ will be set high, which corresponds more edges being detected; this hence corresponds to i being a high-degree node. Therefore, setting $\hat{w}_i$ separately for each gene i allows adaptation to a heterogenous degree distribution in $\mathbf{A}$. As in (Johnstone & Silverman, 2004), I use the posterior median to obtain the estimate $\hat{\mu}_{ij}$. Then I make a conservative estimate of $\mathbf{A}$ as follows:

$$\begin{aligned} \hat{A}_{ij} =& 1 \quad \text{if} \quad \hat{\mu}_{ij} > 0 \text{ and } \hat{\mu}_{ji} > 0 \quad \text{or} \quad \hat{\mu}_{ij} < 0 \text{ and } \hat{\mu}_{ji} < 0, \\ \hat{A}_{ij} =& 0 \quad \text{otherwise.} \end{aligned} \tag{7.11}$$

7.2.3 Community and oncomarker detection

Network nodes can be grouped together according to their propensity to interact with each other, for example groups of friends in a social network, or functional subnetwork modules in a biological network. This statistical method is referred to as community detection (Girvan & Newman, 2002; Newman, 2004). Hence, community detection allows us to find groups of genes in our constructed prognostic network, which interact differently in cancer than in healthy tissue, in a way which is predictive of how advanced the disease is. I term these 'network community oncomarkers'. Within such a detected network community oncomarker, the genes may interact with each other more (relative to healthy tissue) the more serious the disease is (as in Figure 7.6c), or they may interact with each other less the more serious the disease is, (as in Figure 7.6a). I carry out community detection by fitting the degree-corrected stochastic blockmodel (Holland *et al.* , 1983; Bickel & Chen, 2009), by regularised spectral clustering (Qin & Rohe, 2013). I calculate the optimum number of communities to divide the network into using the network histogram method (Olhede & Wolfe, 2014). Each community, or subnetwork module, identified in this way represents a potential network community oncomarker.

For each network community oncomarker, a prognostic score can be calculated for each patient, by summarising the DNA methylation network interaction measure over that community. This prognostic score can be used as a one-number summary of disease prognosis for that patient, according to the network community oncomarker. Some gene-gene interactions will, with worse prognosis, correspond to increasingly negative DNA methylation network interaction measure ρ_{ij} (such as increased inhibitory gene regulation). Whereas some gene-gene interactions will, with worse prognosis, correspond to increasingly positive ρ_{ij} (such as increased activatory gene regulation). This means that care must be taken when summarising the

network interaction measure across the network community oncomarker. Further, the magnitude of the changes in the network interaction measure may be different for different prognostic pairs of genes, for the same amount of prognostic information conveyed. To address these points, I combine the ρ_{ij} across the prognostic pairs of genes of the network community after first multiplying them by the corresponding fitted Cox proportional hazards model coefficients $\hat{\theta}_{ij}$, obtained as described at the start of Section 7.2.2. Under the Cox proportional hazards model, the fitted model coefficient $\hat{\theta}_{ij}$ for a predictor ij gives the log of the hazard-ratio (HR) for that predictor in the model, i.e., $\log(\text{HR}_{ij}) = \hat{\theta}_{ij}$. The hazard ratio is the scale-factor increase in probability of an event (e.g., death) occurring per unit time, relative to the baseline hazard (e.g., compared to a control group). Hence, these coefficients are interpretable in the same way, without scaling issues, across fitted models. This means that, for patient k, we can combine the DNA methylation network interaction measures over a network community oncomarker to generate a one-number prognostic score, as follows:

$$\text{Score}_k = \sum_{i \in C, j \in C, i<j} \hat{A}_{ij} \hat{\theta}_{ij} \rho_{ij}(k),$$

where C is the set of nodes in the network community oncomarker, $\hat{\mathbf{A}}$ is the inferred adjacency matrix, $\rho_{ij}(k)$ is the DNAm network interaction measure for genes/nodes i and j and patient k, and $\hat{\theta}_{ij}$ is the corresponding fitted Cox multivariate proportional-hazards model coefficient. Network edges/DNA methylation network interaction measures ρ_{ij} which increase with poor prognosis (i.e., pairs of genes which interact more as the disease progresses, coloured green in Figure 7.6), will will correspond to $\hat{\theta}_{ij} > 0$. Hence, an increase in such a ρ_{ij} will increase the prognostic score. Equivalently, network edges/DNA methylation network interaction measures ρ_{ij} which decrease with poor prognosis (i.e., pairs of genes which interact less as the disease progresses, coloured red in Figure 7.6), will will correspond to $\hat{\theta}_{ij} < 0$. Hence, a decrease in such a ρ_{ij} will also increase the prognostic score.

7.2.4 An equivalent gene-expression interaction measure

To examine further the hypothesis that the DNA methylation network interaction measure is a reflection of co-regulatory or co-regulated gene-expression patterns (amongst other genomic effects), we need an equivalent measure of gene-gene interaction or association in terms of gene expression. We can calculate such a measure, $\rho^{\text{expr}}_{XY}(k)$, for gene expression measurements

$x^{\text{expr}}(k)$ and $y^{\text{expr}}(k)$ for the genes X and Y and patient k, as follows (equation 7.12):

$$\rho_{XY}^{\text{expr}}(k) = \frac{\left(x^{\text{expr}}(k) - \hat{\mu}_{x^{\text{expr}}}^{(h)}\right)}{\hat{\sigma}_{x^{\text{expr}}}^{(h)}} \cdot \frac{\left(y^{\text{expr}}(k) - \hat{\mu}_{y^{\text{expr}}}^{(h)}\right)}{\hat{\sigma}_{y^{\text{expr}}}^{(h)}} \tag{7.12}$$

where

$$\hat{\mu}_{x^{\text{expr}}}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} x^{\text{expr}}(k) \quad \text{and} \quad \hat{\mu}_{y^{\text{expr}}}^{(h)} = \frac{1}{n_h} \sum_{k \in \text{healthy}} y^{\text{expr}}(k),$$

$$\left(\hat{\sigma}_{x^{\text{expr}}}^{(h)}\right)^2 = \frac{1}{n_h} \sum_{k \in \text{healthy}} \left(x^{\text{expr}}(k) - \hat{\mu}_{x^{\text{expr}}}^{(h)}\right)^2 \quad \text{and} \quad \left(\hat{\sigma}_{y^{\text{expr}}}^{(h)}\right)^2 = \frac{1}{n_h} \sum_{k \in \text{healthy}} \left(y^{\text{expr}}(k) - \hat{\mu}_{y^{\text{expr}}}^{(h)}\right)^2.$$

The intuition of equation 7.12 is that when the gene expression measurements $x^{\text{expr}}(k)$ and $y^{\text{expr}}(k)$ deviate *in the same sample* from the corresponding healthy mean expression levels, this measure will be non-zero. When this occurs in the same samples as the DNAm network interaction measure $\rho_{XY}(k)$ is also non-zero, we will see a correlation between $\rho_{XY}(k)$ and ρ_{XY}^{expr}. These interaction measures for methylation and expression, $\rho_{XY}(k)$ and ρ_{XY}^{expr}, are equivalent because they both measure deviation from typical interactive behaviour in healthy/control samples. I note that, that while ρ_{XY}^{expr} works satisfactorily for this comparison, it would not be expected to be a sensitive statistic to use as a prognostic tool.

7.3 Results

I present the results of the described methodology, to a breast cancer invasive carcinoma (BRCA) data-set, downloaded from the Cancer Genome Atlas (TCGA). I downloaded an initial batch of DNA methylation data for tumour samples from 175 samples/individuals (the training set), together with clinical data relating to patient survival outcome, and the covariates age, disease stage, and residual disease. I also downloaded corresponding DNA methylation data for healthy tissue for 98 individuals, which define the reference DNA methylation profiles. These data were used to detect potential network community oncomarkers. I then downloaded DNA methylation data for a further 528 tumour samples (the test set), together with data for the same clinical features: these independent samples were used to validate the potential network community oncomarkers. I also downloaded gene expression data for 216 of the tumours for which DNAm data were also available.

I inferred the binary prognostic adjacency matrix $\mathbf{A}$ for the 175 samples of the BRCA training data set according to the methods described. DNAm data were available for 14829 genes, and hence the number of nodes/genes m in the inferred adjacency matrix, $\hat{\mathbf{A}}$, is $m = 14829$. The presence of an edge in $\hat{\mathbf{A}}$, i.e., $\hat{A}_{ij} = 1$, implies that the interaction between

genes i and j is, according to the DNA methylation network interaction measure, associated with disease progression. The edge density of $\hat{\mathbf{A}}$ is 0.0035, i.e., $p(\hat{A}_{ij} = 1) = 0.0035$. I then extracted the connected component from this inferred network, and on this carried out community detection as described, resulting in 33 communities, ranging from 116 to 285 nodes in size. The reduced adjacency matrix relating to these communities, with $m = 5668$ and $p(\hat{A}_{ij} = 1) = 0.023$, is shown in Figure 7.3.

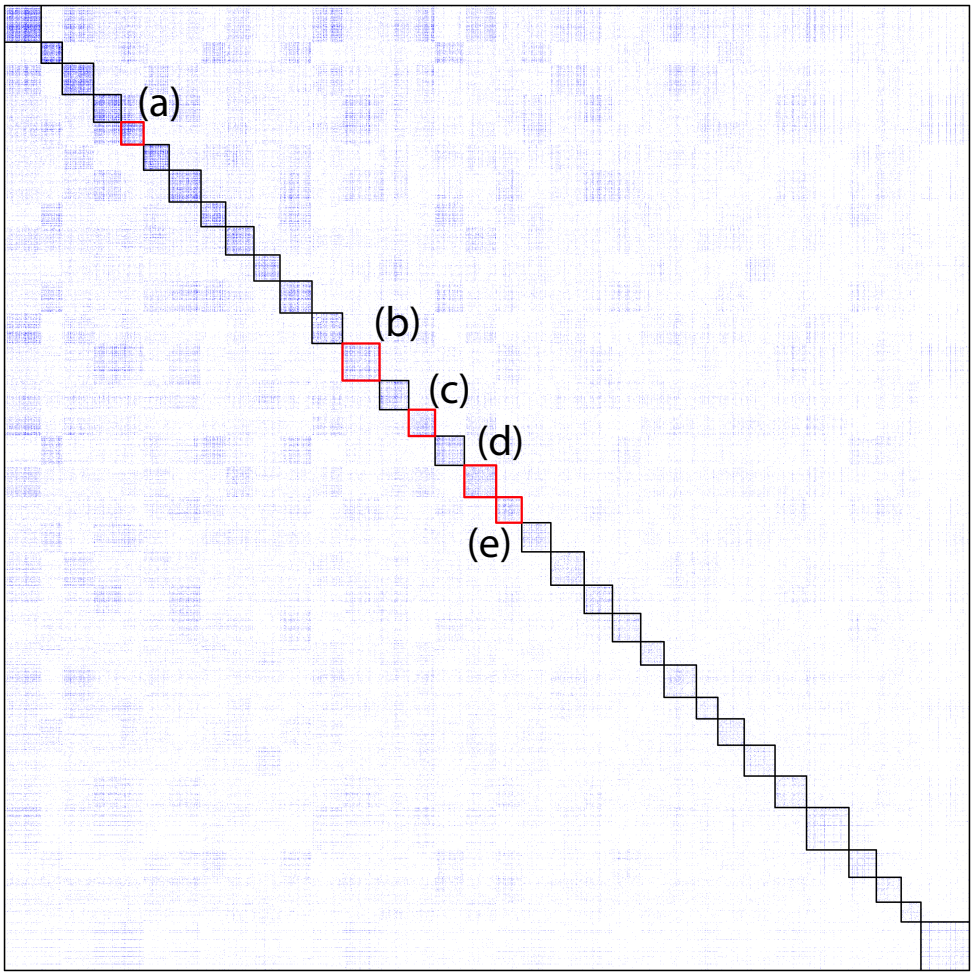


Figure 7.3: *The inferred adjacency matrix, after community detection. Entries in the adjacency matrix equal to 1 (representing a network edge) are coloured blue. Detected communities are outlined. The potential network community oncomarkers which are analysed further in Figures 7.4 - 7.7 and Tables 7.1 - 7.2 and 7.3 - 7.7 are indicated in red, and labelled (a) - (e).*

I validated each of these 33 potential network community oncomarkers in the independent 528 tumour samples of the test/validation set. I note that these 528 samples were not used in any way to identify the 33 potential network community oncomarkers shown in Figure 7.3. To carry out this validation, I calculated the prognostic score for the 528 independent/unseen samples of the test set, based on the inferred adjacency matrix $\hat{\mathbf{A}}$ and the fitted Cox multivariate proportional hazards model coefficients $\hat{\boldsymbol{\theta}}$ derived from the initial 175 samples of the training

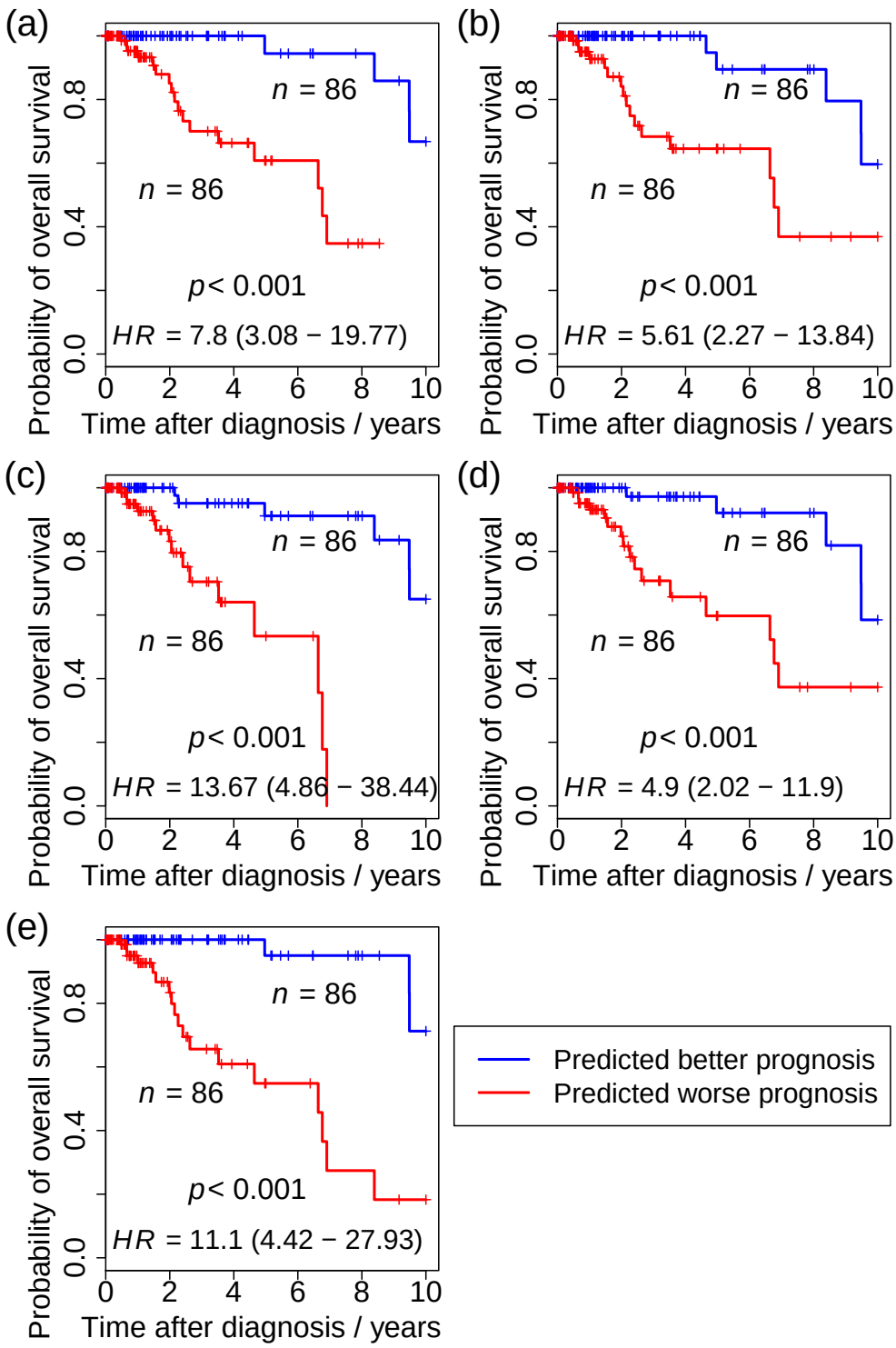


Figure 7.4: *Network community oncomarkers: Kaplan-Meier plots for the training set. Comparison of survival curves for the patient groups defined by the prognostic score for each network community oncomarker. The groups are divided by the median prognostic score in the 175 samples of the initial training data set. The hazard ratio (HR) is displayed with 95% C.I. in brackets, with the corresponding p-value calculated by univariate Cox regression. (a) - (e) indicate network community oncomarkers 1 - 5, as shown in Figure 7.3.*

set. I calculated one prognostic score for each potential network community oncomarker for each of the 528 unseen test-set samples. I then tested the prognostic score, for each potential

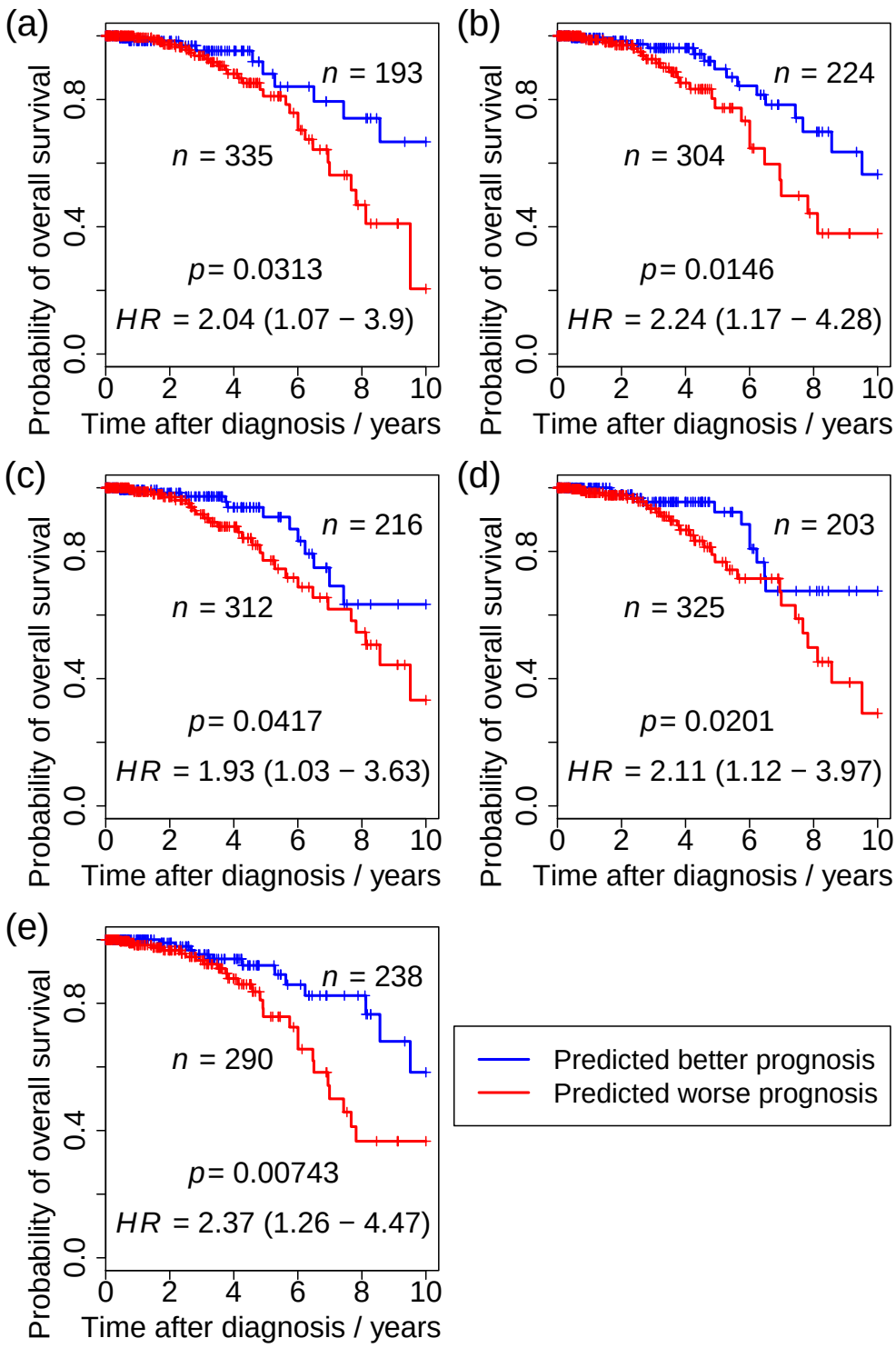


Figure 7.5: *Network community oncomarkers: Kaplan-Meier plots for the test / validation set. Comparison of survival curves for the patient groups defined by the prognostic score for each network community oncomarker. The groups are divided by the median prognostic score in the 175 samples of the initial training data set. The hazard ratio (HR) is displayed with 95% C.I. in brackets, with the corresponding p-value calculated by univariate Cox regression. (a) - (e) indicate network community oncomarkers 1 - 4, as shown in Figure 7.3.*

network community oncomarker, for association with patient survival outcome in these 528 unseen test-set samples. The five potential network community oncomarkers which validated

most significantly in this way are outlined in red in Figure 7.3. The results of univariate and multivariate Cox regression for these five best network community oncomarkers are shown in Figures 7.4 and 7.5, and in Tables 7.1 and 7.2, for the training and test sets respectively. For the multivariate analysis, samples with missing data for any of the clinical covariates were removed, leaving 172 and 396 samples for the training and test sets, respectively.

	HR (95%CI)	*p*	*n*
Prognostic Score	77.1 (10.5-567)	<0.001	172
Age	1.79 (0.66-4.84)	0.249	172
Residual Disease	15.4 (4.68-50.9)	<0.001	172
Stage	2.85 (0.96-8.46)	0.060	172

(a) Network community oncomarker 1.

	HR (95%CI)	*p*	*n*
Prognostic Score	51.3 (8.35-315)	<0.001	172
Age	1.42 (0.48-4.23)	0.53	172
Residual Disease	30.4 (5.82-158)	<0.001	172
Stage	1.95 (0.68-5.54)	0.212	172

(b) Network community oncomarker 2.

	HR (95%CI)	*p*	*n*
Prognostic Score	50.1 (9.77-256)	<0.001	172
Age	2.16 (0.81-5.8)	0.125	172
Residual Disease	13.3 (4.54-39.1)	<0.001	172
Stage	2.41 (0.81-7.18)	0.114	172

(c) Network community oncomarker 3.

	HR (95%CI)	*p*	*n*
Prognostic Score	22.7 (5.52-93.1)	<0.001	172
Age	3.49 (1.3-9.42)	0.0135	172
Residual Disease	16.3 (5.24-50.7)	<0.001	172
Stage	1.05 (0.38-2.91)	0.928	172

(d) Network community oncomarker 4.

	HR (95%CI)	*p*	*n*
Prognostic Score	46.0 (8.17-259)	<0.001	172
Age	2.91 (1-8.44)	0.0493	172
Residual Disease	7.04 (2.68-18.5)	<0.001	172
Stage	3.74 (1.23-11.4)	0.02	172

(e) Network community oncomarker 5.

Table 7.1: *Network community oncomarkers - training set prognosis. Multivariate Cox regression was used to test significance of the prognostic scores derived from the network community oncomarkers. (a) - (e) indicate network community oncomarkers 1 - 5, as shown in Figure 7.3.*

Figure 7.6 shows the five network community oncomarkers which validated most significantly. Green edges indicate gene-gene interactions which become stronger with disease progression. Red edges indicate interactions which become weaker with disease progression. Hence, the network community oncomarkers of Figure 7.6c and 7.6d can be considered to be a functional subnetwork modules which becomes more active as the cancer progresses. On the other hand, the network community oncomarker of Figure 7.6a can be considered to be a functional subnetwork module which becomes less active as the cancer progresses. Then the network community oncomarkers of Figures 7.6b and 7.6c appear to contain a mixture of these effects. However, each of these network community oncomarkers represents a functional subnetwork module which is rewired in a way which is advantageous for the cancer, in favour of proliferation, and against cell death and immune function. The genes/nodes of these network community oncomarkers are shown in Tables 7.3 - 7.7; they list many genes related to cell proliferation (e.g., *CDKL1*, *NKAPL*, *MAPK6*), developmental processes (e.g., *HOXD10*, *HOXB9*, *HOXC10*, *HOXA13*, *HOXC12*, *HOXD13*), and immune function (e.g., *VSIG2*, *IL36B*, *RBPJ*).

	HR (95%CI)	*p*	*n*
Prognostic Score	4.89 (1.65-14.5)	0.00429	396
Age	3.52 (1.46-8.49)	0.00513	396
Residual Disease	12.5 (5.32-29.3)	<0.001	396
Stage	1.62 (0.66-4)	0.294	396

(a) Network community oncomarker 1.

	HR (95%CI)	*p*	*n*
Prognostic Score	5.07 (1.81-14.1)	0.00195	396
Age	3.67 (1.49-9.03)	0.00458	396
Residual Disease	8.72 (3.78-20.1)	<0.001	396
Stage	1.47 (0.6-3.61)	0.406	396

(b) Network community oncomarker 2.

	HR (95%CI)	*p*	*n*
Prognostic Score	2.63 (1.01-6.89)	0.0484	396
Age	2.07 (0.86-5)	0.106	396
Residual Disease	11.3 (4.97-25.5)	<0.001	396
Stage	2.04 (0.76-5.45)	0.157	396

(c) Network community oncomarker 3.

	HR (95%CI)	*p*	*n*
Prognostic Score	4.92 (1.8-13.5)	0.00189	396
Age	1.91 (0.78-4.69)	0.159	396
Residual Disease	17.2 (6.76-43.9)	<0.001	396
Stage	0.92 (0.34-2.48)	0.871	396

(d) Network community oncomarker 4.

	HR (95%CI)	*p*	*n*
Prognostic Score	2.5 (0.94-6.65)	0.0668	396
Age	2.23 (0.94-5.27)	0.0677	396
Residual Disease	8.17 (3.47-19.3)	<0.001	396
Stage	1.59 (0.64-3.95)	0.321	396

(e) Network community oncomarker 5.

Table 7.2: *Network community oncomarkers - test/validation set prognosis. Multivariate Cox regression was used to test significance of the prognostic scores derived from the network community oncomarkers. (a) - (e) indicate network community oncomarkers 1 - 5, as shown in Figure 7.3.*

I also examined further the hypothesis that the DNA methylation network interaction measure is a reflection of co-regulatory or co-regulated gene-expression patterns (amongst other genomic effects). I did this by comparing the DNA methylation network interaction measure ρ_{XY} for a pair of genes XY, equation 7.2, with an equivalent measure of interactive behaviour of these genes in terms of their expression levels, ρ_{XY}^{expr}, equation 7.12. Correlation test p-values for the comparison between ρ_{XY} and ρ_{XY}^{expr} appear in Figure 7.7. We see a concentration of significant p-values close to zero, indicating there is strong association between ρ_{XY} and ρ_{XY}^{expr}, for each network community oncomarker. However, there are also many non-significant p-values in these histograms, indicating that there are other genomic interactive effects present, which cannot be explained in terms of gene expression (as assessed by mRNA levels) alone. Such effects might include the influence of alternatively spliced products or isoforms (Jones, 2012), or the interaction between non-coding transcripts and the epigenome (Lai & Shiekhattar, 2014).

7.4 Discussion

In this chapter, I have presented a measure of pairwise interaction between genes, based on DNA methylation measurements. I have shown how to use this measure to infer prognostic genomic networks, and how it is possible to identify prognostic biomarkers from such networks, using community detection methodology. I call these 'network community oncomarkers'; they are groups of nodes/genes amongst which there is a high density of prognostic genomic interactive or associative behaviour. I have shown that within these communities, the DNA methylation

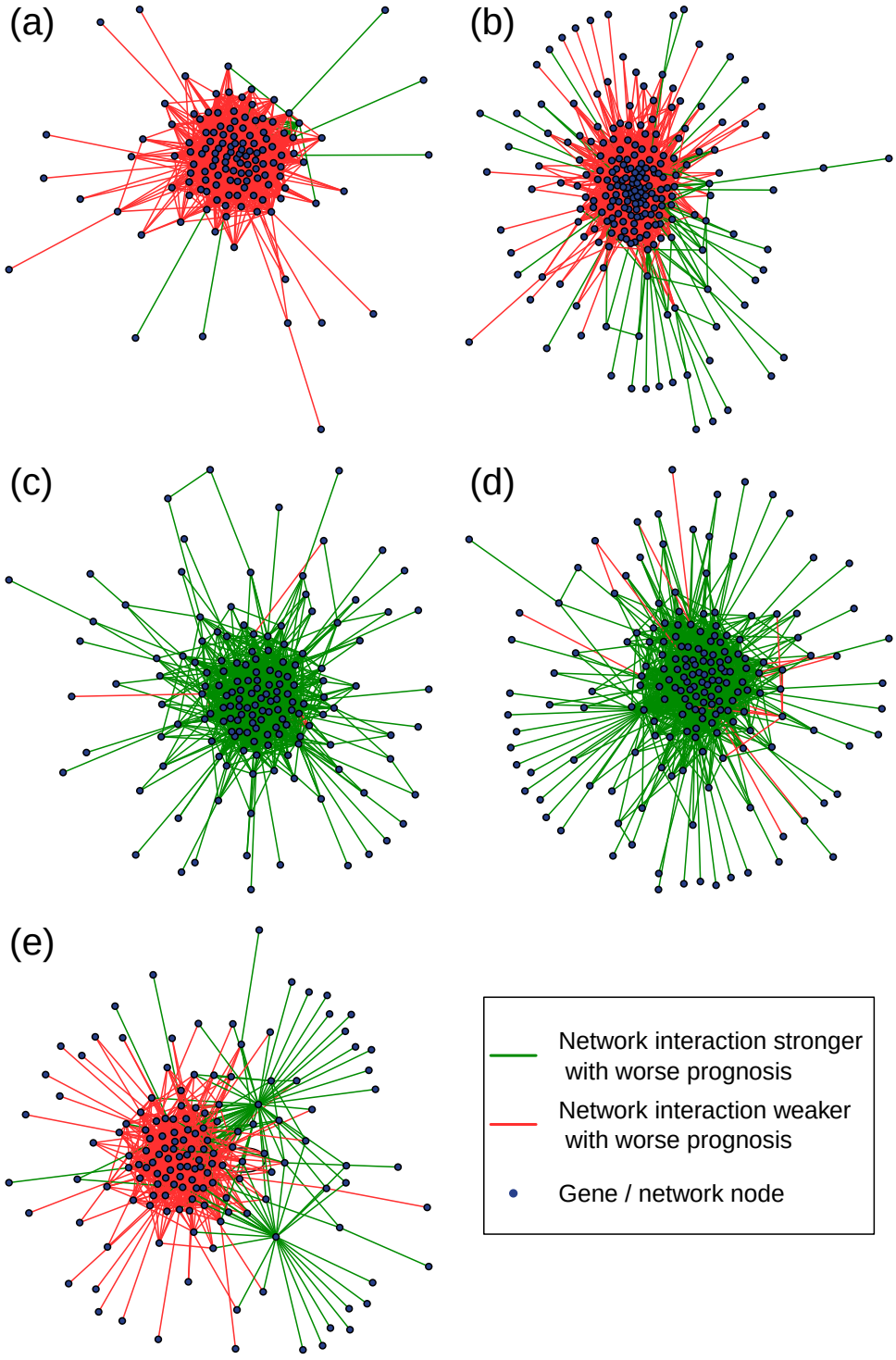


Figure 7.6: *Detected network community oncomarkers.*
(a) - (e) indicate network community oncomarkers 1 - 5, as shown in Figure 7.3.

network interaction measure is highly associated with co-regulatory behaviour linked to gene expression (at the mRNA level), giving functional relevance to the findings. However, there are also likely to be a range of genomic interactive effects present, which are measured by the DNA methylation network interaction measure, but which are not reflected in mRNA levels.

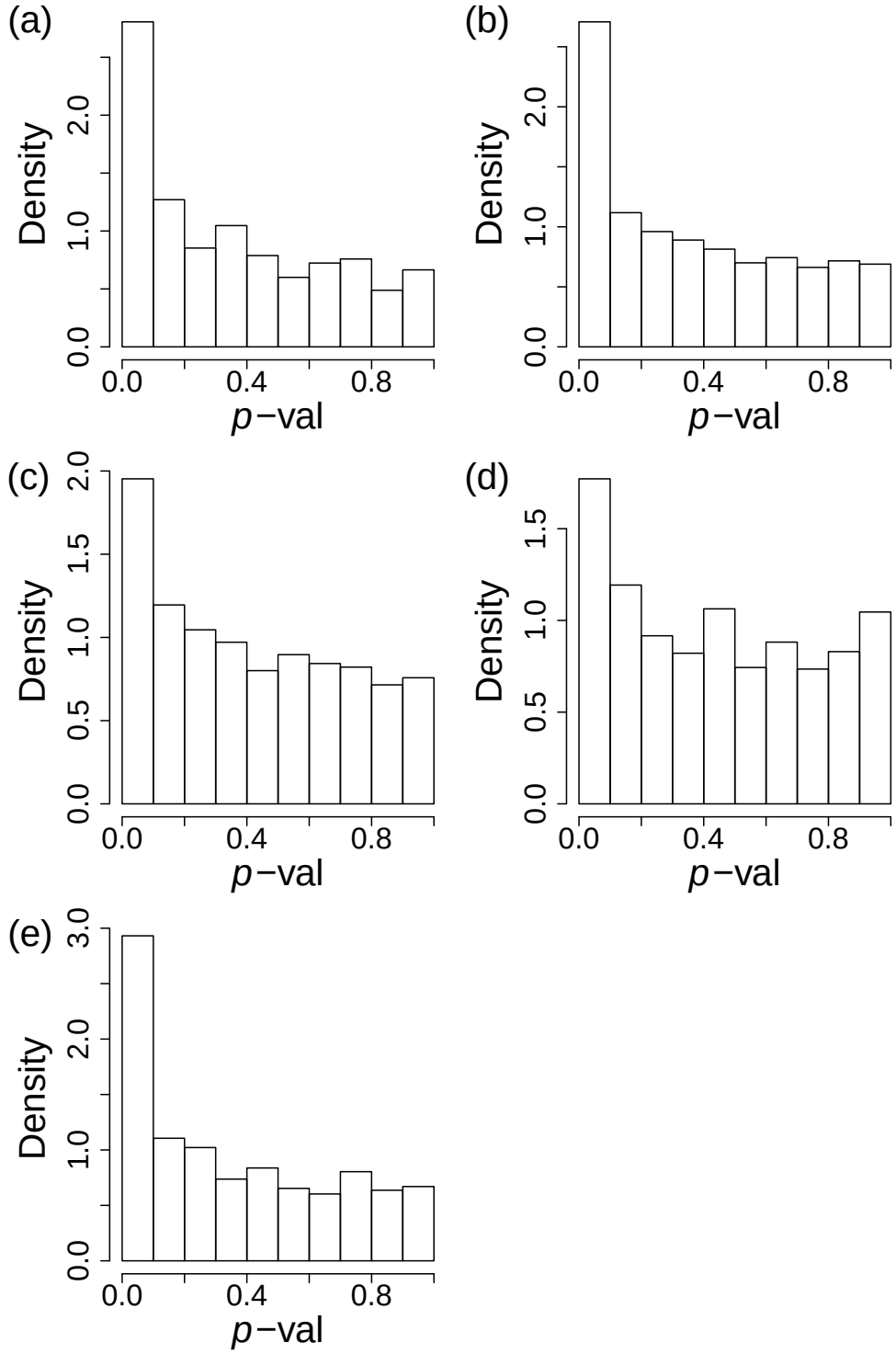


Figure 7.7: *Correlation of DNAm with gene expression for the network community oncomarkers. (a) - (e) indicate network community oncomarkers 1 - 5, as shown in Figure 7.3.*

I have also shown how to derive a one-number prognostic score for a network community oncomarker for each patient/sample. This prognostic score is a measure of disease progression in that patient.

The field of epigenomics is progressing fast, and promises much in the way of insights

into unexplained or undiscovered genomic phenomena, for example relating to the so-called genomic 'dark matter' of the genome (Venters & Pugh, 2013). Epigenomics is also expected to provide many new insights into disease progression: the discovery that some genomic loci gain or lose methylation in ways which may be unique to cancer suggests that understanding changes in DNA methylation machinery may be essential to understanding oncogenesis (Xie *et al.* , 2013). Epigenomics may also provide profound new insights into evolutionary processes: it has been suggested that epigenomic landscapes have shaped the evolution of the basic DNA sequence (Zhu *et al.* , 2013). This is because introns of genes tend to have a higher density of conserved non-coding sequence elements, compared to intergenic regions. These introns also tend to have more accessible chromatin, whereas the chromatin of intergenic regions is more likely to be epigenetically repressed.

The field of network science is also advancing rapidly. Networks are an efficient way to represent and analyse large numbers of variables, which is particularly relevant in modern, large-scale genomic studies. Networks of interactions are also a natural way to represent and analyse genomic interactions, associations and processes. Therefore, the study of genomic and epigenomic networks promises to be a productive field over the coming years, in terms of biology, medicine, and statistics.

7.5 Data-set info

DNA methylation (DNAm) data from breast cancer invasive carcinoma (BRCA) tumour samples, collected via the Illumina Infinium HumanMethylation450 platform, were downloaded from The Cancer Genome Atlas (TCGA) project (Hampton, 2006; Bonetta, 2006; Collins & Barker, 2007) at level 3. These data were pre-processed by first removing probes with non-unique mappings and which map to SNPs (as identified in the TCGA level 3 data); probes mapping to sex chromosomes were also removed; in total 98384 probes were removed in this way from all data sets. After removal of these probes, 270985 probes with known gene annotations remained. Probes were then removed if they had less than 95% coverage across samples; probe values were also replaced if they had corresponding detection p-value greater than 5%, by KNN (k nearest neighbour) imputation ($k = 5$). The loci of analysed CpGs were mapped to genes based on annotation information for the Illumina Infinium platform obtained from the *R* / *Bioconductor* package 'IlluminaHumanMethylation450k'. The data were also checked for batch effects by hierarchical clustering and correlation of the significant principle components with phenotype and batch: no significant batch effects (which would warrant further correction) were found. I downloaded DNA methylation data for tumour samples from 175

samples/individuals, from TCGA in July 2013, with clinical data available for patient survival outcome, and the clinical covariates age, disease stage, and residual disease. At the same time, I also downloaded corresponding DNA methylation data for healthy tissue for 98 individuals. These data were used to detect potential network community oncomarkers. I then downloaded DNA methylation data for a further 528 tumour samples from TCGA in September 2014, with data for the same clinical features available. These independent samples were used to validate the potential network community oncomarkers. I also downloaded gene expression data from TCGA at level 3, for 216 of the tumours for which I also obtained DNAm data.

7.6 Additional tables

Degree	Gene/node	Chr	Gene info
93	*POR*	7	P450 (cytochrome) oxidoreductase
87	*TTF1*	9	transcription termination factor, RNA polymerase I
79	*ZFPM2*	8	zinc finger protein, FOG family member 2
79	*ARHGAP21*	10	Rho GTPase activating protein 21
77	*VSIG2*	11	V-set and immunoglobulin domain containing 2
74	*P4HA1*	10	prolyl 4-hydroxylase, alpha polypeptide I
73	*MSLNL*	16	mesothelin-like
71	*COASY*	17	CoA synthase
71	*FBLL1*	5	fibrillarin-like 1
68	*ANXA2*	15	annexin A2
65	*CERS4*	19	ceramide synthase 4
63	*ZNF469*	16	zinc finger protein 469
63	*SYNGR3*	16	synaptogyrin 3
63	*FXYD1*	19	FXYD domain containing ion transport regulator 1
63	*IZUMO1*	19	izumo sperm-egg fusion 1
61	*EXOC2*	6	exocyst complex component 2
60	*RAP1GAP*	1	RAP1 GTPase activating protein
60	*PAK1*	11	p21 protein (Cdc42/Rac)-activated kinase 1
59	*DRD4*	11	dopamine receptor D4
59	*TAF5L*	1	TAF5-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor, 65kDa
58	*SHOX2*	3	short stature homeobox 2
58	*HOXB9*	17	homeobox B9
57	*TACR1*	2	tachykinin receptor 1
57	*DCHS1*	11	dachsous cadherin-related 1
56	*RTP3*	3	receptor (chemosensory) transporter protein 3
55	*DDX52*	17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 52
54	*SNX32*	11	sorting nexin 32
54	*TLE1*	9	transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)
53	*CNNM4*	2	cyclin M4
53	*CNIH2*	11	cornichon family AMPA receptor auxiliary protein 2
53	*LOC400940*	2	uncharacterized LOC400940
53	*MAPK3*	16	mitogen-activated protein kinase 3
52	*RB1*	13	retinoblastoma 1
52	*FUCA1*	1	fucosidase, alpha-L- 1, tissue
50	*PPP2R5C*	14	protein phosphatase 2, regulatory subunit B', gamma
50	*B3GALTL*	13	beta 1,3-galactosyltransferase-like
48	*JAK3*	19	Janus kinase 3
47	*SLC25A42*	19	solute carrier family 25, member 42
46	*TTC22*	1	tetratricopeptide repeat domain 22
46	*NPDC1*	9	neural proliferation, differentiation and control, 1
45	*ASB4*	7	ankyrin repeat and SOCS box containing 4
45	*ALDH2*	12	aldehyde dehydrogenase 2 family (mitochondrial)
45	*ZNF296*	19	zinc finger protein 296
44	*RBPJ*	4	recombination signal binding protein for immunoglobulin kappa J region
44	*NAT8L*	4	N-acetyltransferase 8-like (GCN5-related, putative)
44	*SMPDL3A*	6	sphingomyelin phosphodiesterase, acid-like 3A
42	*KLHL26*	19	kelch-like family member 26
41	*EBF4*	20	early B-cell factor 4
41	*SLAIN1*	13	SLAIN motif family, member 1
41	*GAMT*	19	guanidinoacetate N-methyltransferase
41	*SH2D3A*	19	SH2 domain containing 3A
40	*BLVRA*	7	biliverdin reductase A
39	*CD36*	7	CD36 molecule (thrombospondin receptor)
39	*BAZ1A*	14	bromodomain adjacent to zinc finger domain, 1A
39	*MLL5*	7	lysine (K)-specific methyltransferase 2E
37	*PIK3AP1*	10	phosphoinositide-3-kinase adaptor protein 1
36	*ITGB1BP1*	2	integrin beta 1 binding protein 1
34	*CMKLR1*	12	chemokine-like receptor 1
33	*TRIM71*	3	tripartite motif containing 71, E3 ubiquitin protein ligase
31	*SMAD3*	15	SMAD family member 3
31	*KIF13B*	8	kinesin family member 13B
30	*ARID3A*	19	AT rich interactive domain 3A (BRIGHT-like)
30	*F2R*	5	coagulation factor II (thrombin) receptor
30	*AMN1*	12	antagonist of mitotic exit network 1 homolog (S. cerevisiae)
29	*LOC100128239*	11	uncharacterized LOC100128239
29	*LRRC8B*	1	leucine rich repeat containing 8 family, member B
29	*ANKRD39*	2	ankyrin repeat domain 39
29	*ARFGAP3*	22	ADP-ribosylation factor GTPase activating protein 3
29	*RBM28*	7	RNA binding motif protein 28
28	*ABR*	17	active BCR-related
28	*CALU*	7	calumenin
28	*BRPF1*	3	bromodomain and PHD finger containing, 1
28	*C17orf104*	17	chromosome 17 open reading frame 104
28	*PAQR3*	4	progestin and adipoQ receptor family member III
27	*RGL2*	6	ral guanine nucleotide dissociation stimulator-like 2
27	*WAC*	10	WW domain containing adaptor with coiled-coil
27	*PMVK*	1	phosphomevalonate kinase
27	*PPP6R3*	11	protein phosphatase 6, regulatory subunit 3
26	*PPP2R1B*	11	protein phosphatase 2, regulatory subunit A, beta
25	*TOLLIP*	11	toll interacting protein
25	*RNASEH2A*	19	ribonuclease H2, subunit A
24	*RERE*	1	arginine-glutamic acid dipeptide (RE) repeats
23	*KRT27*	17	keratin 27
21	*B4GALNT2*	17	beta-1,4-N-acetyl-galactosaminyl transferase 2
21	*MYCBPAP*	17	MYCBP associated protein

Table 7.3: *Network Community Oncomarker 1 (Figure 7.3a) - gene/node info. The 85 highest degree nodes only are shown.*

Degree	Gene/node	Chr	Gene info
137	*TMEM198*	2	transmembrane protein 198
121	*POMP*	13	proteasome maturation protein
108	*GLT25D1*		
107	*HMOX1*	22	heme oxygenase (decycling) 1
100	*STK4*	20	serine/threonine kinase 4
94	*C1orf38*		
90	*XPO4*	13	exportin 4
83	*SOX5*	12	SRY (sex determining region Y)-box 5
82	*ADRA1B*	5	adrenoceptor alpha 1B
81	*RIMKLB*	12	ribosomal modification protein rimK-like family member B
80	*SMG6*	17	SMG6 nonsense mediated mRNA decay factor
72	*PHLDB1*	11	pleckstrin homology-like domain, family B, member 1
72	*PLTP*	20	phospholipid transfer protein
72	*C10orf32*	10	chromosome 10 open reading frame 32
71	*DLG4*	17	discs, large homolog 4 (Drosophila)
67	*SLC27A3*	1	solute carrier family 27 (fatty acid transporter), member 3
66	*KIAA1462*	10	KIAA1462
66	*FES*	15	feline sarcoma oncogene
66	*NDEL1*	17	nudE neurodevelopment protein 1-like 1
65	*ERGIC1*	5	endoplasmic reticulum-golgi intermediate compartment (ERGIC) 1
63	*FTSJD2*	6	cap methyltransferase 1
62	*EEPD1*	7	endonuclease/exonuclease/phosphatase family domain containing 1
61	*KCNA3*	1	potassium voltage-gated channel, shaker-related subfamily, member 3
60	*BREA2*	8	breast cancer estrogen-induced apoptosis 2
59	*MAGI2*	7	membrane associated guanylate kinase, WW and PDZ domain containing 2
59	*NPFF*	12	neuropeptide FF-amide peptide precursor
57	*SPRYD3*	12	SPRY domain containing 3
57	*WDR48*	3	WD repeat domain 48
56	*UHRF1BP1L*	12	UHRF1 binding protein 1-like
55	*ID1*	20	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein
55	*GABRA4*	4	gamma-aminobutyric acid (GABA) A receptor, alpha 4
55	*RNASE1*	14	ribonuclease, RNase A family, 1 (pancreatic)
54	*CDKL1*	14	cyclin-dependent kinase-like 1 (CDC2-related kinase)
54	*MAP4K1*	19	mitogen-activated protein kinase kinase kinase kinase 1
54	*TRADD*	16	TNFRSF1A-associated via death domain
52	*LOXL2*	8	lysyl oxidase-like 2
52	*CARS*	11	cysteinyl-tRNA synthetase
51	*NR3C1*	5	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
51	*SPEF2*	5	sperm flagellar 2
51	*LSM14B*	20	LSM14B, SCD6 homolog B (S. cerevisiae)
50	*LRBA*	4	LPS-responsive vesicle trafficking, beach and anchor containing
50	*LOC440910*	2	uncharacterized LOC440910
49	*SELO*	22	selenoprotein O
46	*TAOK1*	17	TAO kinase 1
46	*DNPEP*	2	aspartyl aminopeptidase
43	*HOXD10*	2	homeobox D10
43	*HGSNAT*	8	heparan-alpha-glucosaminide N-acetyltransferase
43	*ERMAP*	1	erythroblast membrane-associated protein (Scianna blood group)
43	*PPAP2A*	5	phosphatidic acid phosphatase type 2A
40	*MAML3*	4	mastermind-like 3 (Drosophila)
40	*FBXO4*	5	F-box protein 4
40	*SFT2D1*	6	SFT2 domain containing 1
39	*RIN2*	20	Ras and Rab interactor 2
38	*SYCP1*	1	synaptonemal complex protein 1
37	*PLBD1*	12	phospholipase B domain containing 1
36	*PRKCG*	19	protein kinase C, gamma
36	*ANKMY1*	2	ankyrin repeat and MYND domain containing 1
36	*ADAM19*	5	ADAM metallopeptidase domain 19
35	*PARD3*	10	par-3 family cell polarity regulator
35	*EXOC3*	5	exocyst complex component 3
33	*TTYH3*	7	tweety family member 3
33	*PIGG*	4	phosphatidylinositol glycan anchor biosynthesis, class G
33	*PFDN1*	5	prefoldin subunit 1
32	*PCDH8*	13	protocadherin 8
32	*PCBD2*	5	pterin-4 alpha-carbinolamine dehydratase/dimerization cofactor of hepatocyte nuclear factor 1 alpha (TCF1) 2
32	*NR1H3*	11	nuclear receptor subfamily 1, group H, member 3
32	*CTAGE1*	18	cutaneous T-cell lymphoma-associated antigen 1
31	*SOX14*	3	SRY (sex determining region Y)-box 14
31	*LRRC41*	1	leucine rich repeat containing 41
30	*PTPLAD1*	15	protein tyrosine phosphatase-like A domain containing 1
30	*ARMC2*	6	armadillo repeat containing 2
30	*BMP2*	20	bone morphogenetic protein 2
29	*NKAPL*	6	NFKB activating protein-like
28	*CCDC17*	1	coiled-coil domain containing 17
27	*ARL5C*	17	ADP-ribosylation factor-like 5C
27	*CECR6*	22	cat eye syndrome chromosome region, candidate 6
27	*SH3BGRL3*	1	SH3 domain binding glutamate-rich protein like 3
26	*TMEM51*	1	transmembrane protein 51
26	*C1QL3*	10	complement component 1, q subcomponent-like 3
26	*GPANK1*	6	G patch domain and ankyrin repeats 1
25	*KIAA0226*	3	KIAA0226
23	*GGT7*	20	gamma-glutamyltransferase 7
23	*ZNF837*	19	zinc finger protein 837
22	*VPS13D*	1	vacuolar protein sorting 13 homolog D (S. cerevisiae)
22	*SLC12A4*	16	solute carrier family 12 (potassium/chloride transporter), member 4

Table 7.4: *Network Community Oncomarker 2 (Figure 7.3b) - gene/node info. The 85 highest degree nodes only are shown.*

Degree	Gene/node	Chr	Gene info
60	*SOD2*	6	superoxide dismutase 2, mitochondrial
56	*ULK1*	12	unc-51 like autophagy activating kinase 1
56	*IL36B*	2	interleukin 36, beta
47	*GOLGA8A*	15	golgin A8 family, member A
44	*C14orf162*		
44	*DDX27*	20	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27
44	*MRPL35*	2	mitochondrial ribosomal protein L35
43	*ZNF202*	11	zinc finger protein 202
43	*JUND*	19	jun D proto-oncogene
43	*PAPD4*	5	PAP associated domain containing 4
42	*ASF1B*	19	anti-silencing function 1B histone chaperone
41	*SLC35E3*	12	solute carrier family 35, member E3
41	*USF1*	1	upstream transcription factor 1
41	*AXDND1*	1	axonemal dynein light chain domain containing 1
40	*PAFAH1B2*	11	platelet-activating factor acetylhydrolase 1b, catalytic subunit 2 (30kDa)
39	*ZNF2*	2	zinc finger protein 2
39	*KIF2C*	1	kinesin family member 2C
37	*SOX4*	6	SRY (sex determining region Y)-box 4
37	*CNIH4*	1	cornichon family AMPA receptor auxiliary protein 4
37	*TDRD12*	19	tudor domain containing 12
36	*IFNGR2*	21	interferon gamma receptor 2 (interferon gamma transducer 1)
35	*NMI*	2	N-myc (and STAT) interactor
35	*ADAM29*	4	ADAM metallopeptidase domain 29
34	*DNAJC16*	1	DnaJ (Hsp40) homolog, subfamily C, member 16
32	*GSR*	8	glutathione reductase
32	*RPL5*	1	ribosomal protein L5
32	*C16orf79*		
31	*C13orf35*	13	ATP11A upstream neighbor
30	*SLC7A5*	16	solute carrier family 7 (amino acid transporter light chain, L system), member 5
30	*ATXN2*	12	ataxin 2
30	*KLC4*	6	kinesin light chain 4
29	*TMEM8A*	16	transmembrane protein 8A
29	*DCLRE1C*	10	DNA cross-link repair 1C
28	*ORAI1*	12	ORAI calcium release-activated calcium modulator 1
28	*MTHFS*	15	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)
27	*GRIA4*	11	glutamate receptor, ionotropic, AMPA 4
27	*DDA1*	19	DET1 and DDB1 associated 1
27	*SDF2L1*	22	stromal cell-derived factor 2-like 1
27	*HIST1H2AB*	6	histone cluster 1, H2ab
26	*P2RX1*	17	purinergic receptor P2X, ligand-gated ion channel, 1
26	*SLC22A1*	6	solute carrier family 22 (organic cation transporter), member 1
26	*FBXL12*	19	F-box and leucine-rich repeat protein 12
25	*SCLY*	2	selenocysteine lyase
25	*HFM1*	1	HFM1, ATP-dependent DNA helicase homolog (S. cerevisiae)
24	*CHRM3*	1	cholinergic receptor, muscarinic 3
23	*ZNF764*	16	zinc finger protein 764
23	*LEO1*	15	Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)
23	*MARC1*	1	mitochondrial amidoxime reducing component 1
22	*CAPRIN1*	11	cell cycle associated protein 1
22	*RAB11A*	15	RAB11A, member RAS oncogene family
22	*CCNI*	4	cyclin I
22	*PARP14*	3	poly (ADP-ribose) polymerase family, member 14
22	*RIPK3*	14	receptor-interacting serine-threonine kinase 3
22	*VCP*	9	valosin containing protein
21	*SKAP2*	7	src kinase associated phosphoprotein 2
21	*AGTR1*	3	angiotensin II receptor, type 1
21	*TMEM45B*	11	transmembrane protein 45B
21	*NEFL*	8	neurofilament, light polypeptide
21	*TWF2*	3	twinfilin actin-binding protein 2
21	*C6orf141*	6	chromosome 6 open reading frame 141
21	*LOC442308*		
21	*TRIM21*	11	tripartite motif containing 21
20	*ADSL*	22	adenylosuccinate lyase
19	*WDR54*	2	WD repeat domain 54
19	*GMPPB*	3	GDP-mannose pyrophosphorylase B
19	*RECK*	9	reversion-inducing-cysteine-rich protein with kazal motifs
19	*NDUFS5*	1	NADH dehydrogenase (ubiquinone) Fe-S protein 5, 15kDa (NADH-coenzyme Q reductase)
18	*SLC39A7*	6	solute carrier family 39 (zinc transporter), member 7
17	*CPT1C*	19	carnitine palmitoyltransferase 1C
16	*PAFAH2*	1	platelet-activating factor acetylhydrolase 2, 40kDa
16	*NOS2*	17	nitric oxide synthase 2, inducible
15	*ING3*	7	inhibitor of growth family, member 3
14	*HOXC10*	12	homeobox C10
13	*UPF1*	19	UPF1 regulator of nonsense transcripts homolog (yeast)
13	*PKHD1*	6	polycystic kidney and hepatic disease 1 (autosomal recessive)
13	*NCKAP5L*	12	NCK-associated protein 5-like
12	*CEBPE*	14	CCAAT/enhancer binding protein (C/EBP), epsilon
12	*USP20*	9	ubiquitin specific peptidase 20
12	*ST6GALNAC1*	17	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 1
11	*ABHD16B*	20	abhydrolase domain containing 16B
11	*REXO1L2P*	8	REX1, RNA exonuclease 1 homolog (S. cerevisiae)-like 2 (pseudogene)
10	*NHP2L1*	22	NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae)
10	*GANAB*	11	glucosidase, alpha; neutral AB
10	*PKD1L2*	16	polycystic kidney disease 1-like 2
10	*CDR2*	16	cerebellar degeneration-related protein 2, 62kDa

Table 7.5: *Network Community Oncomarker 3 (Figure 7.3c) - gene/node info. The 85 highest degree nodes only are shown.*

Degree	Gene/node	Chr	Gene info
74	*NOL3*	16	nucleolar protein 3 (apoptosis repressor with CARD domain)
72	*NAPRT1*	8	nicotinate phosphoribosyltransferase domain containing 1
72	*PAH*	12	phenylalanine hydroxylase
69	*POU2F3*	11	POU class 2 homeobox 3
67	*SNCA*	4	synuclein, alpha (non A4 component of amyloid precursor)
66	*HOXA13*	7	homeobox A13
63	*HOXC12*	12	homeobox C12
60	*WFDC12*	20	WAP four-disulfide core domain 12
58	*HMG20B*	19	high mobility group 20B
51	*SAMD3*	6	sterile alpha motif domain containing 3
51	*SLC10A2*	13	solute carrier family 10 (sodium/bile acid cotransporter), member 2
49	*ZNF804B*	7	zinc finger protein 804B
48	*NKX6-1*	4	NK6 homeobox 1
48	*TEX19*	17	testis expressed 19
47	*SLC32A1*	20	solute carrier family 32 (GABA vesicular transporter), member 1
47	*DSC3*	18	desmocollin 3
47	*CCDC134*	22	coiled-coil domain containing 134
47	*BDH2*	4	3-hydroxybutyrate dehydrogenase, type 2
46	*ABCG4*	11	ATP-binding cassette, sub-family G (WHITE), member 4
46	*VPS41*	7	vacuolar protein sorting 41 homolog (S. cerevisiae)
45	*SIX1*	14	SIX homeobox 1
45	*O3FAR1*		
45	*MIR219-2*	9	microRNA 219-2
44	*LHX5*	12	LIM homeobox 5
44	*TARS*	5	threonyl-tRNA synthetase
44	*C6orf221*		
44	*C1orf100*	1	chromosome 1 open reading frame 100
43	*PDX1*	13	pancreatic and duodenal homeobox 1
42	*VSX2*	14	visual system homeobox 2
41	*ACTR2*	2	ARP2 actin-related protein 2 homolog (yeast)
41	*ASZ1*	7	ankyrin repeat, SAM and basic leucine zipper domain containing 1
40	*E2F8*	11	E2F transcription factor 8
40	*DPPA2*	3	developmental pluripotency associated 2
39	*LINC00461*	5	long intergenic non-protein coding RNA 461
38	*AP4E1*	15	adaptor-related protein complex 4, epsilon 1 subunit
38	*GPR150*	5	G protein-coupled receptor 150
37	*LOC440461*	17	Rho GTPase activating protein 27 pseudogene
37	*C15orf55*	15	NUT midline carcinoma, family member 1
37	*WTH3DI*	2	RAB6C-like
36	*GAD1*	2	glutamate decarboxylase 1 (brain, 67kDa)
36	*TUBA1C*	12	tubulin, alpha 1c
36	*FAM123A*		
36	*TULP2*	19	tubby like protein 2
36	*C19orf53*	19	chromosome 19 open reading frame 53
35	*LHX9*	1	LIM homeobox 9
35	*LINC00520*	14	long intergenic non-protein coding RNA 520
34	*HOXD13*	2	homeobox D13
34	*KCTD17*	22	potassium channel tetramerization domain containing 17
33	*HPX*	11	hemopexin
32	*CD8B*	2	CD8b molecule
31	*SH2D1B*	1	SH2 domain containing 1B
30	*GDNF*	5	glial cell derived neurotrophic factor
30	*RXFP3*	5	relaxin/insulin-like family peptide receptor 3
30	*CHMP4B*	20	charged multivesicular body protein 4B
29	*SPRED1*	15	sprouty-related, EVH1 domain containing 1
29	*MAPK6*	15	mitogen-activated protein kinase 6
28	*SLC15A1*	13	solute carrier family 15 (oligopeptide transporter), member 1
28	*HTR1D*	1	5-hydroxytryptamine (serotonin) receptor 1D, G protein-coupled
27	*VSTM2L*	20	V-set and transmembrane domain containing 2 like
27	*HIST3H2BB*	1	histone cluster 3, H2bb
26	*CLGN*	4	calmegin
26	*MCART2*		
25	*DNAJC19*	3	DnaJ (Hsp40) homolog, subfamily C, member 19
24	*ZNF710*	15	zinc finger protein 710
24	*CDH7*	18	cadherin 7, type 2
23	*POLR2C*	16	polymerase (RNA) II (DNA directed) polypeptide C, 33kDa
22	*LIMS2*	2	LIM and senescent cell antigen-like domains 2
22	*ZSWIM6*	5	zinc finger, SWIM-type containing 6
21	*FAM174A*	5	family with sequence similarity 174, member A
21	*MIR130A*	11	microRNA 130a
20	*ZFAND6*	15	zinc finger, AN1-type domain 6
19	*ACTA1*	1	actin, alpha 1, skeletal muscle
19	*TXNDC11*	16	thioredoxin domain containing 11
19	*ARF3*	12	ADP-ribosylation factor 3
19	*SNAI1*	20	snail family zinc finger 1
19	*C11orf20*		
18	*FAM195A*	16	family with sequence similarity 195, member A
18	*PPP1R3G*	6	protein phosphatase 1, regulatory subunit 3G
17	*ADAP2*	17	ArfGAP with dual PH domains 2
16	*PAQR5*	15	progestin and adipoQ receptor family member V
16	*GLTPD1*	1	glycolipid transfer protein domain containing 1
16	*SLC4A9*	5	solute carrier family 4, sodium bicarbonate cotransporter, member 9
14	*BMP7*	20	bone morphogenetic protein 7
14	*NOP58*	2	NOP58 ribonucleoprotein
14	*MPND*	19	MPN domain containing

Table 7.6: *Network Community Oncomarker 4 (Figure 7.3d) - gene/node info. The 85 highest degree nodes only are shown.*

Degree	Gene/node	Chr	Gene info
62	*C3orf18*	3	chromosome 3 open reading frame 18
49	*FAR1*	11	fatty acyl CoA reductase 1
43	*DDAH1*	1	dimethylarginine dimethylaminohydrolase 1
43	*L3MBTL4*	18	l(3)mbt-like 4 (Drosophila)
41	*NOVA1*	14	neuro-oncological ventral antigen 1
40	*SPRY1*	4	sprouty homolog 1, antagonist of FGF signaling (Drosophila)
39	*MAP3K8*	10	mitogen-activated protein kinase kinase kinase 8
39	*SERTAD4*	1	SERTA domain containing 4
38	*TPPP3*	16	tubulin polymerization-promoting protein family member 3
37	*IGFLR1*	19	IGF-like family receptor 1
35	*KANSL1L*	2	KAT8 regulatory NSL complex subunit 1-like
35	*LOC100130417*	1	uncharacterized LOC100130417
31	*CCHCR1*	6	coiled-coil alpha-helical rod protein 1
31	*LOX*	5	lysyl oxidase
31	*PLK3*	1	polo-like kinase 3
31	*RSL1D1*	16	ribosomal L1 domain containing 1
30	*KIAA0825*	5	KIAA0825
30	*SEC22A*	3	SEC22 vesicle trafficking protein homolog A (S. cerevisiae)
29	*VGLL3*	3	vestigial-like family member 3
29	*MFN2*	1	mitofusin 2
29	*TRHR*	8	thyrotropin-releasing hormone receptor
29	*SIGLECP3*		
29	*TAAR9*	6	trace amine associated receptor 9 (gene/pseudogene)
28	*LMO4*	1	LIM domain only 4
28	*POLE2*	14	polymerase (DNA directed), epsilon 2, accessory subunit
27	*SNX18*	5	sorting nexin 18
26	*PHACTR2*	6	phosphatase and actin regulator 2
26	*SCARB2*	4	scavenger receptor class B, member 2
26	*PGLS*	19	6-phosphogluconolactonase
26	*MIR365B*	17	microRNA 365b
25	*C16orf80*	16	chromosome 16 open reading frame 80
25	*CDK17*	12	cyclin-dependent kinase 17
24	*GPM6A*	4	glycoprotein M6A
24	*SLC35D1*	1	solute carrier family 35 (UDP-GlcA/UDP-GalNAc transporter), member D1
24	*PMM2*	16	phosphomannomutase 2
24	*C8orf45*		
23	*SOAT1*	1	sterol O-acyltransferase 1
22	*KIFC1*	6	kinesin family member C1
22	*ZNF8*	19	zinc finger protein 8
22	*TXNDC15*	5	thioredoxin domain containing 15
22	*FLJ26850*	19	FLJ26850 protein
21	*STAT5A*	17	signal transducer and activator of transcription 5A
21	*ST6GALNAC6*	9	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 6
21	*KCTD10*	12	potassium channel tetramerization domain containing 10
20	*ANKS1A*	6	ankyrin repeat and sterile alpha motif domain containing 1A
20	*XYLT2*	17	xylosyltransferase II
20	*NUCB1*	19	nucleobindin 1
19	*LOC440040*	11	glutamate receptor, metabotropic 5 pseudogene
18	*HBM*	16	hemoglobin, mu
18	*ENPP2*	8	ectonucleotide pyrophosphatase/phosphodiesterase 2
18	*SNX21*	20	sorting nexin family member 21
18	*KLF1*	19	Kruppel-like factor 1 (erythroid)
17	*HMGB2*	4	high mobility group box 2
17	*FOXD1*	5	forkhead box D1
16	*WDR43*	2	WD repeat domain 43
15	*STK10*	5	serine/threonine kinase 10
15	*GSX1*	13	GS homeobox 1
14	*GRM3*	7	glutamate receptor, metabotropic 3
14	*LOC285548*	4	long intergenic non-protein coding RNA 1096
14	*HIST1H4G*	6	histone cluster 1, H4g
13	*RTN3*	11	reticulon 3
13	*ATG14*	14	autophagy related 14
13	*TBC1D3C*	17	TBC1 domain family, member 3C
12	*FNDC3B*	3	fibronectin type III domain containing 3B
12	*DNAJC6*	1	DnaJ (Hsp40) homolog, subfamily C, member 6
11	*YWHAG*	7	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma
11	*NUPL1*	13	nucleoporin like 1
10	*FHL3*	1	four and a half LIM domains 3
10	*ANKRD34A*	1	ankyrin repeat domain 34A
10	*GNB3*	12	guanine nucleotide binding protein (G protein), beta polypeptide 3
9	*GABRA1*	5	gamma-aminobutyric acid (GABA) A receptor, alpha 1
9	*MLXIP*	12	MLX interacting protein
9	*ADORA3*	1	adenosine A3 receptor
9	*TRIM17*	1	tripartite motif containing 17
9	*MFGE8*	15	milk fat globule-EGF factor 8 protein
9	*FAM86C2P*	11	family with sequence similarity 86, member A pseudogene
9	*LACTB*	15	lactamase, beta
8	*CYB5R3*	22	cytochrome b5 reductase 3
8	*ATOH7*	10	atonal homolog 7 (Drosophila)
8	*PRKAA1*	5	protein kinase, AMP-activated, alpha 1 catalytic subunit
7	*ATXN7*	3	ataxin 7
7	*IGFN1*	1	immunoglobulin-like and fibronectin type III domain containing 1
7	*HDHD2*	18	haloacid dehalogenase-like hydrolase domain containing 2
7	*C14orf39*	14	chromosome 14 open reading frame 39
6	*VAMP3*	1	vesicle-associated membrane protein 3

Table 7.7: *Network Community Oncomarker 5 (Figure 7.3e) - gene/node info. The 85 highest degree nodes only are shown.*

Chapter 8

Conclusions

8.1 Summary

Epigenetic processes - including DNA methylation - are increasingly seen as having a fundamental role in chronic diseases like cancer. Traditionally, methylation levels at particular genes or loci have been shown to differ between normal and diseased tissue. In chapter 2, I investigated stochastic processes in intra-gene DNA methylation patterns. I considered whether the intra-gene methylation architecture is corrupted in cancer and whether the variability of levels of methylation of individual CpGs within a defined gene is able to discriminate cancerous from normal tissue. I analysed 270985 CpGs annotated to 18272 genes in 681 normal and 3284 cancerous samples taken from 14 different cancer entities. I found novel differences in intra-gene methylation pattern across phenotypes, particularly in those genes which are crucial for stem cell biology; my measures of intra-gene methylation architecture are a better determinant of phenotype than measures based on mean methylation level alone (K-S test $p < 10^{-3}$ in all 14 cancer entities tested). These findings strongly support the view that in addition to mean methylation levels of linked CpGs (as analysed in methylation specific PCR), intra-gene methylation architecture has great clinical potential for the development of DNA-based cancer biomarkers.

Glioblastoma is a particularly aggressive cancer, with very poor prognosis. Gliobastomas are thought to be driven by stem-like cells, motivating the study of epigenetic changes which occur when gliolblastoma stem-like cells are caused to differentiate. In chapter 3, I developed statistical network methodology to analyse DNA methylation time-course experimental data from differentiating healthy human neural stem cells and human glioblastoma stem-like cells. In doing so, I identified a characteristic differential epigenotype of glioblastoma stem-like cells, which is normalised towards the epigenotype of healthy human neural stem-cells during differentiation. This glioblastoma stem-like cell differential epigenotype contains several genes which are very relevant to tumour, glioblastoma, and stem cell biology, including *WT1*, *STAT3*,

HOXD4, *EZH2*, *P73*, *PAX6*, *VIMENTIN*, and *CBP*.

In chapter 4, I presented methodology to enable estimation of binary adjacency matrices, from a range of measures of the strength of association between pairs of network nodes or, more generally, pairs of variables. This strength of association can be quantified in terms of sample covariance / correlation matrices, and more generally by test-statistics / hypothesis test p-values from arbitrary distributions. Binary adjacency matrices inferred in this way are then ideal for community detection, for example by fitting the stochastic blockmodel. I showed that this methodology works well in a range of data-sets, including a simulation study, and several gene expression data-sets. This methodology performs well on large datasets, and is based on commonly available and computationally efficient algorithms.

In chapter 5, I introduced the notion of co-modularity, to co-cluster observations of bipartite networks into co-communities. The task of co-clustering is to group together nodes of one type, whose interaction with nodes of another type are the most similar. The novel measure of co-modularity was introduced to assess the strength of co-communities, as well as to arrange the representation of nodes and clusters for visualisation. The existing non-parametric understanding of co-clustering was generalised in this chapter, by introducing an anisotropic graphon class for realisations of bipartite networks. By modelling the smoothness of the anisotropic graphon directly, it is possible to obtain a quantitative measure to determine the number of groups to be used when fitting co-communities, subsequently using the co-modularity measure to do so. I illustrated the power of the proposed methodology on simulated data, as well as an example based on IGV (intra-gene variability of DNA methylation) data with linked gene expression data.

In chapter 6, I further investigated IGV, finding that it is prognostic independently of known clinical factors. Using IGV, based on raw data, I derived a robust gene-panel prognostic signature for ovarian cancer (OC, $n = 221$), which validated in two independent data sets from Mayo Clinic ($n = 198$) and TCGA ($n = 358$), with significance of $p = 0.004$ in both sets. The OC prognostic signature gene-panel is comprised of four gene groups, which may represent distinct biological processes. I showed that the IGV of these gene groups is likely a surrogate measure of transcription factor (TF) binding/activity. Analysing linked DNA methylation and gene expression data, I also found co-clusters by using the methodology of chapter 5. These represent groups of genes with highly associated expression and IGV patterns, and provide a starting-point for further investigation into the mechanistic roles of the observed IGV patterns in disease. I concluded that IGV is a self-calibrating measure of methylation variability which can be used to predict clinical outcome in patients individually, providing a surrogate read-out

of hard-to-measure disease processes.

In chapter 7, I presented a DNA methylation-based measure of genomic interaction and association, called the 'DNA methylation network interaction measure'. I showed how to use the DNA methylation network interaction measure to infer prognostic genomic networks, and how to identify prognostic biomarkers from such networks, which I term 'network community oncomarkers'. I also showed that the DNA methylation network interaction measure, between a pair of genes, is highly associated with gene expression correlation between the same pair of genes. However, it also appears likely that other genomic effects in addition to those measured by gene expression may be included in the genomic interactive behaviour quantified by the DNA methylation network interaction measure. The methods presented in this chapter represent a foundation for the development of cancer biomarkers based on genomic networks derived from DNA methylation data.

8.2 Discussion and directions for further work

In this thesis, I have developed novel statistical methodology, which is also useful as a tool for use in cell biology, for discovering new molecular patterns and associations, and informing biomarker development. However, much of what I have found using this methodology has only suggested possible new associations between DNA methylation patterns and disease, without proving biological mechanisms for the role of these observed patterns in disease. To do so would require further experiments to be carried out, which are beyond the scope of this work. However, I hope that some of what I have done here will provide motivation and direction for others to carry out such experimental investigation. I also hope that this work will provide mathematical tools with which others may inform the direction of their own experimental and computational investigations.

DNA methylation is a conduit for environmental risk factors of disease. Although I have found DNA methylation patterns which are associated with disease, I have not investigated which environmental risk factors are associated with these observed DNA methylation patterns. Similarly, I have not investigated how those environmental factors could give rise to these observed DNA methylation patterns. Investigations of that type could provide valuable information for understanding the relevance and importance of a range of environmental risk factors in relation to disease. Such information could be very valuable from the perspective of public health and disease prevention.

Bibliography

Agrawal, Pooja, Yu, Kebing, Salomon, Arthur R, & Sedivy, John M. 2010. Proteomic profiling of Myc-associated proteins. *Cell Cycle*, **9**(24), 4908–4921.

Airoldi, Edoardo M, Costa, Thiago B, & Chan, Stanley H. 2013. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *Pages 692–700 of: Advances in Neural Information Processing Systems*.

Alberts, B. 2002. *Molecular Biology of the Cell*. 4th edn. Garland Science, New York.

Aldous, David J. 1985. *Exchangeability and related topics*. Springer.

Almlöf, Tova, Wallberg, Annika E, Gustafsson, Jan-Åke, & Wright, Anthony PH. 1998. Role of important hydrophobic amino acids in the interaction between the glucocorticoid receptor τ1-core activation domain and target factors. *Biochemistry*, **37**(26), 9586–9594.

Alter, O., Brown, P.O., & Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, **97**(18), 10101–10106.

Armstrong, Jane F, Pritchard-Jones, Kathryn, Bickmore, Wendy A, Hastie, Nicholas D, & Bard, Jonathan BL. 1993. The expression of the Wilms' tumour gene, WT1, in the developing mammalian embryo. *Mechanisms of development*, **40**(1), 85–97.

Attanasio, Catia, Nord, Alex S, Zhu, Yiwen, Blow, Matthew J, Biddie, Simon C, Mendenhall, Eric M, Dixon, Jesse, Wright, Crystal, Hosseini, Roya, Akiyama, Jennifer A, *et al.* . 2014. Tissue-specific SMARCA4 binding at active and repressed regulatory elements during embryogenesis. *Genome research*, **24**(6), 920–929.

Ballman, Karla V, Buckner, Jan C, Brown, Paul D, Giannini, Caterina, Flynn, Patrick J, LaPlant, Betsy R, & Jaeckle, Kurt A. 2007. The relationship between six-month progression-free survival and 12-month overall survival end points for phase II trials in patients with glioblastoma multiforme. *Neuro-oncology*, **9**(1), 29–38.

Barabási, Albert-László, & Albert, Réka. 1999. Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Barabási, Albert-László, & Oltvai, Zoltan N. 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2), 101–113.

Barré, Benjamin, Avril, Sylvie, & Coqueret, Olivier. 2003. Opposite regulation of Myc and p21 waf1 transcription by STAT3 proteins. *Journal of Biological Chemistry*, **278**(5), 2990–2996.

Bartlett, Thomas E, Zaikin, Alexey, Olhede, Sofia C, West, James, Teschendorff, Andrew E, & Widschwendter, Martin. 2013. Corruption of the Intra-Gene DNA Methylation Architecture Is a Hallmark of Cancer. *PloS One*, **8**(7), e68285.

Bartlett, Thomas E, Olhede, Sofia C, & Zaikin, Alexey. 2014. A DNA Methylation Network Interaction Measure, and Detection of Network Oncomarkers. *PloS One*, **9**(1), e84573.

Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A., & Weinberg, R.A. 2008. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature Genetics*, **40**(5), 499–507.

Benjamini, Yoav, & Hochberg, Yosef. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Bernstein, B.E., Meissner, A., & Lander, E.S. 2007. The mammalian epigenome. *Cell*, **128**(4), 669–681.

Bickel, Peter J, & Chen, Aiyou. 2009. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, **106**(50), 21068–21073.

Bickel, Peter J, & Levina, Elizaveta. 2008. Covariance regularization by thresholding. *The Annals of Statistics*, 2577–2604.

Bickmore, W. 2012 (March). *Chromatin compaction and nuclear organisation in development and disease*. Seminar given at the UCL Cancer Institute.

Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes & development*, **16**(1), 6.

Blelloch, R., Wang, Z., Meissner, A., Pollard, S., Smith, A., & Jaenisch, R. 2006. Reprogramming efficiency following somatic cell nuclear transfer is influenced by the differentiation and methylation state of the donor nucleus. *Stem Cells*, **24**(9), 2007–2013.

Blondel, Vincent D, Guillaume, Jean-Loup, Lambiotte, Renaud, & Lefebvre, Etienne. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), P10008.

Bollobás, Béla, Janson, Svante, & Riordan, Oliver. 2007. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, **31**(1), 3–122.

Bonetta, Laura. 2006. Genome sequencing in the fast lane. *Nature Methods*, **3**(2), 141.

Brocks, David, Assenov, Yassen, Minner, Sarah, Bogatyrova, Olga, Simon, Ronald, Koop, Christina, Oakes, Christopher, Zucknick, Manuela, Lipka, Daniel Bernhard, Weischenfeldt, Joachim, *et al.* . 2014. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell reports*, **8**(3), 798–806.

Carén, Helena, Pollard, Steven M, & Beck, Stephan. 2013. The good, the bad and the ugly: Epigenetic mechanisms in glioblastoma. *Molecular aspects of medicine*, **34**(4), 849–862.

Chan, Stanley H, & Airoldi, Edoardo M. 2014. A Consistent Histogram Estimator for Exchangeable Graph Models. *arXiv preprint arXiv:1402.1888*.

Chen, Mike Y, Clark, Aaron J, Chan, Dana C, Ware, Joy L, Holt, Shawn E, Chidambaram, Archana, Fillmore, Helen L, & Broaddus, William C. 2011. Wilms tumor 1 silencing decreases the viability and chemoresistance of glioblastoma cells in vitro: a potential role for IGF-1R de-repression. *Journal of neuro-oncology*, **103**(1), 87–102.

Choi, David, Wolfe, Patrick J, *et al.* . 2014. Co-clustering separately exchangeable network data. *The Annals of Statistics*, **42**(1), 29–63.

Collins, F., & Barker, A. 2007. Mapping the cancer genome. *Scientific American Magazine*, **296**(3), 50–57.

Consortium, ENCODE Project, *et al.* . 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**(5696), 636–640.

Consortium, ENCODE Project, *et al.* . 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

Coulson, Judy M. 2005. Transcriptional regulation: cancer, neurons and the REST. *Current biology*, **15**(17), R665–R668.

Cox, David R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, **34**, 187–220.

Danon, Leon, Diaz-Guilera, Albert, Duch, Jordi, & Arenas, Alex. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, **2005**(09), P09008.

De La Iglesia, Núria, Konopka, Genevieve, Puram, Sidharth V, Chan, Jennifer A, Bachoo, Robert M, You, Mingjian J, Levy, David E, DePinho, Ronald A, & Bonni, Azad. 2008. Identification of a PTEN-regulated STAT3 brain tumor suppressor pathway. *Genes & development*, **22**(4), 449–462.

Dhillon, Inderjit S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. *Pages 269–274 of: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Diaconis, Persi. 1977. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, **36**(2), 271–281.

Duong, Tarn, & Hazelton, Martin. 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15**(1), 17–30.

Easwaran, Hariharan, Johnstone, Sarah E, Van Neste, Leander, Ohm, Joyce, Mosbruger, Tim, Wang, Qiuju, Aryee, Martin J, Joyce, Patrick, Ahuja, Nita, Weisenberger, Dan, *et al.* . 2012. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Research*, **22**(5), 837–849.

Edgar, R., Domrachev, M., & Lash, A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**(1), 207–210.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*, **27**(8), 861–874.

Feil, Robert, & Fraga, Mario F. 2012. Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics*, **13**(2), 97–109.

Feinberg, A.P., Ohlsson, R., & Henikoff, S. 2006. The epigenetic progenitor origin of human cancer. *Nature Reviews Genetics*, **7**(1), 21–33.

Fisher, Ronald A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 507–521.

Flynn, Cheryl J, & Perry, Patrick O. 2012. Consistent biclustering. *arXiv preprint arXiv:1206.6927*.

Friedman, J., Hastie, T., & Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Gartel, AL. 2006. A new mode of transcriptional repression by c-Myc: methylation. *Oncogene*, **25**(14), 1989–1990.

Gelman, Laurent, Zhou, Gaochao, Fajas, Lluis, Raspé, Eric, Fruchart, Jean-Charles, & Auwerx, Johan. 1999. p300 interacts with the N-and C-terminal part of PPARγ2 in a ligand-independent and-dependent manner, respectively. *Journal of Biological Chemistry*, **274**(12), 7681–7688.

Gill, Zahidah P, Perks, Claire M, Newcomb, Paul V, & Holly, Jeff MP. 1997. Insulin-like growth factor-binding protein (IGFBP-3) predisposes breast cancer cells to programmed cell death in a non-IGF-dependent manner. *Journal of Biological Chemistry*, **272**(41), 25602–25607.

Girvan, Michelle, & Newman, Mark EJ. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**(12), 7821–7826.

Golub, Todd R, Slonim, Donna K, Tamayo, Pablo, Huard, Christine, Gaasenbeek, Michelle, Mesirov, Jill P, Coller, Hilary, Loh, Mignon L, Downing, James R, Caligiuri, Mark A, *et al.* . 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Gosden, R.G., & Feinberg, A.P. 2007. Genetics and epigeneticsnature's pen-and-pencil set. *New England Journal of Medicine*, **356**(7), 731–733.

Greenlee, Robert T, Hill-Harmon, Mary Beth, Murray, Taylor, & Thun, Michael. 2001. Cancer statistics, 2001. *CA: A Cancer Journal for Clinicians*, **51**(1), 15–36.

Gustems, Montse, Woellmer, Anne, Rothbauer, Ulrich, Eck, Sebastian H, Wieland, Thomas, Lutter, Dominik, & Hammerschmidt, Wolfgang. 2014. c-Jun/c-Fos heterodimers regulate cellular genes via a newly identified class of methylated DNA sequence motifs. *Nucleic Acids Research*, **42**(5), 3059–3072.

Hampton, Tracy. 2006. Cancer genome atlas. *JAMA: The Journal of the American Medical Association*, **296**(16), 1958–1958.

Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., *et al.* . 2011. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, **43**(8), 768–775.

Harrell, Frank E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Herzberg, AM, & Tsukanov, AV. 1986. The design of experiments for model selection. *Pages 175–178 of: Proc. 1st World Congress Bernoulli Soc*, vol. 2.

Holland, Paul W, Laskey, Kathryn Blackmond, & Leinhardt, Samuel. 1983. Stochastic blockmodels: First steps. *Social networks*, **5**(2), 109–137.

Hom, Roger A, & Johnson, Charles R. 1991. Topics in matrix analysis. *Cambridge UP, New York*.

Hotelling, Harold. 1936. Relations between two sets of variates. *Biometrika*, **28**(3/4), 321–377.

Izumoto, Shuichi, Tsuboi, Akihiro, Oka, Yoshihiro, Suzuki, Tsuyoshi, Hashiba, Tetsuo, Kagawa, Naoki, Hashimoto, Naoya, Maruno, Motohiko, Elisseeva, Olga A, Shirakata, Toshiaki, *et al.* . 2008. Phase II clinical trial of Wilms tumor 1 peptide vaccination for patients with recurrent glioblastoma multiforme. *Journal of neurosurgery*, **108**(5), 963–971.

Jacob, Laurent, Neuvial, Pierre, Dudoit, Sandrine, *et al.* . 2012. More power via graph-structured tests for differential expression of gene networks. *The Annals of Applied Statistics*, **6**(2), 561–600.

Jaffe, Andrew E, Feinberg, Andrew P, Irizarry, Rafael A, & Leek, Jeffrey T. 2012. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, **13**(1), 166–178.

Jelinic, Petar, Mueller, Jennifer J, Olvera, Narciso, Dao, Fanny, Scott, Sasinya N, Shah, Ronak, Gao, JianJiong, Schultz, Nikolaus, Gonen, Mithat, Soslow, Robert A, *et al.* . 2014. Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. *Nature Genetics*, **46**(5), 424–426.

Jemal, Ahmedin, Bray, Freddie, Center, Melissa M, Ferlay, Jacques, Ward, Elizabeth, & Forman, David. 2011. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, **61**(2), 69–90.

Jirtle, Randy L, & Skinner, Michael K. 2007. Environmental epigenomics and disease susceptibility. *Nature Reviews Genetics*, **8**(4), 253–262.

Johnstone, Iain M, & Silverman, Bernard W. 2004. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 1594–1649.

Jojic, Vladimir, Shay, Tal, Sylvia, Katelyn, Zuk, Or, Sun, Xin, Kang, Joonsoo, Regev, Aviv, Koller, Daphne, & Immunological Genome Project Consortium. 2013. Identification of transcriptional regulators in the mouse immune system. *Nature Immunology*, **14**(6), 633–643.

Jones, P.A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, **13**(7), 484–492.

Jones, P.A., & Baylin, S.B. 2002. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, **3**(6), 415–428.

Kallenberg, Olav. 2005. *Probabilistic symmetries and invariance principles*. Vol. 9. Springer.

Katenka, Natallia, Kolaczyk, Eric D, *et al.* . 2012. Inference and characterization of multi-attribute networks with application to computational biology. *The Annals of Applied Statistics*, **6**(3), 1068–1094.

Kirillov, Andrei, Kistler, Barbara, Mostoslavsky, Raul, Cedar, Howard, Wirth, Thomas, & Bergman, Yehudit. 1996. A role for nuclear NF–κB in B–cell–specific demethylation of the Igκ locus. *Nature Genetics*, **13**(4), 435–441.

Kleinsmith, Lewis J, & Pierce, G Barry. 1964. Multipotentiality of single embryonal carcinoma cells. *Cancer research*, **24**(9), 1544–1551.

Koller, Daphne, & Friedman, Nir. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Lai, Fan, & Shiekhattar, Ramin. 2014. Where long noncoding RNAs meet DNA methylation. *Cell research*, **24**(3), 263–264.

Lannoy, Vincent J, Rodolosse, Annie, Pierreux, Christophe E, Rousseau, Guy G, & Lemaigre, Frédéric P. 2000. Transcriptional stimulation by hepatocyte nuclear factor-6. *Journal of Biological Chemistry*, **275**(29), 22098–22103.

Larremore, Daniel B, Clauset, Aaron, & Jacobs, Abigail Z. 2014. Efficiently inferring community structure in bipartite networks. *arXiv preprint arXiv:1403.2933*.

Latouche, Pierre, Birmelé, Etienne, Ambroise, Christophe, *et al.* . 2011. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, **5**(1), 309–336.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., *et al.* . 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell*, **125**(2), 301–313.

Lee, Y, Kim, D, Kim, Y, Park, K, Shong, Minho, Seong, H, Ha, H, & Choi, H. 2008. Orphan nuclear receptor SHP interacts with and represses hepatocyte nuclear factor-6 (HNF-6) transactivation. *Biochem. J*, **413**, 559–569.

Lee, Yoon-Kwang, Dell, Helen, Dowhan, Dennis H, Hadzopoulou-Cladaras, Margarita, & Moore, David D. 2000. The orphan nuclear receptor SHP inhibits hepatocyte nuclear factor 4 and retinoid X receptor transactivation: two mechanisms for repression. *Molecular and Cellular Biology*, **20**(1), 187–195.

Lehoucq, Richard B, & Sørensen, Danny C. 1996. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, **17**(4), 789–821.

Lemasson, Isabelle, & Nyborg, Jennifer K. 2001. Human T-cell leukemia virus type I Tax repression of p73β is mediated through competition for the C/H1 domain of CBP. *Journal of Biological Chemistry*, **276**(19), 15720–15727.

Lerner, Lorena, Henriksen, Melissa A, Zhang, Xiaokui, & Darnell, James E. 2003. STAT3-dependent enhanceosome assembly and disassembly: synergy with GR for full transcriptional increase of the α2-macroglobulin gene. *Genes & development*, **17**(20), 2564–2577.

Li, Caiyan, & Li, Hongzhe. 2010. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, **4**(3), 1498.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., *et al.* . 2010. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biology*, **8**(11), e1000533.

Liu, Bingrong, Lee, Ho-Young, Weinzimer, Stuart A, Powell, David R, Clifford, John L, Kurie, Jon M, & Cohen, Pinchas. 2000. Direct functional interactions between insulin-like growth

factor-binding protein-3 and retinoid X receptor-α regulate transcriptional signaling and apoptosis. *Journal of Biological Chemistry*, **275**(43), 33607–33613.

Lunardi, Andrea, Di Minin, Giulio, Provero, Paolo, Dal Ferro, Marco, Carotti, Marcello, Del Sal, Giannino, & Collavin, Licio. 2010. A genome-scale protein interaction profile of Drosophila p53 uncovers additional nodes of the human p53 network. *Proceedings of the National Academy of Sciences*, **107**(14), 6322–6327.

Madeira, Sara C, & Oliveira, Arlindo L. 2004. Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **1**(1), 24–45.

Makismovic, Jovana, Gordon, Lavinia, & Oshlack, Alicia. 2012. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, **13**(6), 1–12.

Malone, Cindy Sue, Miner, Maurine D, Doerr, Jeanette R, Jackson, James P, Jacobsen, Steven E, Wall, Randolph, & Teitell, Michael. 2001. CmC (A/T) GG DNA methylation in mature B cell lymphoma gene silencing. *Proceedings of the National Academy of Sciences*, **98**(18), 10404–10409.

Mohseni-Zadeh, Sarah, & Binoux, Michel. 1997. Insulin-like growth factor (IGF) binding protein-3 interacts with the type 1 IGF receptor, reducing the affinity of the receptor for its ligand: an alternative mechanism in the regulation of IGF action. *Endocrinology*, **138**(12), 5645–5648.

Monti, Stefano, Tamayo, Pablo, Mesirov, Jill, & Golub, Todd. 2003. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, **52**(1-2), 91–118.

Nakahara, Yukiko, Okamoto, Hiroaki, Mineta, Toshihiro, & Tabuchi, Kazuo. 2004. Expression of the Wilms' tumor gene product WT1 in glioblastomas and medulloblastomas. *Brain tumor pathology*, **21**(3), 113–116.

Network, Cancer Genome Atlas, *et al.* . 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

Newman, Mark EJ. 2004. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, **38**(2), 321–330.

Newman, Mark EJ, & Girvan, Michelle. 2004. Finding and evaluating community structure in networks. *Physical review E*, **69**(2), 026113.

Newman, MEJ. 2013. Spectral methods for network community detection and graph partitioning. *arXiv preprint arXiv:1307.7729*.

Ng, Christopher W, Yildirim, Ferah, Yap, Yoon Sing, Dalin, Simona, Matthews, Bryan J, Velez, Patricio J, Labadorf, Adam, Housman, David E, & Fraenkel, Ernest. 2013. Extensive changes in DNA methylation are associated with expression of mutant huntingtin. *Proceedings of the National Academy of Sciences*, **110**(6), 2354–2359.

Nitzsche, Anja, Paszkowski-Rogacz, Maciej, Matarese, Filomena, Janssen-Megens, Eva M, Hubner, Nina C, Schulz, Herbert, de Vries, Ingrid, Ding, Li, Huebner, Norbert, Mann, Matthias, *et al.* . 2011. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PloS One*, **6**(5), e19470.

Ohm, Joyce E, McGarvey, Kelly M, Yu, Xiaobing, Cheng, Linzhao, Schuebel, Kornel E, Cope, Leslie, Mohammad, Helai P, Chen, Wei, Daniel, Vincent C, Yu, Wayne, *et al.* . 2007. A stem cell–like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature Genetics*, **39**(2), 237–242.

Oji, Yusuke, Suzuki, Tsuyoshi, Nakano, Yoko, Maruno, Motohiko, Nakatsuka, Shin-ichi, Jomgeow, Tanyarat, Abeno, Sakie, Tatsumi, Naoya, Yokota, Asumi, Aoyagi, Sayaka, *et al.* . 2005. Overexpression of the Wilms' tumor gene WT1 in primary astrocytic tumors. *Cancer science*, **95**(10), 822–827.

Olhede, Sofia C, & Wolfe, Patrick J. 2012. Degree-based network models. *arXiv preprint arXiv:1211.6537*.

Olhede, Sofia C., & Wolfe, Patrick J. 2014. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, **111**(41), 14722–14727.

Page, Lawrence, Brin, Sergey, Motwani, Rajeev, & Winograd, Terry. 1999. The PageRank citation ranking: Bringing order to the web.

Parker, Joel S, Mullins, Michael, Cheang, Maggie CU, Leung, Samuel, Voduc, David, Vickery, Tammi, Davies, Sherri, Fauron, Christiane, He, Xiaping, Hu, Zhiyuan, *et al.* . 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, **27**(8), 1160–1167.

Peng, Jie, Zhu, Ji, Bergamaschi, Anna, Han, Wonshik, Noh, Dong-Young, Pollack, Jonathan R, & Wang, Pei. 2010. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, **4**(1), 53.

Perry, Patrick O, & Wolfe, Patrick J. 2012. Null models for network data. *arXiv preprint arXiv:1201.5871*.

Peterson, Craig L, & Workman, Jerry L. 2000. Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Current opinion in genetics & development*, **10**(2), 187–192.

Piccirillo, SGM, Reynolds, BA, Zanetti, N, Lamorte, G, Binda, E, Broggi, G, Brem, H, Olivi, A, Dimeco, F, & Vescovi, AL. 2006. Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. *Nature*, **444**(7120), 761–765.

Pierreux, Christophe E, Stafford, John, Demonte, Dominique, Scott, Donald K, Vandenhaute, Jean, OBrien, Richard M, Granner, Daryl K, Rousseau, Guy G, & Lemaigre, Frederic P. 1999. Antiglucocorticoid activity of hepatocyte nuclear factor-6. *Proceedings of the National Academy of Sciences*, **96**(16), 8961–8966.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., & Stanley, H.E. 2002. Random matrix approach to cross correlations in financial data. *Physical Review E*, **65**(6), 066126.

Prisco, Marco, Peruzzi, Francesca, Belletti, Barbara, & Baserga, Renato. 2001. Regulation of Id gene expression by type I insulin-like growth factor: roles of Stat3 and the tyrosine 950 residue of the receptor. *Molecular and Cellular Biology*, **21**(16), 5447–5458.

Pujadas, E., & Feinberg, A.P. 2012. Regulated noise in the epigenetic landscape of development and disease. *Cell*, **148**(6), 1123–1131.

Qin, Tai, & Rohe, Karl. 2013. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Pages 3120–3128 of: Advances in Neural Information Processing Systems*.

Ramos, Pilar, Karnezis, Anthony N, Craig, David W, Sekulic, Aleksandar, Russell, Megan L, Hendricks, William PD, Corneveaux, Jason J, Barrett, Michael T, Shumansky, Karey, Yang, Yidong, *et al.* . 2014. Small cell carcinoma of the ovary, hypercalcemic type, displays frequent inactivating germline and somatic mutations in SMARCA4. *Nature genetics*, **46**(5), 427–429.

Riolo, Maria A, & Newman, MEJ. 2012. First-principles multiway spectral partitioning of graphs. *arXiv preprint arXiv:1209.5969*.

Rohe, Karl, & Yu, Bin. 2012. Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. *arXiv preprint arXiv:1204.2296*.

Rohe, Karl, Chatterjee, Sourav, Yu, Bin, *et al.* . 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, **39**(4), 1878–1915.

Sandoval, Juan, Heyn, Holger, Méndez-González, Jesús, Gomez, Antonio, Moran, Sebastian, Baiget, Montserrat, Melo, Montserrat, Badell, Isabel, Nomdedéu, Josep F, & Esteller, Manel. 2013. Genome-wide DNA methylation profiling predicts relapse in childhood B-cell acute lymphoblastic leukaemia. *British Journal of Haematology*, **160**(3), 406–409.

Scharnhorst, Volkher, Dekker, Patrick, van der Eb, Alex J, & Jochemsen, Aart G. 2000. Physical interaction between Wilms tumor 1 and p73 proteins modulates their functions. *Journal of Biological Chemistry*, **275**(14), 10202–10211.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E., *et al.* . 2006. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature Genetics*, **39**(2), 232–236.

Shen, Wei-fang, Krishnan, Keerthi, Lawrence, HJ, & Largman, Corey. 2001. The HOX homeodomain proteins block CBP histone acetyltransferase activity. *Science Signalling*, **21**(21), 7509.

Shen-Orr, Shai S, Milo, Ron, Mangan, Shmoolik, & Alon, Uri. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, **31**(1), 64–68.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. 2011. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **39**(5), 1–13.

Singh, Sheila K, Hawkins, Cynthia, Clarke, Ian D, Squire, Jeremy A, Bayani, Jane, Hide, Takuichiro, Henkelman, R Mark, Cusimano, Michael D, & Dirks, Peter B. 2004. Identification of human brain tumour initiating cells. *nature*, **432**(7015), 396–401.

Smyth, G.K., *et al.* . 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, **3**(1), 3.

Solomin, Ludmila, Johansson, Clas B, Zetterström, Rolf H, Bissonnette, Reid P, Heyman, Richard A, Olson, Lars, Lendahl, Urban, Frisén, Jonas, & Perlmann, Thomas. 1998. Retinoid-X receptor signalling in the developing spinal cord. *Nature*, **395**(6700), 398–402.

Sørensen, Danny C. 1992. Implicit application of polynomial filters in ak-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, **13**(1), 357–385.

Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., & Davis, R.W. 2005. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(36), 12837–12842.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* . 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545.

Sueoka, Naoko, Lee, Ho-Young, Wiehle, Sandra, Cristiano, Richard J, Fang, BingLiang, Ji, Lin, Roth, Jack A, Hong, Waun Ki, Cohen, Pinchas, Kurie, Jonathan M, *et al.* . 2000. Insulin-like growth factor binding protein-6 activates programmed cell death in non-small cell lung cancer cells. *Oncogene*, **19**(38), 4432–4436.

Surawicz, Tanya S, Davis, Faith, Freels, Sally, Laws, Edward R, & Menck, Herman R. 1998. Brain tumor survival: results from the National Cancer Data Base. *Journal of neuro-oncology*, **40**(2), 151–160.

Surawicz, Tanya S, McCarthy, Bridget J, Kupelian, Varant, Jukich, Patti J, Bruner, Janet M, Davis, Faith G, *et al.* . 1999. Descriptive epidemiology of primary brain and CNS tumors: results from the Central Brain Tumor Registry of the United States, 1990-1994. *Neuro-oncology*, **1**(1), 14–25.

Taylor, Ian W, Linding, Rune, Warde-Farley, David, Liu, Yongmei, Pesquita, Catia, Faria, Daniel, Bull, Shelley, Pawson, Tony, Morris, Quaid, & Wrana, Jeffrey L. 2009. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, **27**(2), 199–204.

Teschendorff, A.E., & Widschwendter, M. 2012. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, **28**(11), 1487–1494.

Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Gayther, S.A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I.J., *et al.* . 2009. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One*, **4**(12), e8274.

Teschendorff, Andrew E, Jones, Allison, Fiegl, Heidi, Sargent, Alexandra, Zhuang, Joanna J, Kitchener, Henry C, & Widschwendter, Martin. 2012. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Medicine*, **4**(3), 1–14.

Triche, Tim, & Jr. 2012. *IlluminaHumanMethylation450k.db: Illumina Human Methylation 450k annotation data*. R package version 1.4.6.

Valk, Peter JM, Verhaak, Roel GW, Beijen, M Antoinette, Erpelinck, Claudia AJ, van Doorn-Khosrovani, Sahar Barjesteh van Waalwijk, Boer, Judith M, Beverloo, H Berna, Moorhouse, Michael J, van der Spek, Peter J, Löwenberg, Bob, *et al.* . 2004. Prognostically useful gene-expression profiles in acute myeloid leukemia. *New England Journal of Medicine*, **350**(16), 1617–1628.

Venters, Bryan J, & Pugh, B Franklin. 2013. Genomic organization of human transcription initiation complexes. *Nature*, **502**(7469), 53–58.

Vittinghoff, Eric, & McCulloch, Charles E. 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, **165**(6), 710–718.

Von Luxburg, Ulrike. 2007. A tutorial on spectral clustering. *Statistics and computing*, **17**(4), 395–416.

Wagner, Andreas. 2002. Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Research*, **12**(2), 309–315.

Wand, MP, & Jones, MC. 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, **88**(422), 520–528.

Wang, Kai, Li, Mingyao, & Hakonarson, Hakon. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**(16), e164–e164.

Wang, Rui, Cherukuri, Pratima, & Luo, Jianyuan. 2005. Activation of Stat3 sequence-specific DNA binding and transcription by p300/CREB-binding protein-mediated acetylation. *Journal of Biological Chemistry*, **280**(12), 11528–11534.

Wang, Weihong, Lee, Sean Bong, Palmer, Rachel, Ellisen, Leif W, & Haber, Daniel A. 2001. A functional interaction with CBP contributes to transcriptional activation by the Wilms tumor suppressor WT1. *Journal of Biological Chemistry*, **276**(20), 16810–16816.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D.J., Campan, M., Young, J., Jacobs, I., *et al.* . 2006. Epigenetic stem cell signature in cancer. *Nature Genetics*, **39**(2), 157–158.

Williams, Kristine, Christensen, Jesper, Pedersen, Marianne Terndrup, Johansen, Jens V, Cloos, Paul AC, Rappsilber, Juri, & Helin, Kristian. 2011. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, **473**(7347), 343–348.

Wilson, James D, Wang, Simi, Mucha, Peter J, Bhamidi, Shankar, & Nobel, Andrew B. 2013. A Testing Based Extraction Algorithm for Identifying Significant Communities in Networks. *arXiv preprint arXiv:1308.0777*.

Witkowski, Leora, Carrot-Zhang, Jian, Albrecht, Steffen, Fahiminiya, Somayyeh, Hamel, Nancy, Tomiak, Eva, Grynspan, David, Saloustros, Emmanouil, Nadaf, Javad, Rivera, Barbara, *et al.* . 2014. Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nature Genetics*, **46**(5), 438–443.

Wolfe, Patrick J, & Olhede, Sofia C. 2013. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*.

Wu, Yongzhong, Diab, Iman, Zhang, Xueping, Izmailova, Elena S, & Zehner, Zendra E. 2004. Stat3 enhances vimentin gene expression by binding to the antisilencer element and interacting with the repressor protein, ZBP-89. *Oncogene*, **23**(1), 168–178.

Xie, Wei, Schultz, Matthew D, Lister, Ryan, Hou, Zhonggang, Rajagopal, Nisha, Ray, Pradipta, Whitaker, John W, Tian, Shulan, Hawkins, R David, Leung, Danny, *et al.* . 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**(5), 1134–1148.

Xu, Bin, Zeng, De-quan, Wu, Yuan, Zheng, Rong, Gu, Le, Lin, Xiao, Hua, Xianxin, & Jin, Guang-Hui. 2011. Tumor Suppressor Menin Represses Paired Box Gene 2 Expression via Wilms Tumor Suppressor Protein-Polycomb Group Complex. *Journal of Biological Chemistry*, **286**(16), 13937–13944.

Yang, Xiaojing, Han, Han, De Carvalho, Daniel D, Lay, Fides D, Jones, Peter A, & Liang, Gangning. 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell*, **26**(4), 577–590.

Yen, Ray-Whay Chiu, Vertino, Paula M, Nelkin, Barry D, Jane, J Yu, El-Deiry, Wafik, Cumaraswamy, Arunthathi, Lennon, Gregory G, Trask, Barbara J, Celano, Paul, & Baylin, Stephen B. 1992. Isolation and characterization of the cDNA encoding human DNA methyltransferase. *Nucleic Acids Research*, **20**(9), 2287–2291.

Yokomori, Norihiko, Tawata, Masato, Saito, Tukasa, Shimura, Hiroki, & Onaya, Toshimasa. 1998. Regulation of the rat thyrotropin receptor gene by the methylation-sensitive transcription factor GA-binding protein. *Molecular Endocrinology*, **12**(8), 1241–1249.

Zakany, J., & Duboule, D. 2007. The role of Hox genes during vertebrate limb development. *Current opinion in genetics & development*, **17**(4), 359–366.

Zanghi, Hugo, Picard, Franck, Miele, Vincent, Ambroise, Christophe, *et al.* . 2010. Strategies for online inference of model-based clustering in large and growing networks. *The Annals of Applied Statistics*, **4**(2), 687–714.

Zechner, U, Seifert, D, Schneider, E, El Hajj, N, Navarro, B, Kondova, I, Bontrop, RE, Bartsch, O, & Haaf, T. 2012 (November). Different DNA methylation of FOXP2 target genes in adult cortices of humans and chimpanzees. *Page 3266W of: Proceedings of the Annual Meeting of the American Society of Human Genetics*. American Society of Human Genetics.

Zhang, Zhixin, Jones, Simon, Hagood, James S, Fuentes, Nelson L, & Fuller, Gerald M. 1997. STAT3 acts as a co-activator of glucocorticoid receptor signaling. *Journal of Biological Chemistry*, **272**(49), 30607–30610.

Zhao, Yunpeng, Levina, Elizaveta, Zhu, Ji, *et al.* . 2012. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, **40**(4), 2266–2292.

Zhu, Jiang, Adli, Mazhar, Zou, James Y, Verstappen, Griet, Coyne, Michael, Zhang, Xiaolan, Durham, Timothy, Miri, Mohammad, Deshpande, Vikram, De Jager, Philip L, *et al.* . 2013. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*, **152**(3), 642–654.

Zhu, Wei-Guo, Srinivasan, Kanur, Dai, Zunyan, Duan, Wenrui, Druhan, Lawrence J, Ding, Haiming, Yee, Lisa, Villalona-Calero, Miguel A, Plass, Christoph, & Otterson, Gregory A.

2003. Methylation of adjacent CpG sites affects Sp1/Sp3 binding and activity in the p21Cip1 promoter. *Molecular and Cellular Biology*, **23**(12), 4056–4065.

Zhuang, J., Jones, A., Lee, S.H., Ng, E., Fiegl, H., Zikan, M., Cibula, D., Sargent, A., Salvesen, H.B., Jacobs, I.J., *et al.* . 2012. The Dynamics and Prognostic Potential of DNA Methylation Changes at Stem Cell Gene Loci in Women's Cancer. *PLoS Genetics*, **8**(2), e1002517.

Zou, H., & Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.