

Trans-Study Projection of Genomic Biomarkers in Analysis of Oncogene Deregulation and Breast Cancer

Dan Merl, Joseph E. Lucas, Joseph R. Nevins, Haige Shen & Mike West

Duke University

June 9, 2008

Abstract

In cancer studies as in many areas of human disease research, gene expression microarray technology has been central to the emergent field of genomic medicine. Expression profiles of physiological states and clinical outcomes play increasing roles as biomarkers in both experimental and human observational studies. Central challenges in moving towards clinical applications include hard questions of how to link and combine such measures across contexts: from laboratory experiments with cultured cells, to animal model experiments, to human outcome studies and clinical trials. The question of how to translate and transfer experimental, laboratory findings to the context of human observational studies sits at the core of current translational research agendas. This case study focuses on precisely this question in cancer genomics, where the *in vitro* laboratory results involve gene expression signatures of changes in human cells in response to a set of interventions on cancer related genes, the *oncogene intervention experiments*, and the *in vivo* context is gene expression studies with data generated from human breast tumours.

The analyses involve a series of applications of sparse Bayesian latent factor regression models, and are illustrative of the use of these models for large-scale multivariate data arising from both designed experiments and observational studies.

We exploit a range of posterior summaries from analysis of the oncogene intervention data to project the resulting *in vitro*-defined signatures of biological responses to the interventions into the *in vivo* data from breast cancers. Bayesian latent factor analysis of gene expression linked to these signatures in the breast cancer data then reveals the greater complexity of patterns of expression evidenced *in vivo*, and evolutionary model search links the oncogene signatures to a number of cancer-relevant biological pathways not initially represented in the experimental context. We follow this with a study of how latent factors estimated in the breast data project back to the oncogene experimental context, highlighting and providing interpretation of some of the inferred factors. Further, using posterior estimates of the latent factors as covariates, we examine Bayesian survival models for recurrence of breast cancer that identify several key latent factors that clearly have value as clinical biomarkers with respect to recurrence. Bayesian pathway annotation analysis provides clear evidence of the biological relevance of these factors by linking them to known biological pathways of relevance in cancer progression and development; beyond immediate interpretation, this has led to follow-on biological investigations.

This case study in integrative, trans-study Bayesian analysis of gene expression data sets is illustrative of the use of the overall strategy and approach – enabled by relevant Bayesian concepts, models and computational tools – in a number of other studies in genomics.

1 Oncogene Pathway Deregulation and Human Cancers

The overall biological focus of this study is to investigate the extent to which patterns of gene expression associated with artificially induced oncogene pathway deregulation can be used to query oncogene pathway activity in real human cancers. This is often referred to as the *in vitro* to *in vivo* translation problem: results drawn from controlled intervention studies using cancer cells cultured in clinical settings (*in vitro*) must be made relevant for cancer cells growing in the highly heterogeneous micro-environment of a living host (*in vivo*). We address this using Bayesian sparse factor regression analysis for model-based translation and refinement of *in vitro* generated signatures of oncogene pathway activity into the domain of human breast tumour tissue samples. The promise of this avenue of research is the ability to directly query the degree of functionality of relevant cellular pathways in such cancers, thereby facilitating individualized diagnosis, prognosis, and treatment (Nevins et al. 2003; Pittman et al. 2004; West et al. 2006). Our study here is an example application of an overall strategy developed through a series of studies and being utilised in a number of areas (Huang et al. 2003; Bild et al. 2006; Seo et al. 2007; Chen et al. 2007; Carvalho et al. 2008; Lucas et al. 2008).

1.1 Problem Context and Goals

Since the late 1990s, the increasing availability of microarray data across the spectrum of human cancers has created new opportunities for understanding the molecular basis of cancer pathogenesis. Microarray technology, as exemplified by the Affymetrix GeneChip platform, can provide effective, broad snapshots of the state of protein manufacture within a cell or tissue sample by measuring mRNA, the molecular precursor to protein, associated with each of some 25-30,000 known protein-coding genes in the human genome. The development of methodologies for associating observed manifestations of cancer progression with patterns extracted from such high-dimensional data has led to several key prognostic innovations over the past decade, including the identification of relatively small subsets of genes whose collective patterns of expression comprise protein-level “signatures” of clinically relevant phenotypes (West et al. 2001; Bild et al. 2006; Huang et al. 2002; Huang et al. 2003; Miller et al. 2005).

Such gene sets and their associated patterns of expression, referred to simply as signatures, are thought to represent core regulatory or inhibitory elements of one or more cellular pathways whose micro-level behavior is sufficient – though not necessarily necessary – for presentation of the macro-level phenotype. However, due to the complexity of cellular pathway interactions, the immediate applicability of any such signature beyond the context/data in which it is derived is always questionable. Various signatures have been derived *in vitro* from controlled experiments involving cell cultures subject to some combination of interventions whose effects on gene expression are of interest; good examples are targeted over-expression of a particular gene, or manipulation of the surrounding micro-environment of the cell culture. Prior to intervention, the cultured cells are in a state of quiescence, or inactivity, thus permitting all cellular reaction to the intervention – relative to some control group – to be attributed to the intervention. Though necessary for interpretability of the experiment, cells living outside of culture are rarely in such a state, and therefore were it possible to perform the same intervention on cells living *in vivo*, the observed response, in terms of gene expression, would almost certainly differ as a result of interactions with other normally active cellular pathways.

Our *in vitro* data concerns 9 experimentally derived signatures, each associated with the effects of over-expression of a different known oncogene involved in the progression of human cancer; we explore the projection of these signatures as a group into expression samples from a set of human

breast tumour tissues. Given that chemotherapies can be designed to block cancer progression by interrupting specific components of a targeted cellular pathway, it is of interest for purposes of drug discovery to evaluate the relative activity levels of these oncogene pathways in human breast cancer tissue samples, and furthermore to associate those pathway activity levels with clinical outcomes, such as long term survival. This aims to improve the ability to identify high-priority pathways, the disruption of which may be most likely to result in improved patient prognosis.

The key to the problem of *in vitro* to *in vivo* translation of the oncogene pathway signatures and their subsequent evaluation lies in allowing a more flexible notion of a biological “pathway”. Rather than attempting to evaluate the activity level of an entire oncogene pathway as it is observed in the experimental setting, we use Bayesian sparse factor analysis to simultaneously decompose and refine the original signatures in the context of human breast cancer samples. Through directed latent factor search, the original pathway signatures are mapped onto the tissue data and decomposed into smaller sub-pathway modules represented by latent factors; some factors may be common to multiple oncogene pathways, while some may represent sections of pathways uniquely linked to a particular oncogene. In this way, the original collection of intersecting oncogene pathways can be effectively modelled in the biologically more diverse and heterogeneous human cancer data context; analysis allows inference on the activity levels of each sub-pathway, so aiding focused study of any associations of sub-pathway activity with clinical outcomes.

1.2 *In Vitro* Oncogene Experiments

[Bild et al. \(2006\)](#) devised an experimental methodology for investigating the effects of artificially induced activation of oncogene pathways in human mammary epithelial cell cultures (HMECs). Quiescent cell cultures were infected with recombinant adenovirus constructs containing an oncogene insert. Upon replication in the cultured cells, the virus expresses the oncogene, thereby initiating within the otherwise inactive cell the sequence of responses to oncogene production that collectively comprises the oncogene pathway. After infection, RNA is extracted from the culture and prepared for subsequent gene expression analysis on the Affymetrix Human U133 2.0 Plus GeneChip microarray.

A total of 9 oncogene pathways were queried in this way, representing a variety of different aspects of cancer progression (Figure 1). The original study reported on the pathway signatures associated with the oncogenes human c-MYC, activated H-RAS, human c-SRC, human E2F3, and activated β -Catenin. Subsequently 4 additional interventions were conducted for the oncogenes P63, ATK, E2F1, and P110. Interventions were replicated across approximately 9 separate cultures per oncogene. Gene expression was also measured for 10 control cultures associated with the initial 5 interventions, and 5 control cultures associated with the latter 4 interventions, producing a total sample size of 97 microarrays, each quantifying hybridization levels of over 50,000 oligonucleotide probes.

1.3 *In Vivo* Breast Cancer Studies

[Miller et al. \(2005\)](#) presented a collection of 251 surgically removed human breast tumour tissue samples gathered during a study conducted between the years 1987 and 1989 in Uppsala County, Sweden. Following RNA extraction and processing, gene expression measurements for each tumour sample were obtained using the Affymetrix Human U133A and U133B GeneChip microarrays.

Several clinico-pathological variables were collected as part of standard surgical procedure, including age at diagnosis, tumour size, lymph node status (an indicator of metastases), and the

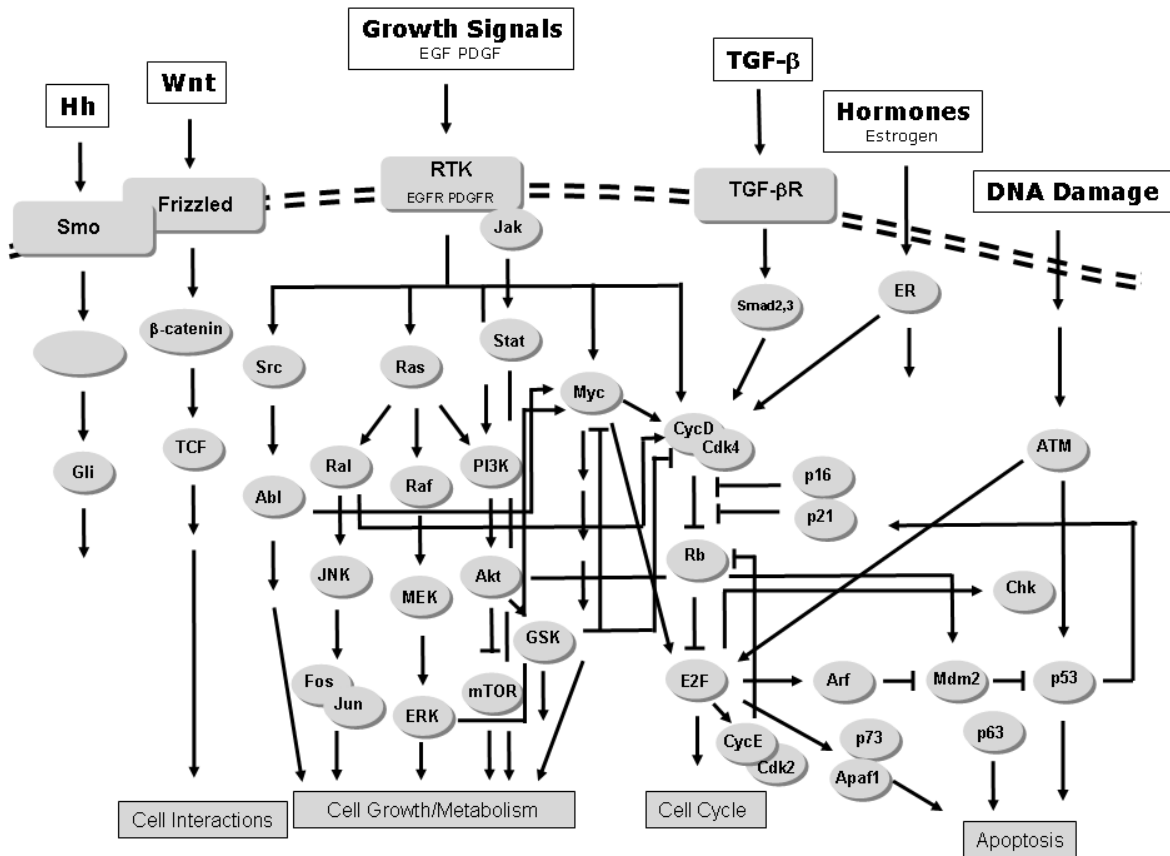


Figure 1: Schematic network structure depicting interactions between oncogenes in the context of the broader cell developmental and fate network. Arrows indicate biologically defined causal influences that are activating, all of which are heavily context dependent and may be direct, indirect and reliant on activation of other pathways, and on the roles of multiple other transcription factors, binding proteins and phosphorylating agents. The edges terminating as \perp indicate repressive links that are similarly well-defined and also similarly heavily context dependent.

Elston histologic grade of the tumour (a categorical rating of malignancy based on observed features of the tumour). Molecular assays were subsequently conducted to identify mutations in the estrogen receptor (ER), P53, and progesterone receptor (PgR) genes, the presence or absence of mutations in which have formed the basis of classification into several breast cancer subtypes. Patient survival histories were also monitored following surgery.

2 Modelling and Data Analysis

The flow of analysis is as follows. Section 2.1 concerns evaluation of the oncogene pathway signatures from the Bild data using sparse multivariate regression (Lucas et al. 2006). Section 2.2 defines initial pathway signature scores for each oncogene on each tumour tissue sample, providing measures of the activity level of each pathway. These signature scores, that represent pathway activity in the same way that gene expression represents activity of a single gene, are then used as seeds for evolutionary latent factor search (Carvalho et al. 2008), described in Section 2.3. By initializing the search in this way, the 9 leading factors are by construction associated the original oncogene pathway signatures, and subsequent factor discovery can be viewed as providing refinement of the fit of the signature-inspired factors to the tumour data. Section 2.4 develops and details exploratory analysis of the fitted latent factor model, and Section 2.5 takes further the *in vitro-in vivo* comparison by back-projecting the estimated factors to the experimental oncogene context. Section 2.6 incorporates posterior estimates of latent factors in the breast data as candidate covariates, together with traditional clinical variables, in a cancer survival analysis. This uses Bayesian survival analysis and shotgun stochastic search (SSS) to explore model uncertainty and identify factors predictive of clinical outcomes (Hans et al. 2007; Hans et al. 2007). Section 3 explores biological interpretation of a few key factors emerging from this analysis, and links them to known biological pathways and oncogenic phenomena using summary results from our Bayesian probabilistic pathway annotation analysis, or PROPA (Shen 2007; Shen and West 2008).

2.1 Generation of Oncogene Pathway Signatures

Evaluation of oncogene pathway signatures from the Bild et al. (2006) data is based on sparse multivariate regression analysis (Lucas et al. 2006) applied to the RMA expression values on $p = 8,509$ probes (referred to from here on as genes) showing non-trivial variation across the $n = 97$ samples. A key aspect is the use of sparsity priors. It is a biological surety that a relatively small number of genes are expected to be involved in any one oncogene pathway, and to therefore show differential expression in response to activation of that pathway. Sparse regression augments a standard linear model with hierarchical point mass mixture priors on regression coefficients (West 2003). This allows identification of the subset of differentially expressed genes that show a response to each intervention, quantified by the posterior probability of a non-zero regression coefficient for the gene \times intervention effect. For each intervention, the set of genes for which this probability is high, along with the estimated effects, comprise the pathway signature gene set.

Let X^{vitro} denote the $8,509 \times 97$ dimensional gene expression matrix, with elements $x_{g,i}^{vitro}$ measuring expression of gene g on array sample i . Expression is on the \log_2 scale, so a unit change in x corresponds to a doubling or halving, i.e., a one-fold change in mRNA levels. Let H^{vitro} denote the 18×97 design matrix where the first 10 rows contain binary indicators associating samples with their intervention effects (9 oncogene effects and 1 for the second control group), and the final 8 rows are covariates constructed as artifact correction factors. The latter, derived from the first 8 principal components of the Affymetrix housekeeping/control probe expression levels

on each sample array, are for gene-sample specific normalisation. The construction and efficacy of these artifact control covariates in aiding model-based, automatic correction for the effects of experimental and assay biases, has been abundantly demonstrated (Carvalho et al. 2008) and is crucial in the context of this analysis where the two separate regimes of experimentation led to very major artifactual differences Lucas et al. (2006).

The regression model for each gene $g = 1 : p$ on any sample $i = 1 : n$ is

$$x_{g,i}^{vitro} = \mu_g + \sum_{k=1}^{18} \beta_{g,k} h_{k,i}^{vitro} + \nu_{g,i},$$

or in matrix form,

$$X^{vitro} = \mu \iota' + BH + N \quad (1)$$

where the element μ_g of the $p \times 1$ vector μ is baseline expression of gene g , ι is the $n \times 1$ vector of ones, the elements $\beta_{g,k}$ of the $p \times 18$ matrix B are the effects of interventions and coefficients of artifact correction covariates, and the element $\nu_{g,i}$ of the $p \times n$ error matrix N represents independent residual noise for gene g on sample array i . Careful consideration is given to assignment of prior distributions. Baseline expression effects are modelled as $\mu_g \sim N(8, 100)$, based on known properties of Affymetrix RMA gene expression indices. Also, based on years of experience with thousands of Affymetrix samples in dozens of studies, noise terms are modelled as $\nu_{g,i} \sim N(0, \psi_g)$ with $\psi_g \sim IG(2, 0.1)$; this reflects the view that noise standard deviations will be gene-specific and range between 0.05 – 0.7 with median values across genes near 0.2 – 0.25. The hierarchical sparsity prior on elements $\beta_{g,k}$ of B is

$$\begin{aligned} \beta_{g,k} &\sim (1 - \pi_{g,k})\delta_0 + \pi_{g,k}N(0, \tau_k) \quad \text{with} \quad \tau_k \sim IG(1, 5), \\ \pi_{g,k} &\sim (1 - \rho_k)\delta_0 + \rho_k Be(9, 1) \quad \text{with} \quad \rho_k \sim Be(2.5, 497.5). \end{aligned} \quad (2)$$

Here δ_0 represents a point mass at zero. Thus the effect of regressor k on gene g is non-zero and drawn from the $N(0, \tau_k)$ prior with probability $\pi_{g,k}$. The variance τ_k is regressor-specific, reflecting the fact that levels of gene activation/suppression will vary across pathways. The choice of prior for τ_k reflects expected ranges of expression changes supporting several orders of magnitude, reflecting relevant biological expectations and again based on extensive prior experience with similar data sets. The prior structure on the π and ρ terms govern the expected sparsity of the pathway signatures and reflect the prior view that approximately 99.5% of all monitored genes will show no response to a given intervention, and the pathway inclusion probability of all such genes should be exactly 0. This ensures that the data must provide overwhelming evidence of expression change for a gene to achieve high posterior probability of inclusion in the pathway.

Model fitting uses the highly optimized BFRM Markov chain Monte Carlo (MCMC) algorithms now widely applied in such models (Wang et al. 2007); see Appendix 3.2 for more details and links to software. The analysis is based on 25,000 posterior samples following an initial burn-in of 2,500. All summaries and further evaluation in this study are based on the posterior means of model parameters generated from BFRM.

Figure 2 depicts the sparsity structure of the estimated oncogene pathway signatures (columns 2 through 10 of B) using a posterior probability threshold of 0.95 (i.e. $Pr(\beta_{g,k} \neq 0 | X^{vitro}) > 0.95$). At this threshold level, the RAS pathway involves the largest number of probes (2,202), followed by the P110 pathway (1,630), with the SRC pathway involving the smallest number of probes (417). The size of the RAS signature gene set indicates significant opportunity for modular decomposition of this pathway via latent factor analysis, as the downstream effects of intervention on RAS clearly involve many genes and therefore, presumably, a large number of intersecting pathways.

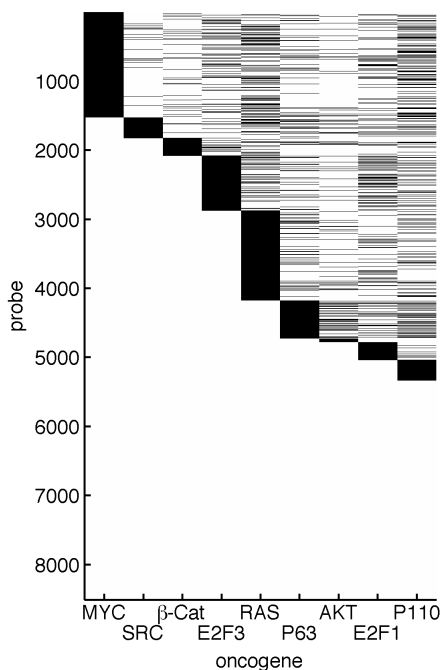


Figure 2: Estimated sparsity patterns of oncogene pathway signatures. Black indicates genes \times oncogene intervention pairs g, k such that $P(\beta_{g,k} \neq 0 | X^{vitro}) > 0.95$. Genes are ordered to highlight the relative size of the pathway signature gene sets across oncogenes, and also the cross-talk in terms of genes appearing responsive in multiple pathways.

2.2 Initial Quantification of Pathway Activity by Signature Projection

Translation to the breast cancer data is eased due to compatibility of the microarrays used, as the $p = 8,509$ probes from the U133 2.0 Plus array are all present on the U133AB arrays on which the breast data are generated. Let X^{vivo} denote the matrix of RMA measures of gene expression for the tumour data. Initial measures of the levels of activation of the 9 oncogene pathways in each of the breast tumour samples are calculated using posterior means of parameters, denoted by $\beta_{g,k}^{vitro}$ and ψ_g^{vitro} and estimated based on the MCMC results. The raw *projected signature score* of pathway k in tumour i is defined, following Lucas et al. (2008), as $s_{k,i} = \sum_{g=1}^p \beta_{g,k}^{vitro} x_{g,i}^{vivo} / \psi_g^{vitro}$. To map to the same scale as genes in the breast expression data, each signature score is then simply transformed to match the sample means and variances, viz $s_{k,i}^* = m + u(s_{k,i} - m_k) / u_k$ where m_k, u_k are the mean and standard deviation of $s_{k,i}$ values over the tumour samples $i = 1 : 251$, m is the sample average gene expression over all genes and tumour samples, and u is the average of the set of p gene-specific standard deviations of gene expression across tumour samples. This mapping is only for ease of visual comparisons and serves no technical purpose otherwise.

Signature scores for the tumour tissue samples are shown in Figure 3. Clear gradients emerge across MYC, SRC, β -Catenin, RAS, and P63 signature scores when samples are ordered by rank along the first principle component, indicating possible patterns of activity of these 5 oncogene pathways across breast cancers. High correlations exist between MYC and β -Catenin status, and between SRC and RAS status, with the former two negatively correlated with the latter. Latent factor analysis will provide the methodology for uncovering the common structure underlying these clearly inter-connected pathways.

2.3 Latent Factor Models for Breast Tumour Gene Expression

Latent factor analysis of breast expression data on a set of genes that are related to the oncogene pathways aids in the development of an understanding of the complexity of patterns of expression

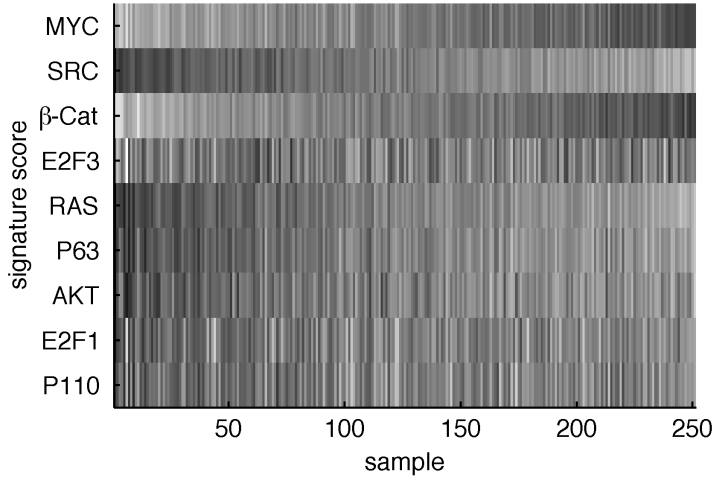


Figure 3: Initial measures of oncogene pathway activity in tumour samples. Black indicates low signature scores, or pathway suppression; white indicates higher signature scores, or pathway activation. Tumor tissue samples have been reordered to emphasize dominant patterns of pathway activity, especially the correlated gradients of MYC and β -Catenin status, and SRC and RAS status.

linked to those pathways when evidenced in the *in vivo* context of human tumours. The now standard sparse latent factor extension of the earlier multivariate regression model (West 2003; Lucas et al. 2006; Wang et al. 2007; Carvalho et al. 2008) has the form

$$x_{g,i} = \mu_g + \sum_{k=1}^r \beta_{g,k} h_{k,i} + \sum_{\ell=1}^s \alpha_{g,\ell} \lambda_{\ell,i} + \nu_{g,i} \quad (3)$$

where each h_i now contains any known design or covariate data of interest in the breast cancer analysis, and the $\alpha_{g,\ell}$ are coefficients that define the *loadings* of gene g on the values $\lambda_{\ell,i}$ on a set of s latent factors evidence across the $i = 1 : n$ tumours; the value $\lambda_{\ell,i}$ is referred to as the *factor score* for factor ℓ on sample i . The general structure of the priors for μ , B , and ν terms remains as above. The prior structure on latent factor loadings $\alpha_{g,k}$ is of the same form as that for the $\beta_{g,k}$, as the same notion of expected sparsity applies; multiple factors represent the complexity of expression patterns in tumours, but genes will be only selectively related to factors as described by a sparsity prior over the factor loadings. One technical modification is that the first $s \times s$ block of the implied loading matrix is constrained to have positive diagonal values and an upper triangle of zeros; this ensures model identification and identifies the first s variables as named *founders* of the s factors.

In order to represent non-Gaussian distributions of factor scores, as well as to facilitate clustering of samples based on estimated factor profiles, the $s \times 1$ vector of latent factor scores $\Lambda_{:,i}$ on tumour sample i are modelled jointly via a Dirichlet process mixture model (Carvalho et al. 2008); that is,

$$\begin{aligned} \Lambda_{:,i} &\sim F(\cdot) \quad \text{with} \quad F \sim DP(\alpha_0 G_0), \\ \alpha_0 &\sim Ga(e, f) \quad \text{and} \quad G_0 = N(0, I) \end{aligned} \quad (4)$$

where I is the $s \times s$ identity.

A key novelty of the subsequent analysis lies in applying the above latent factor model in such a way as to preserve the connection between identified factors and aspects of the oncogene pathways. We do this by defining the extended data matrix that has the 9×251 matrix S^* of projected signature scores as its first 9 rows, viz

$$X^{vivo*} = \begin{pmatrix} S^* \\ X^{vivo} \end{pmatrix},$$

now a $44,601 \times 251$ matrix. To ease notation, the model equation (3) can be reexpressed as

$$x_{g,i}^{vivo*} = \mu_g + \sum_{\ell=1}^r \alpha_{g,\ell} \lambda_{\ell,i} + \sum_{\ell=r+1}^{s+r} \alpha_{g,\ell} \lambda_{\ell,i} + \nu_{g,i} \quad (5)$$

or in matrix form,

$$X^{vivo*} = \mu\mu' + A\Lambda + N \quad (6)$$

where the first r columns of A are the regression coefficients associated with the first r rows of Λ , the known design effects, and the remaining columns relate to the latent factors. The only covariates used are 4 artifact control covariates are in the first 4 rows of Λ , again being computed from the housekeeping genes on the tumour sample arrays.

Prior distributions are as follows. Baseline expression $\mu_g \sim N(8, 100)$, again reflecting known properties of Affymetrix expression indices, while $\nu_{g,i} \sim N(0, \psi_g)$ with $\psi_g \sim IG(2, 0.005)$ reflecting the view that residual noise should be somewhat less than that in the oncogene signature analysis due to the increased opportunity for explaining variance via an arbitrary number of latent factors. Indeed, some latent factors may reflect patterns of shared variation that are due to experimental artifacts above those captured by the artifact control covariate strategy. The sparsity prior over elements of A is

$$\begin{aligned} \alpha_{g,\ell} &\sim (1 - \pi_{g,\ell})\delta_0 + \pi_{g,\ell}N(0, \tau_\ell) \quad \text{with} \quad \tau_\ell \sim IG(2, 1), \\ \pi_{g,\ell} &\sim (1 - \rho_\ell)\delta_0 + \rho_\ell Be(9, 1) \quad \text{with} \quad \rho_\ell \sim Be(0.2, 199.8). \end{aligned} \quad (7)$$

The prior is modified for $g = 1 : s$ in view of the identifying constraints that $\alpha_{g,r+g} > 0$ and $\alpha_{g,\ell} = 0$ when $g < \ell - r$ for $\ell = (r + 2) : (r + s)$. This is done simply by taking $\alpha_{g,r+g} \sim N(0, \tau_g)$ truncated to $\alpha_{g,r+g} > 0$ for $g = 1 : s$, and fixing $\pi_{g,\ell} = 0$ for $g < \ell - r, \ell = (r + 2) : (r + s)$.

The prior on τ_ℓ anticipates that variation explained by any single factor will tend to be rather less than that explained by the design effects of the oncogene study, though hedged by considerable uncertainty. The prior on inclusion probabilities now reflects an assumption that approximately 99.9% of genes will show no association with a given factor, though this assumption is enforced with less certainty than that of the 99.5% assumption made in the oncogene study. The prior on the Dirichlet process precision α_0 of equation (4) is a standard $Ga(1, 1)$ reflecting considerable uncertainty about the complexity of structure expected in the factor distribution.

Latent factor discovery and estimation is accomplished using the same BFRM software (Wang et al. 2007) used for the *in vitro* oncogene analysis. For latent factor analysis, the software implements a novel evolutionary search algorithm for incrementally growing and fitting the factor model, based on specification of an initial set of genes and factors. The idea is that most of the 40,000+ probes in the full dataset will show no association with any design variable or latent factor relevant to the set of oncogene pathways, and thus only genes that do show associations should be incorporated into the model; genes not incorporated into the model can be viewed as having $\pi_{g,\ell} \approx 0$ for all ℓ . At each step of the iterative *evolutionary search* (Carvalho et al. 2008), gene inclusion probabilities are calculated for all unincorporated genes, and those with highest inclusion probabilities become candidates for inclusion in the model. Increasing the set of genes within the model can then introduce additional structure in the expression data that requires expansion of the number of latent factors, and hence model refitting. Analysis terminates after a series of such steps to expand the set of genes and number of factors, subject to some specified controls on the numbers of each and guided by thresholds on gene \times factor inclusion probabilities. Finally, the

factor loadings and scores are estimated by MCMC for the final model (refer to Appendix 3.2 for references and more details on BFRM).

Analysis is initiated with a 9-factor model incorporating only the oncogene signature scores as the initial data set. After evolutionary analysis this will expand to a model with at least 9 factors and genes that are added that, initially, are most highly related to the evolving estimates of these 9 factors. Thus, in an expanded model the loadings of the initial 9 factors correspond to the changes in gene expression of those genes most associated with changes in signature scores in the tumour tissues. Factors beyond the first 9 will successively improve model fit by accounting for variation in gene expression beyond that explicitly linked with the patterns of pathway activation represented by the signature scores, thus incrementally augmenting and refining the first 9 core pathway factors.

2.4 Exploring Latent Factor Structure in the Breast Data

BFRM evolved the 9-gene, 9-factor model to ultimately include 500 genes (the maximum we allowed) in a model with 33 estimated factors. This final model is an expanded latent factor representation of gene expression patterns that (a) link to the initial oncogene signatures, (b) evidence the greater complexity seen in the tumour data in genes defining the oncogene pathways, (c) link in numerous genes that relate to the factor representation of these initial pathways as well as (d) many other genes that reflect biological activity in intersecting pathways in the broader gene network that the initial oncogenes play roles in. Since the breast data analysis allowed exploration of over 40,000 probes on the microarray, it had the potential to identify genes that were not in the 8,509 used in the oncogene analysis; at termination, the model identified a set of 213 probes among the 8,509 and an additional 287 not appearing in the *in vitro* analysis.

Figure 4 shows the intensity image of corrected gene expression for these 500 genes, ordered by ranking along first principle components in order to accentuate dominant structure. Corrected expression is calculated by subtracting baseline expression and the contribution of artifact control factors from the raw expression values, using posterior means of model parameters; i.e., $X^{vitro*} - \mu' - \alpha_{:,1:4}\Lambda_{1:4,:}$, where posterior means are used for μ , A , and Λ . A weak structure emerges, with some 100 strongly differentially expressed genes characterizing a subgroup of samples (probes 1–25 and 475–500 approximately). Inspection of clinical data on these tumours reveals that this pattern of expression defines a high risk tumour subgroup, correlated with ER negativity, Elston grade 3, P53 positivity, and PgR negativity (Figure 5). This simply serves to demonstrate that the relatively small subset of genes incorporated into the factor model is sufficient to begin to discriminate clinically relevant tumour subtypes from one another, though the analysis was purely expression-data based and not at all trained on the clinical data. This gives initial strength to the view that global summaries, in terms of inferred factors, in expression related to key oncogenic pathways may emerge as relevant in defining biomarkers of clinical use.

The structure of the latent factor loadings matrix gives additional insight to the factor basis of this high risk subgroup. The left frame of Figure 6 depicts the “factor skeleton” identifying high probability gene-factor relationships in terms of non-zero loadings. The emphatic ordering of probes also serves to indicate blocks of genes that became incorporated upon addition of each new factor, including many that are distant from the initial 9 factors. Factor 1, associated with the MYC signature, is the most heavily loaded factor, with 223 nonzero loadings. Factor 7, associated with the RAS signature, and factors 20, 22, and 28 are relatively sparsely loaded. Replotting the image of corrected gene expression with probes now ordered by factor incorporation (Figure 6) reveals that many of the genes defining the clinically high-risk cancer subgroup are those heavily associated with latent factor 19.

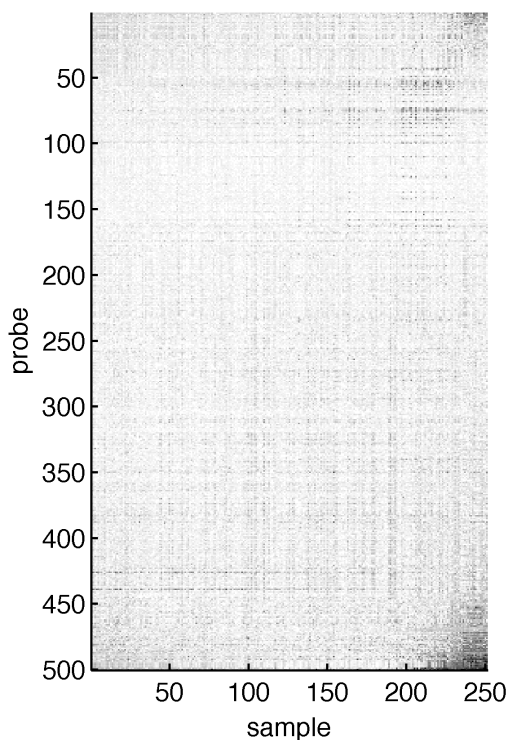


Figure 4: Artifact corrected gene expression of breast tumour samples for the 500 genes incorporated into the latent factor model. Darker shades indicate higher expression than baseline. Probes and samples have been ordered by rank along first principle components in order to emphasize patterns of expression. The patterns of expression associated with the first and last 50 probes characterize a subgroup of approximately 50 tumours samples having clinical variables indicative of poor prognosis.

Figure 7 displays estimated posterior means of the factor scores of factor 19 on tumours, coded by clinical outcomes and ordered as Figure 4. The figure demonstrates an important aspect of latent factor analysis: the need for non-Gaussian modelling of latent factors. Representation of the bimodal nature of factor scores, and thus the clear separation of distinct tumour subgroups, emerges naturally under the Dirichlet process mixture prior on factor scores, but would be nearly impossible under a normal model.

Comparing estimated factor scores for the first 9 factors and their associated oncogene signature scores (Figure 8) illustrates varying levels of “intactness” of the original pattern of pathway activity as represented by the signature scores. High correlation, as is the case for MYC and E2F1, indicates that the patterns of MYC and E2F1 pathway activity predicted by their *in vitro* signature scores are still largely captured by the single factors 1 and 6, respectively. The relationship between factor 4 and the SRC signature score indicates that pattern of SRC pathway activity predicted by the signature score is not evident at all in the tumour tissue samples, implying that the SRC pathway may not play a significant role in characterising breast cancers. The relationship between factor 7 and the RAS signature score indicates that very little of the pattern of initial RAS pathway activity remains explained by factor 7, though the estimated scores for factor 7 show some variation (relative to that of the factor 4) implying that the RAS pathway may be active but that activity is now captured as the combined effect of multiple factors (as hypothesized above). We see this further below. We can also reflect the contrast between *in vitro* signatures and their immediate *in vivo* factor counterparts by comparing estimated weights on all genes defining the signature scores with the estimated loadings on the factor counterparts (Figure 9) with similar conclusions.

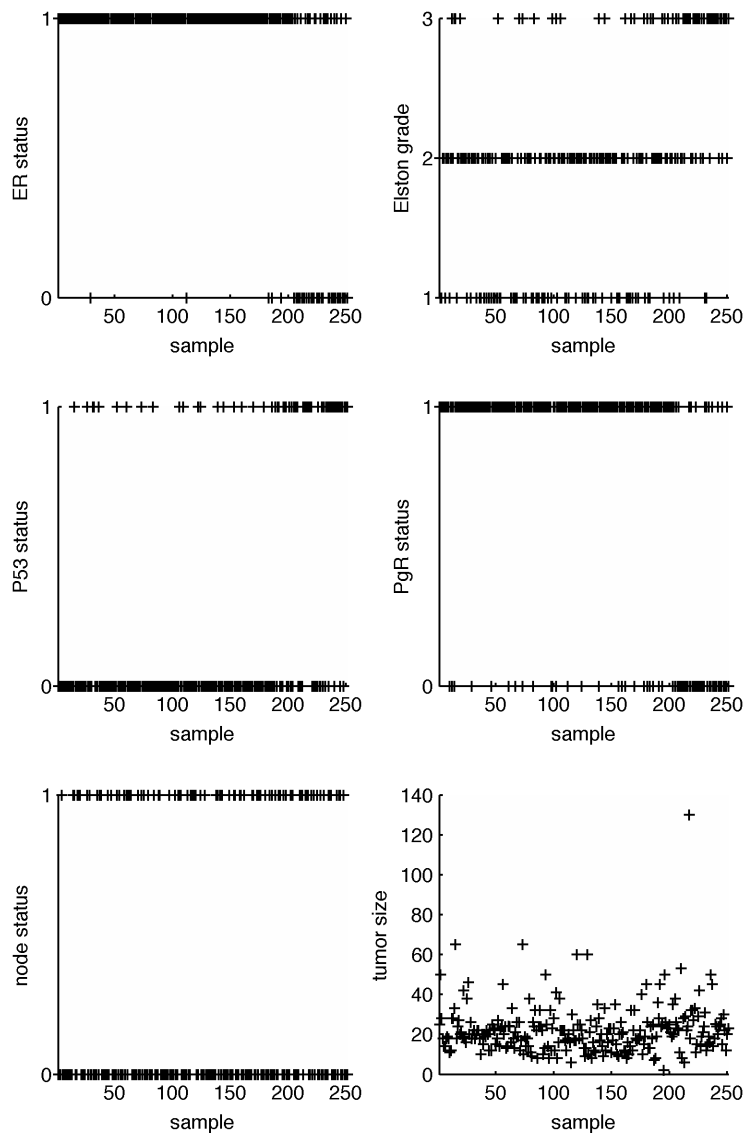


Figure 5: Clinical variables associated with tumour tissue samples, with samples ordered as in Figure 4. The subgroup consisting of samples 200-250 (approximately) is consistently associated with ER negative, Elston grade 3, P53 positive, PgR negative tumours, all of which are primary indicators of tumour aggressiveness/malignancy. The subgroup does not have an obvious correlation with lymph node status or tumour size.

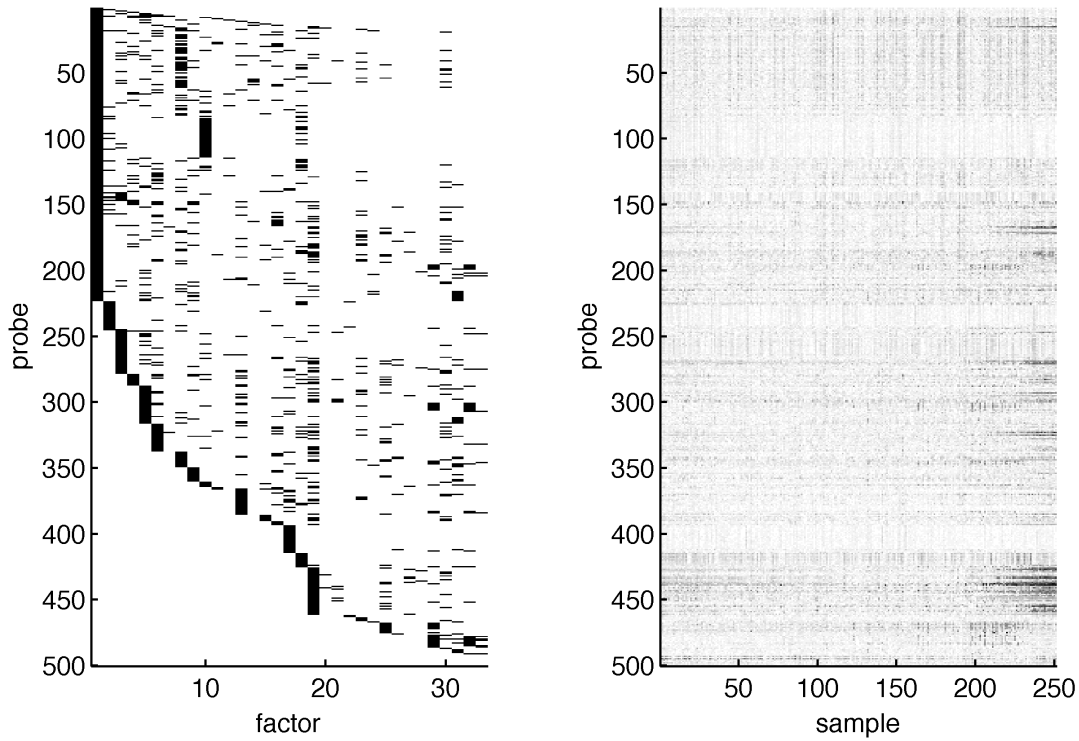


Figure 6: *Left frame:* Sparsity structure of latent factors discovered during analysis of tumour tissue data. Black bars indicate genes with significant factor loadings in terms of $P(\alpha_{g,l} \neq 0 | X^{vivo}) > 0.95$. Probes have been reordered to emphasize structure of loadings. The first 9 factors are by construction associated with the MYC, β -Catenin, AKT, SRC, P63, E2F1, RAS, P110, and E2F3 signatures respectively. Note that the RAS pathway factor (factor 7) now involves significantly fewer genes than the original RAS signature. *Right frame:* Artifact corrected gene expression of breast tumour samples for the 500 genes incorporated into the latent factor model, with samples ordered as in Figure 4 and probes now ordered as in the factor skeleton in the left frame here. Darker shades indicate expression higher than baseline. This ordering reveals that the set of approximately 50 differentially expressed genes characterizing the high-risk tumour subgroup are some of those heavily associated with factor 19.

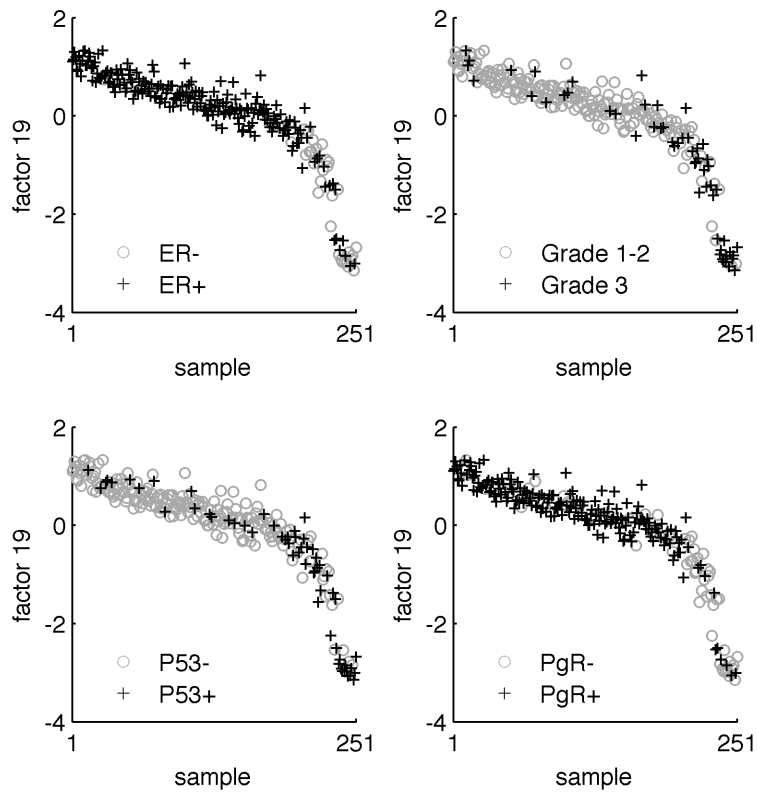


Figure 7: Factor scores associated with latent factor 19 (i.e. $\Lambda_{19,:}$), coded by ER, Elston Grade, P53, and PgR status and with samples ordered as in Figure 4. Decreased factor 19 scores show a clear association with ER negative, Elston grade 3, P53 positive, PgR negative tumours.

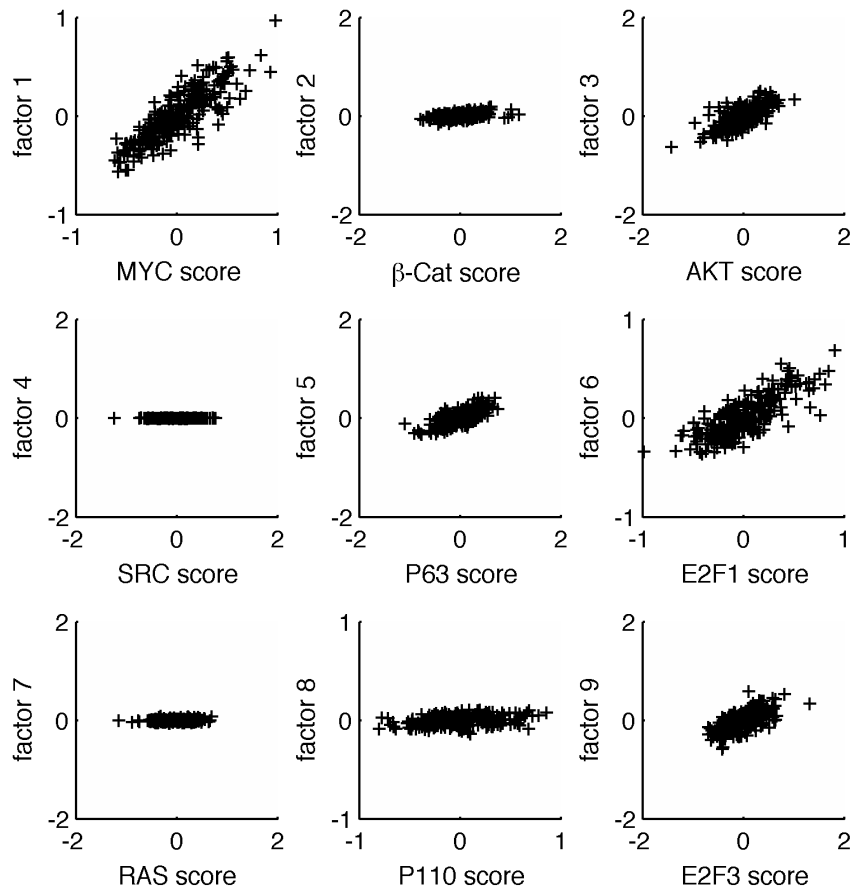


Figure 8: Scatter plots depicting correlations between original oncogene signature scores and associated posterior latent factor scores ($\Lambda_{1:9,\cdot}$). Higher correlations indicate higher levels of “intactness” of original pattern of pathway activity. MYC pathway activity in the breast tumour samples appears to be largely as predicted by the original MYC signature score, while SRC pathway activity in the breast tumour samples appears to be inactive, despite the predicted variation indicated by by the original SRC signature scores.

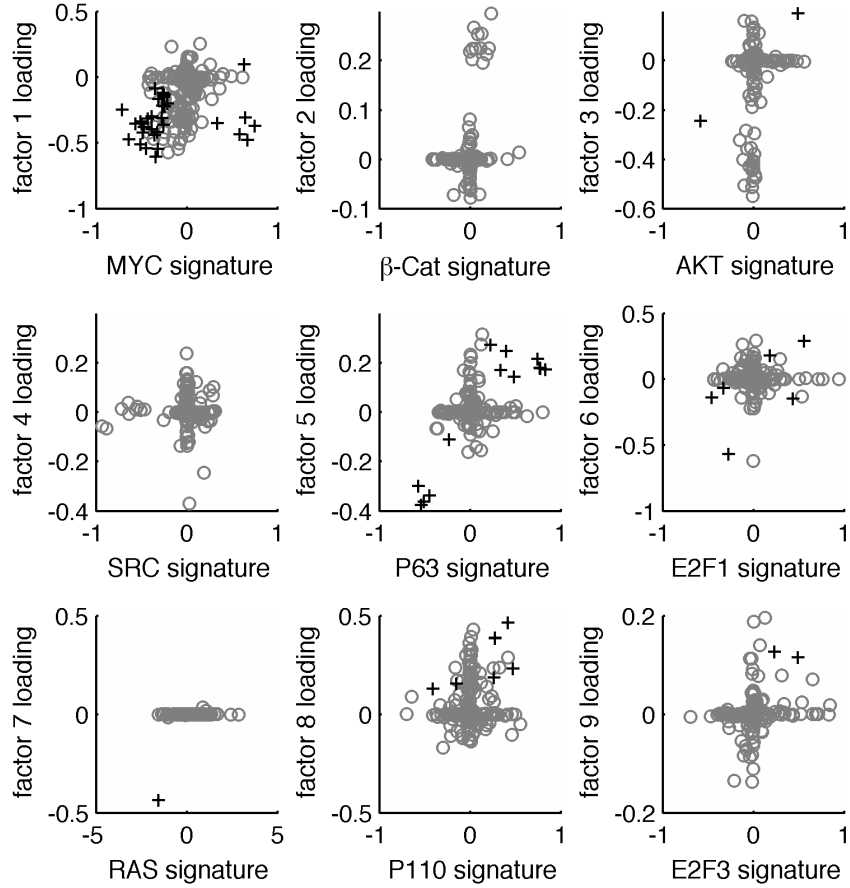


Figure 9: Scatter plots of gene-specific weights in the oncogene signatures ($\beta_{g,k}^{vitro} / \psi_g^{vitro}$) and associated posterior latent factor loadings ($\alpha_{g,\ell}$) for the 213 genes common to both the *in vitro* sparse regression and *in vivo* latent factor analyses. Grey circles indicate genes with probabilities < 0.95 of nonzero weights in at least one study, and black crosses indicate genes with probabilities > 0.95 of nonzero weights in both studies. The specific genes involved in each oncogene-associated latent factor largely differ from those involved in the original oncogene signature, thus indicating the degree to which the levels of biological activation of these genes in the *in vivo* pathway differs from those in the *in vitro* context.

2.5 Projection of Breast Tumour Latent Factors from *In Vivo* to *In Vitro*

It is possible to establish more concrete connections between latent factors and oncogene pathway activity via the same strategy as was used to produce the initial signature scores, but now in reverse. That is, to now regard each sample in the oncogene intervention study as a new datum in the tumour data context and to impute factor scores associated with each of the *in vitro* samples for the inferred set of latent factors. This was done using modified expression values $x_{g,i}^{vitro}$ from the oncogene regression model analysis now with the estimated contributions of the baselines and artifact control covariates subtracted from the raw data values. Write $x_{:,i}^{vitro}$ for the resulting vector on oncogene sample i , noting that we need to restrict to the 213 genes that are in common in the two analyses. Denote by A_* only the 251×33 component of the *in vivo* loadings matrix related to these genes and the 33 factors, and the corresponding residual variances as the elements of the diagonal matrix Ψ_* .

The imputed factor score vectors for these *in vitro* samples are calculated, following [Lucas et al. \(2008\)](#), as approximate posterior predictive means of latent factors $\Lambda_{:,i}$ under the Dirichlet process prior; that is,

$$\Lambda_{:,i}^{vitro} = c_{i,0}f_i + \sum_{j=1}^{251} c_{i,j}\Lambda_{:,j}$$

where

$$f_i = (I + A_*' \Psi_* A_*)^{-1} A_*' \Psi_* x_{:,i}^{vitro}$$

while

$$c_{i,0} \propto \alpha_0 N(x_{:,i}^{vitro} | 0, A_* A_*' + \Psi_*) \quad \text{and} \quad c_{i,j} \propto N(x_{:,i}^{vitro} | A_* \Lambda_{:,j}, \Psi_*), \quad (j = 1 : 33),$$

and the $c_{i,j}$ sum to one over $j = 0 : 251$.

Figures 10, 11, and 12 present the imputed factor scores for all 33 discovered factors across the complete set of observations from the oncogene study. Among the first 9 factors, factors 1, 5, 6, 7, and 9, retain clear associations with their respective founding oncogene pathways. Across all 33 factors, 20 different factors, including factor 19, separate the RAS activated subgroup from the others, illustrating the degree to which the original RAS pathway signature has been dissected into constituent sub-pathways and also emphasising the major role RAS plays in cancer gene networks as reflected in complexity of expression patterns. Some oncogene subgroups, such as β -Catenin and SRC, are not distinguished by any factors, indicating these pathways may play less of a role in explaining patterns of gene expression in tumour tissue than does the RAS pathway. Many of the factors identify multiple oncogene subgroups, i.e., likely intersections between pathways.

2.6 Factor-based Prediction of Clinical Outcome

For over a decade a driving interest in expression genomics has been in the potential for expression-based biomarkers of prognostic and diagnostic clinical use ([West et al. 2001](#); [van't Veer et al. 2002](#); [Nevins et al. 2003](#); [Chang et al. 2004](#); [Pittman et al. 2004](#); [Miller et al. 2005](#); [Bild et al. 2006](#); [West et al. 2006](#); [Lucas et al. 2008](#)), and we have already seen here that an estimated biomarker – factor 19 – strongly associates with some of the central clinical markers in regard to breast cancer recurrence risk. Identification of clinically relevant *pathways* represented by discovered factors can be explored by considering statistical models that use estimated factor scores as candidate covariates in models to predict clinical outcomes; any factors found to play significant roles in such models will warrant further investigation. We do this now in connection with survival outcomes in

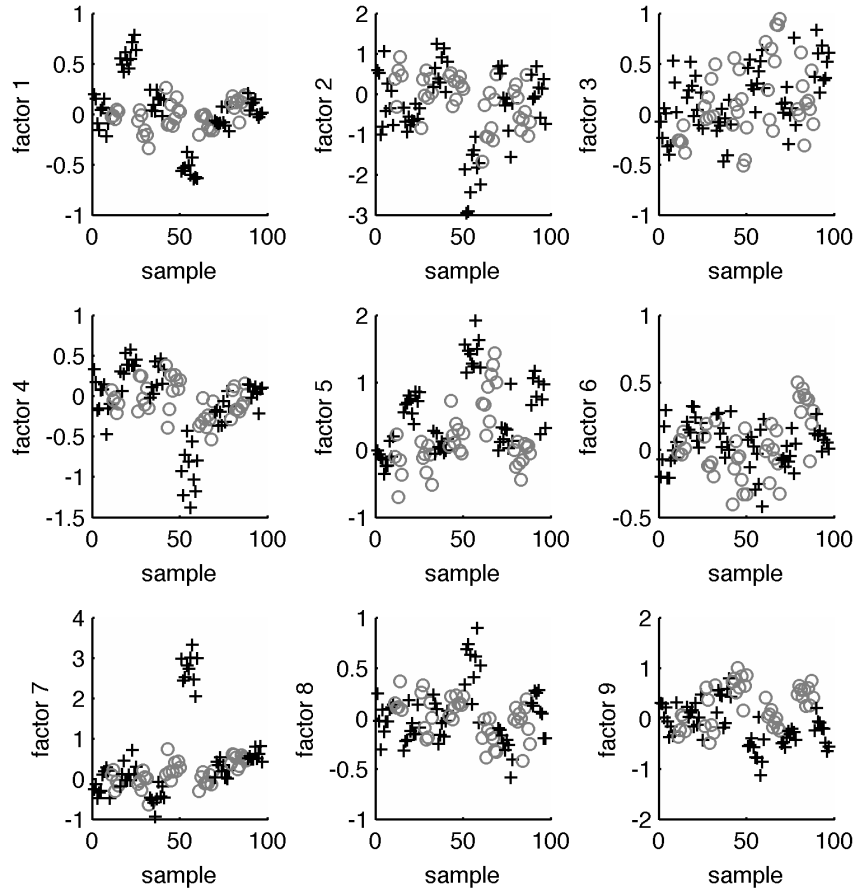


Figure 10: Imputed latent factor scores for oncogene intervention data. Latent factors 1–9 are those founded by the MYC, β -Catenin, AKT, SRC, P63, E2F1, RAS, P110, E2F3 oncogene signature scores respectively. Oncogene interventions are ordered as control 1, control 2, MYC, SRC, β -Catenin, E2F3, RAS, P63, AKT, E2F1, P110, with samples from the same experimental intervention appearing consecutively, designated by the same color/marker. Separate interventions are depicted by alternating black crosses and grey circles. Note the factor-based separation of the RAS activated experimental group achieved by factors 1, 2, 4, 5, 7, and 8.

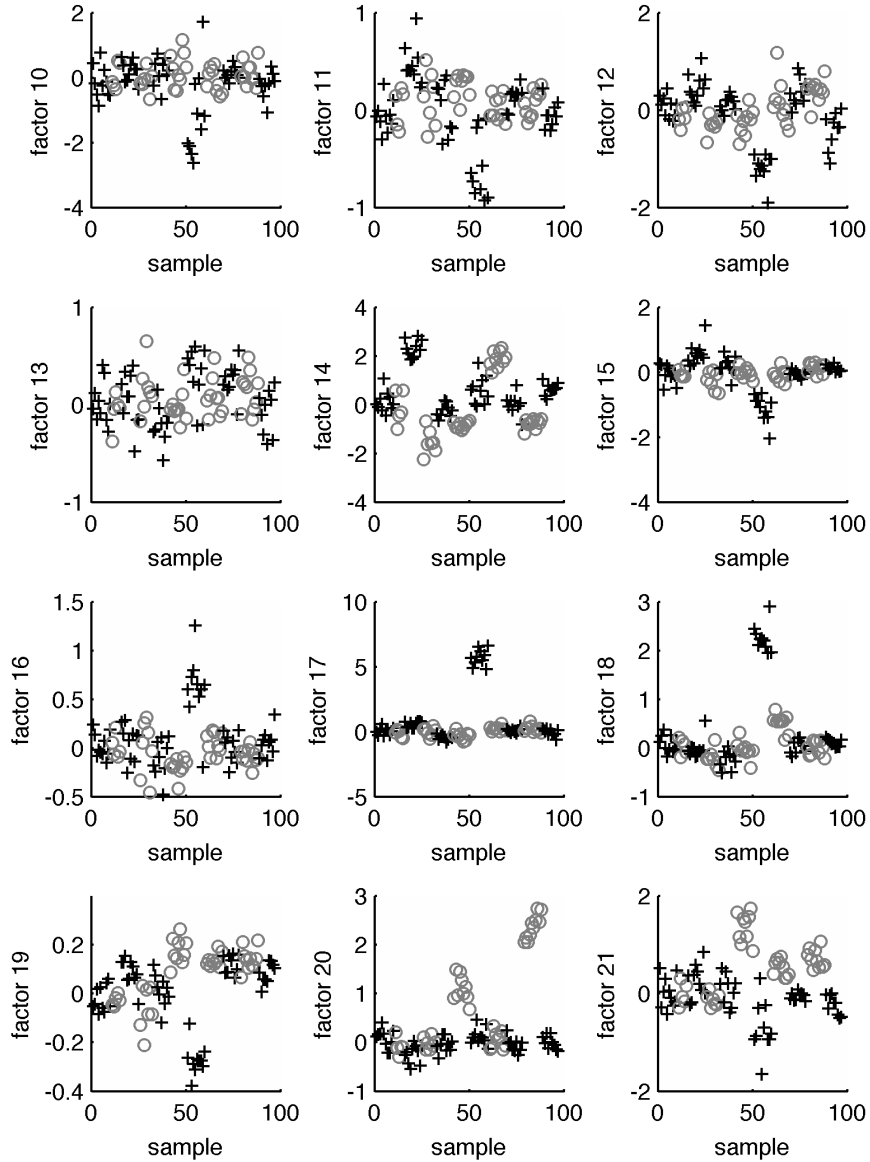


Figure 11: Imputed latent factor scores for oncogene intervention data, latent factors 10-21. Oncogene interventions are ordered as control 1, control 2, MYC, SRC, β -Catenin, E2F3, RAS, P63, AKT, E2F1, P110, with samples from the same experimental intervention appearing consecutively, designated by the same color/marker. Separate interventions are depicted by alternating black crosses and grey circles. Continued factor-based separation of the RAS group is evident. Factor 20 distinguishes the E2F3 and E2F1 interventions.

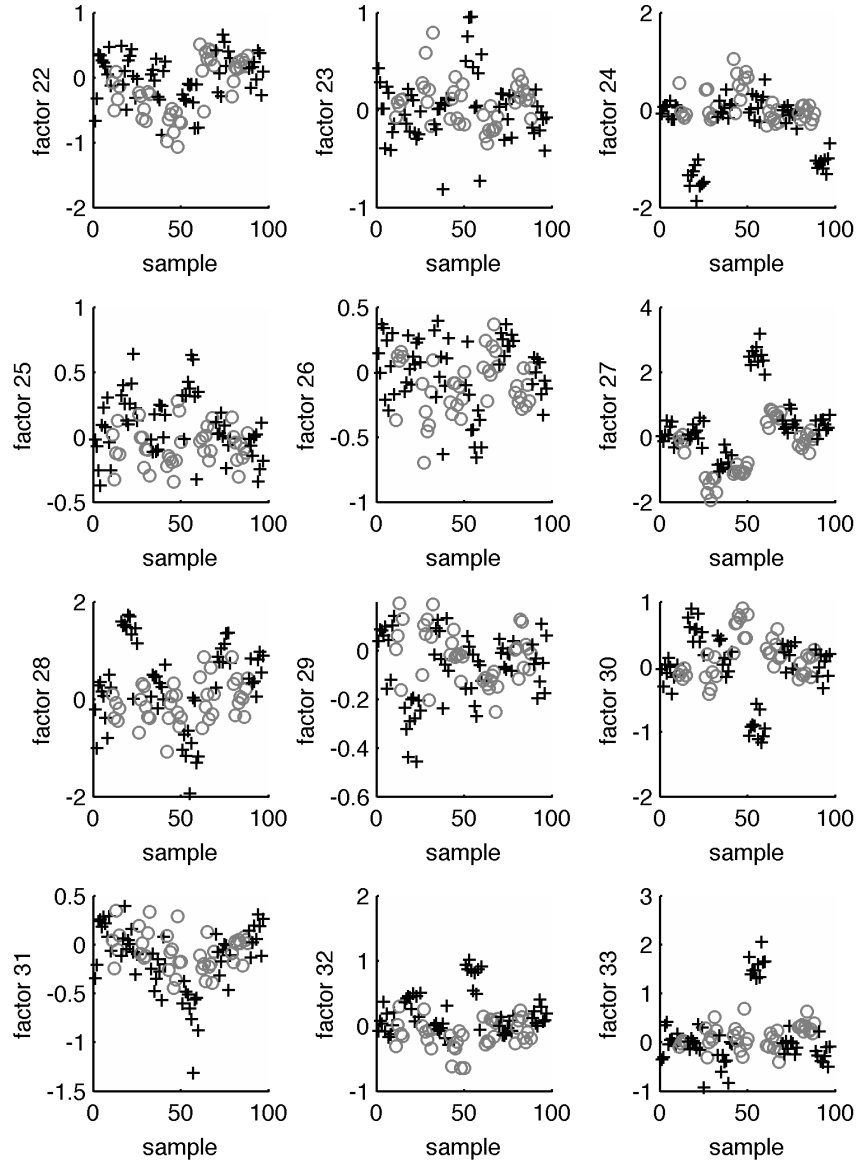


Figure 12: Imputed latent factor scores for oncogene intervention data, latent factors 22-33. Oncogene interventions are ordered as control 1, control 2, MYC, SRC, β -Catenin, E2F3, RAS, P63, AKT, E2F1, P110, with samples from the same experimental intervention appearing consecutively, designated by the same color/marker. Separate interventions are depicted by alternating black crosses and grey circles. In addition to further factor-based separation of the RAS group, factor 24 distinguishes the MYC and P110 interventions.

the breast cancer study using Weibull survival models. With t_i denoting the survival time of breast cancer patient i , the Weibull density function is $p(t_i|a, \gamma) = at_i^{a-1} \exp(\eta_i - t_i^a e^{\eta_i})$ where $\eta_i = \gamma' y_i$ is the linear predictor based on covariate vector y_i and regression parameter vector γ , and a the Weibull index. We have previously used these models, and the Bayesian model uncertainty analysis using shotgun stochastic search (SSS) (Hans et al. 2007; Hans et al. 2007), with some success in applications in cancer genomics (Rich et al. 2005; Dressman et al. 2006), and now apply the approach in the breast cancer context here. Based on an overall set of candidate predictors and specified prior distributions, SSS generates a search over the space of subset regression and delivers posterior probabilities on all models visited together with posterior summaries for each model, the latter including approximate inferences on (γ, a) in any one of the set of evaluated models (see Appendix 3.2 for more details and links to software).

We explored multiple Weibull survival models drawing on the 33 estimated factors (posterior means of factor scores) as covariates. The left panel of Figure 13 depicts the marginal Weibull-model inclusion probabilities for the 33 latent factors from the breast cancer survival analysis. Three factors, 5, 12, and 25, appear with probabilities greater than 20%, with factor 5 being of clear interest. The right panel in Figure 13 displays the pairwise inclusion probabilities among factors 5, 12, and 25, and further indicates that factor 5 alone may have significant use for predicting survival. Note that factor 19, though marginally associated with clinical risk factors, does not score at all highly – relative to these other 3 factors – in terms of its appearance in high probability survival models.

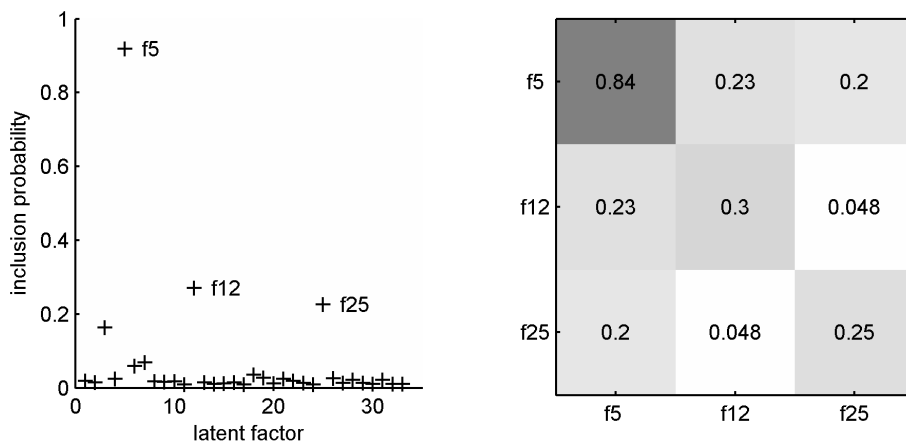


Figure 13: *Left frame*: Marginal posterior inclusion probabilities for the 33 factor covariates in the Weibull model for survival time, as determined by shotgun stochastic search. *Right frame*: Probabilities of pairs of factors appearing together in the Weibull models. Darker tiles indicate higher probabilities. Factor 5 is often accompanied by either factor 12 or factor 25, but very rarely both, indicating some predictive redundancy associated with the 12,25 pair.

The median survival time m for a patient with covariate vector y satisfies $m^a = \exp(-\gamma' y) \log(2)$ so that posterior inference on (a, γ) yields inference on m . From the top 1,000 models visited over the course of 10,000 model-search steps we compute approximate posterior means for (a, γ) in each model as well as approximate model probabilities, and average the plug-in estimates of m over models. The resulting estimated median survival time can then be evaluated at any specified covariate vector y . In particular, we evaluate this at the covariate vectors for each of the $n = 251$ patients in the breast data set, so generating predictions of median survival times for a future hypothetical cohort of 251 patients sharing these covariate values. Kaplan-Meier curves can provide

useful visual displays for simply displaying survival data on a number of patient subgroups, and Figure 14 shows separate KM curves for the two groups of patients whose estimated median survival lies above/below the median value across all 251 patients. This data display clearly indicates the potential clinical value of the biological sub-pathways represented by factors 5, 12, and 25, at least.

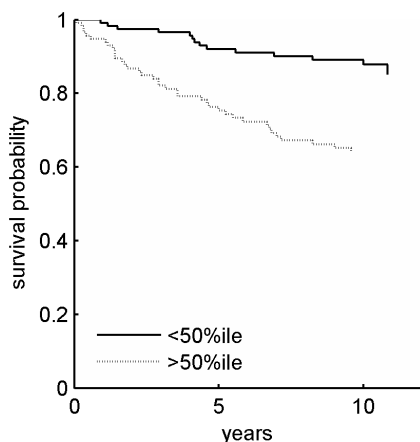


Figure 14: Kaplan-Meier curves representing predicted median patient survival probabilities when patients are divided into two cohorts based on Weibull model prediction of survival time. The solid black curve represents patients for whom the posterior predictive estimate of median survival time exceeds the median of that value across the 251 patients, and the dotted curve represents the other 50% higher-risk patients.

This study was repeated with an expanded covariate set that now adds to the 33 factors the following candidates: (a) the set of 9 projected oncogene pathway signature scores, and (b) the traditional clinical covariates used in breast cancer prognosis: Elston grade, ER, lymph node status, P53, PgR, tumour size and age at diagnosis. Figure 15 summarizes the resulting inclusion probabilities. Although the key clinical variables – lymph node status and tumour size – are identified as the top two predictors of survival time, factors 5, 12, and 25, are the third, fourth, and fifth most relevant. It is noteworthy that none of the original signature scores are among the top predictors, demonstrating that the refinement of the original signatures achieved by latent factor analysis hones the representation of biological pathway activation to deliver a refined and, clinically, more predictive representation of the complexity of pathway structure in the *in vivo* setting.

Kaplan-Meier curves for this expanded analysis (Figure 16) show greater separation between good and poor prognosis patient groups than was evident in the factor-only analysis, though this may be expected given the inclusion of clinico-pathological variables such as lymph node status that are highly indicative of poor outcome and tend to be poorly predicted by expression data.

More formally and more incisively, we can directly compute estimates of the posterior predictive survival function for any future patient represented by a candidate covariate vector – i.e., address the question of using the factor-based sub-pathway signatures for personalised prognosis – again by averaging over all models visited by the SSS survival model analysis. We do this using vectors y that set some covariates at their median values across the 251 samples and others at more extreme values, i.e., upper or lower 5% values with respect to the data. This enables an investigation of the predictive effect of individual covariates with the other fixed at average values. Some examples appear in Figures 17 and 18, obtained by varying the values of factors 5, 12, and 25, and then with respect to lymph node status and tumour size in Figure 18). In the factors-only analysis (Figure 17), predicted survival changes dramatically for extreme values of factor 5. In fact, the change in predicted survival associated with variation in factor 5 in the factors-only analysis is greater than the change in predicted survival associated with variation in lymph node status in the combined factors and clinical variables analysis.

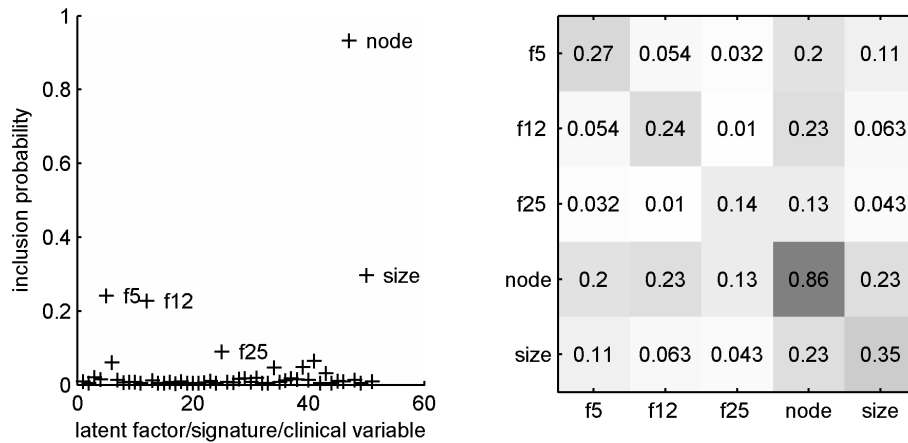


Figure 15: *Left frame:* Marginal posterior inclusion probabilities for candidate covariate in the Weibull survival analysis when drawing on factors, signatures and clinical covariates. *Right frame:* Pairwise inclusion probabilities for the top 5 covariates. Again factors 12 and 25 rarely appear in conjunction. Factor 25 appears with lymph node status with substantially lower probability than do the other factors, suggesting that factor 25 can play a role as a surrogate for node status.

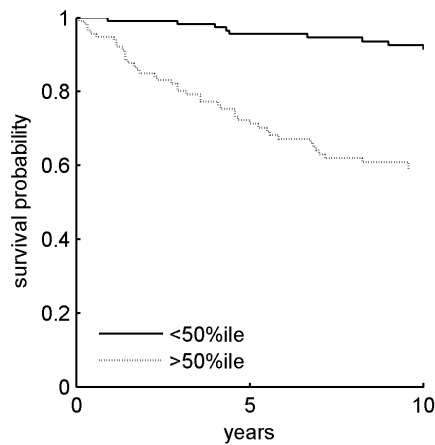


Figure 16: Kaplan-Meier curves representing median patient survival from survival analysis incorporating latent factors, original signatures, and clinical variables. The solid black curve represents patients with predicted median survival time above the median over patients and the dotted curve represents the other 50% higher-risk patients. Greater stratification is achieved relative to Figure 14, mainly due to the use of lymph node status as a covariate.

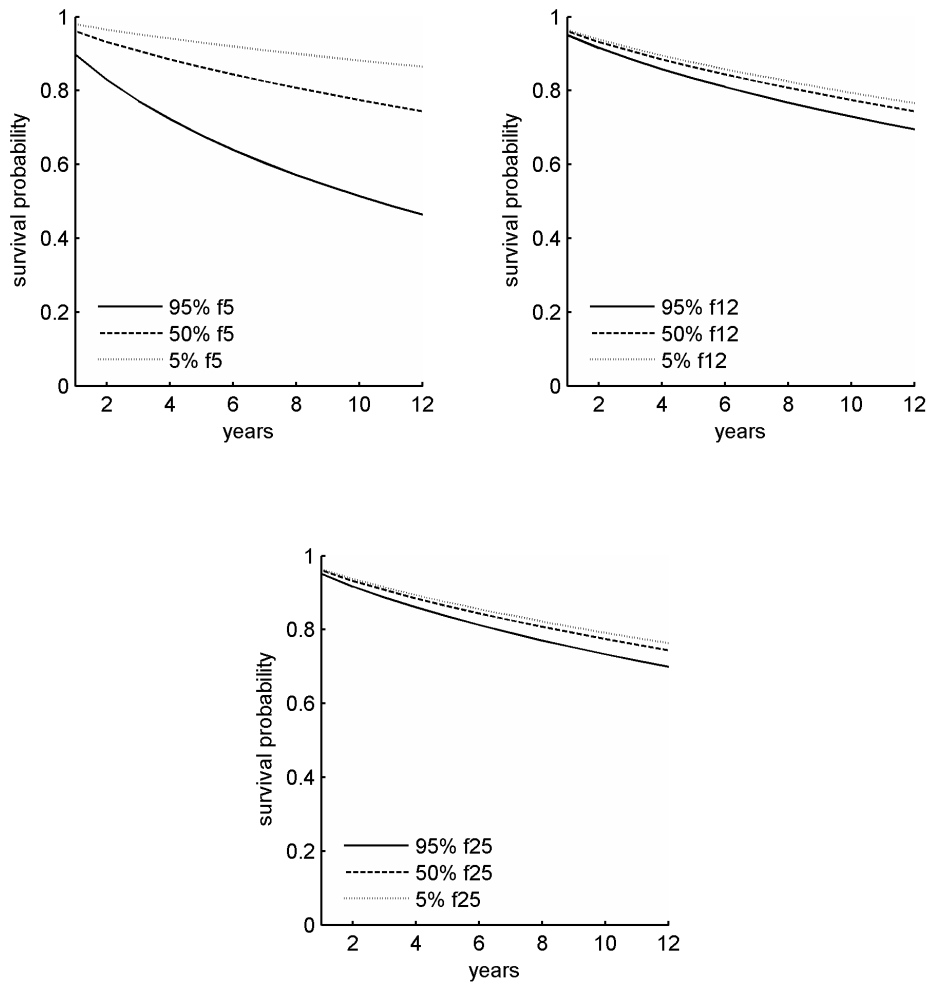


Figure 17: Estimates of posterior predicted Weibull survival curves from factor-only survival analysis. The dashed lines represent survival when all factor values are fixed at their posterior median value. The solid and dotted lines in each figure represent predictions achieved by individually varying values of factors 5, 12, and 25 to values chosen as the 5th and 95th sample quantiles in the $n = 251$ patient data set. Stratification on basis of factor 5 appears to be a useful predictive diagnostic.

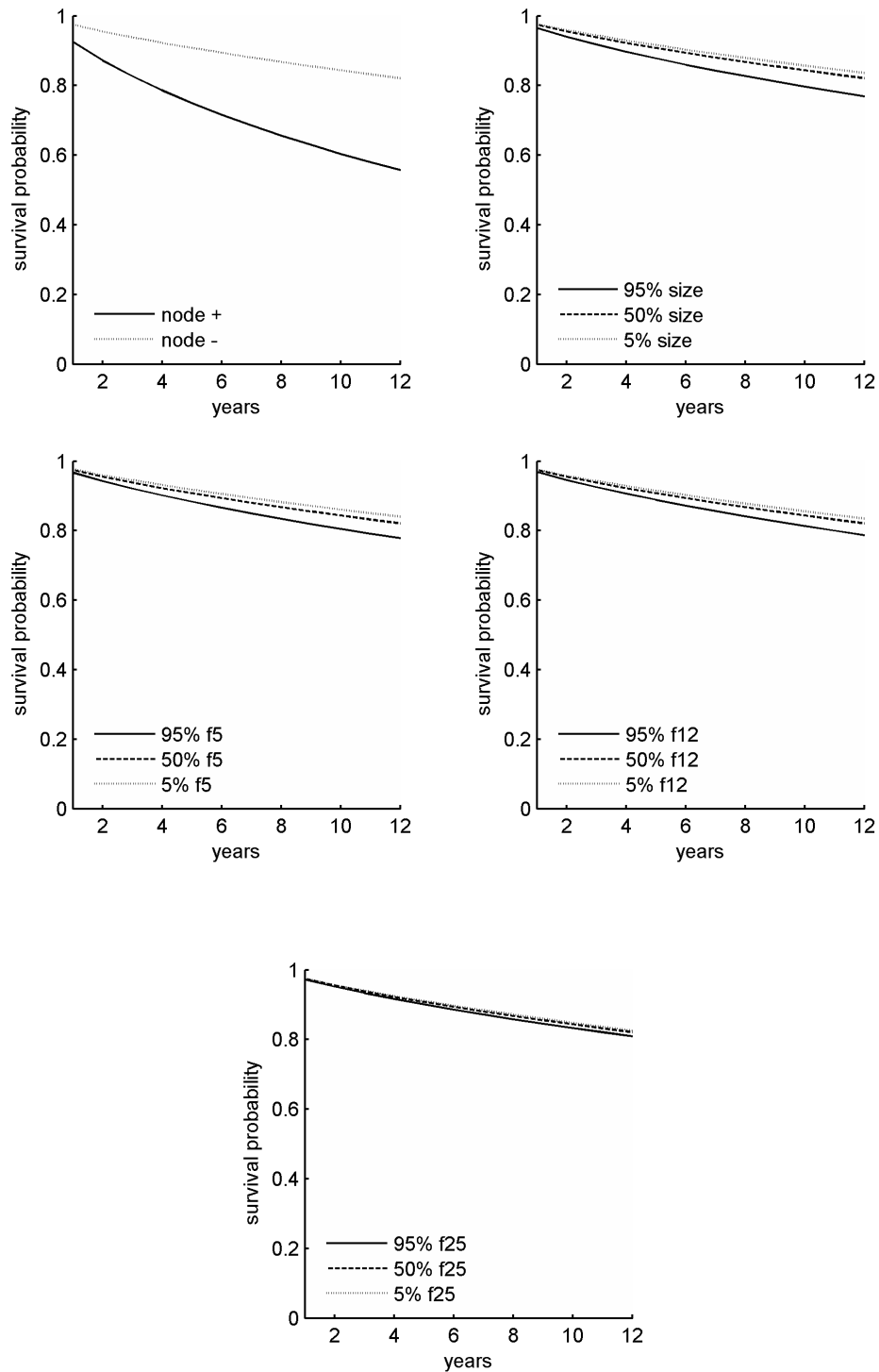


Figure 18: Estimates of posterior predicted Weibull survival curves from factor/signature/clinical variable survival analysis. Dashed lines represent median covariate predictions. Solid and dotted lines represent predictions achieved by individually varying values of lymph node status, tumour size, and factors 5, 12, and 25 to their 5th and 95th quantiles. The predictive utility of factor 5 is diminished in light of lymph node status, though the separation achieved by lymph node status is comparable to that achieved by factor 5 in the Figure 17.

3 Biological Evaluation and Pathway Annotation Analysis

3.1 Initial Discussion

In light of their obvious clinical significance, placing factors 5, 12 and 25 in their proper biological context is of primary importance. The discussion in Section 2.3 gave some clues as to the relationship between these factors and the oncogene pathways. Factor 5 appears to represent some intersection between the MYC, RAS, P63, and P110 pathways, and factor 12 some sub-component of the RAS pathway, but factor 25 does not appear to associate with any of the 9 *in vitro* defined oncogene pathways. We are also interested in the biological relevance of factor 19 that, although not appearing competitive in survival prediction, evidently associates with traditional clinical risk factors.

Gene-specific connections between factors and the original oncogene pathways can be explored by identifying genes with high probabilities of non-zero factor loadings and relating them to the original oncogene pathway signatures. The top 20 genes in factors 5, 12, and 25 were identified by absolute values of the estimated factor loadings. Figure 19 presents the probabilities of inclusion in the original oncogene signatures for each gene among the three sets of top genes. Predictably, many of the top genes in factor 5 were strongly associated with the P63 signature. This makes biological sense, given that factor 5 was founded by the P63 signature score. Top genes in factor 12 show association primarily with the RAS and P63 signatures; in particular the founder gene of factor 12 has high probability of involvement in both the RAS and P63 pathways. Very few of the top genes in factor 25 were among those included in the oncogene study, and of the 4 genes common to both studies, only two genes showed significant probability of the involvement in oncogene pathways: probe 210761_s_at in the RAS pathway, and probe 216836_s_at in the P63 and P110 pathways. Such connections further emphasize the involvement of RAS and P63 sub-pathway activity in affecting tumour malignancy and, therefore, long-term survival.

Critically, this already focuses attention on the ability of factor profiling to zoom-in on critical pathway intersections; it turns out that these two probes identify genes well-known to play central roles in breast cancer oncogenesis. Probe 210761_s_at is a sequence within the gene GRB7, a well-known correlate and co-expression partner of the epidermal growth factor receptor ERB-B2 that is a fundamentally causative oncogene in breast cancer and one of the two major drug targets in breast cancer chemotherapy; ERB-B2 is also known as HER-2- ν (Sørli et al. 2001; Bertucci et al. 2004; Badache and Gonçalves 2006). Remarkably, probe 216836_s_at is in fact a probe sequence for oncogene ERB-B2 itself. Factor analysis seeded by a set of individual oncogenic signatures has identified pathway interconnections that clearly lead to ERB-B2 as a determining player represented, at least in part, by the clinically relevant factor 25.

3.2 Bayesian Pathway Annotation Analysis

Further interpretation of factors can be developed using annotated biological databases that catalogue known cellular pathways and other versions of pathways. One standard is the Molecular Signatures database (MSigDB 2008) that contains thousands of annotated and curated lists of genes representing verified cellular pathways and also others based on gene expression studies. Such *pathway gene lists* are, of course, incomplete and typically error-prone, but nevertheless provide key representations of known pathways. We apply Probabilistic Pathway Annotation (PROPA) analysis to the results from our factor analysis in order to assess factors against the set of over 1000 annotated pathway gene lists from MSigDB (2008). PROPA is a fully Bayesian analysis that provides – among other things – formal assessment and rankings of pathways putatively linked to

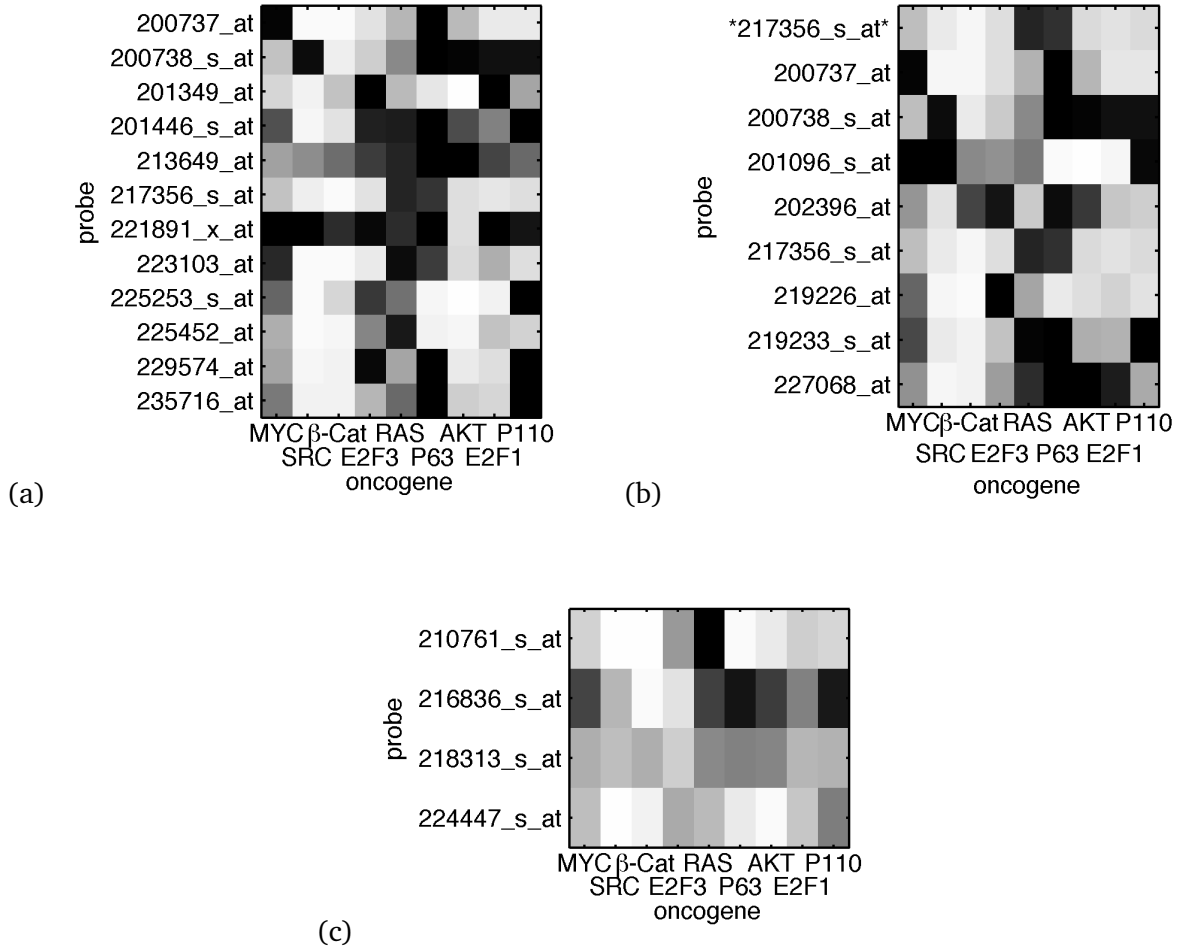


Figure 19: Heatmaps showing oncogene signature inclusion probabilities $P(\beta_{g,k}^{vitro} \neq 0 | X^{vitro})$ for subsets of the top 20 genes in factors 5 (a), 12 (b), and 25 (c) also present in the oncogene signature study. Probabilities are coded by grayscale, with black designating probability 1. The top genes in each factor were determined by sorted absolute value of mean posterior factor loading. 9 of the 12 top genes in factor 5 appear in the P63 pathway signature with high probability. Several top genes in factor 12 show association with the RAS and P63 signatures, including founder probe 217356_s.at defining gene PGK1. Factor 25 shares only 4 genes in common with the oncogene study, and of them only two played significant roles in oncogene pathways: probe 210761_s.at in the RAS pathway is a key oncogenic marker, gene GRB7 highly correlated in expression patterns generally with ERB-B2, and probe 216836_s.at in the P63 and P110 pathways is the fundamentally causative breast cancer oncogene ERB-B2 itself.

factors based on the posterior gene×factor probabilities $Pr(\beta_{g,k} \neq 0|X^{vivo})$ and signs of estimated $\beta_{g,k}$. Full details and examples appear in Shen (2007) and Shen and West (2008); additional comments are given here in Appendix 3.2.

Application of PROPA to the gene sets comprising factors 5, 12, 19 and 25, yielded the connections below to previously annotated cellular pathways of immediate relevance in cancer. Similar evaluations can be performed on other factors; such studies, as well as other follow-on biological studies, were initiated as a result of this statistical analysis.

Factor 5

PROPA identifies the serum fibroblast cell cycle¹ pathway gene list as representing the most likely candidate biological pathway underlying factor 5. This pathway gene list relates to the gene expression program of human cells in response to serum exposure and represents links between cancer progression and wound healing – “cancer invasion and metastasis have been likened to wound healing gone awry” (Chang et al. 2004). The “wound-healing” expression signature that this pathway gene list is based on is a well-known and powerful predictor of risk in a number of cancers and particularly in breast cancer, a biomarker of risk of increased increased risk of invasion, metastasis and death. This pathway very substantially dominates the other 1000+ in PROPA analysis of factor 5, and hence we identify factor 5 as reproducing this wound-healing signature and overlaying the corresponding cell cycle pathways that are induced by serum exposure. This is consonant with our findings that factor 5 alone is the most potent predictor of survival in our breast survival studies and that it derives in our analysis from genes linked to the MYC signature, and apparently represents intersections of the MYC, RAS, P63, and P110 pathways. That many of the top genes in factor 5 were strongly associated with the P63 signature raises the potential for further studies to advance our currently limited understanding of the connectivities between P63 and other components of the broader cell cycle and cell growth regulatory machinery (Figure 1).

Factor 12

PROPA identifies a number of hypoxia-related pathways, the most highly scoring by far being the HIF-1 target pathway², a set of genes regulated by Hypoxia-inducible factor 1 (HIF-1) and that define pathway activities that mediate adaptive responses of cells to reduced oxygen availability. The HIF-1 pathway responses are well-known to play major roles in the pathogenesis of cancer (Semenza 2001; Chi et al. 2006). Less highly scored by PROPA but of note is a well-known pathway gene list representing a breast cancer “poor prognosis” signature (van’t Veer et al. 2002) that consists primarily of genes regulating cell cycle, invasion, metastasis and angiogenesis. That our analysis identifies factor 12 as most likely overlaying hypoxia-response pathways while tying into cancer invasion mechanisms connects aspects of the tumour micro-environment (oxygenation) with gene pathway responses in cell regulation. Further biological investigation may be suggested to evaluate genes scoring highly in factor 12, especially in view of the fact that, while factor 12 arises initially as a sub-component of the RAS pathway and has many top genes that show association with the RAS signature, there are strong connections also with the P63 signature; in particular the founder gene of factor 12 has high probability of involvement in both the RAS and P63 pathways. To our knowledge, these connections of P63 with primary cancer mechanisms and pathways have not been previously explored.

¹http://www.broad.mit.edu/gsea/msigdb/cards/SERUM_FIBROBLAST_CELLCYCLE.html

²http://www.broad.mit.edu/gsea/msigdb/cards/HIF1_TARGETS.html

Factor 19

Factor 19 is a primary ER ((o)estrogen receptor) factor; the two top PROPA pathways are expression-based ER pathways involving genes whose expression levels are either consistently positively or negatively correlated with estrogen receptor status in breast cancer ³. This is concordant with our findings of the strong association of factor 19 with ER in the breast data set (Figure 5) and with the names of some of the genes that score highly on factor 19 in the BFRM analysis that can be recognised as ER related, ER targets or otherwise synergistic with ER. That an ER factor arises naturally as part of the evolutionary BFRM model refinement is no surprise given the dominance of the ER pathway in breast cancer – in terms of very many genes influenced by ER activity – and also as ER is intimately interconnected with the cell cycle and progression machinery through cyclin-D and other known pathway interactions (Figure 1 and the ER example and references in [Carvalho et al. \(2008\)](#)).

Factor 25

The top PROPA pathway linked to factor 25 is also ER, that gene list representing genes consistently positively correlated with estrogen receptor status also identified for factor 19. The second and third most highly scoring pathways also link to ER but are very different from those defined by factor 19 and also seem to indicate a rather different biological function for the pathway activities factor 25 represents. In particular, they are pathway gene lists derived from experiments to investigate the effects on ER positive cells of drugs including Tamoxifen, the currently most used hormonal therapy in breast cancer care, and importantly related to questions of cellular responses and resistance to Tamoxifen ⁴. In an earlier study we had identified such a factor as a “predictor of resistance to Tamoxifen” and had verified its ability to discriminate breast patients with respect to survival outcomes according to whether they were or were not resistant in the sense of high versus low values of the estimated factor ([Lucas et al. 2008](#)). That analysis identified this factor from a very different set of initial genes in the BFRM analysis, and that it emerges again here in an analysis initiated by the set of oncogene signatures indicates that this is clearly a key clinically relevant factor in breast cancer genomics, linking to critical underlying pathways. It is an important ER-related factor that is quite distinct from the dominant ER measures provided here by factor 19.

We have already noted that very few of the top genes in factor 25 were among those included in the oncogene study, but key among those that were ERB-B2. There is no annotated ERB-B2 pathway in the gene lists data bases, but this suggests a role for the ERB-B2 activity in mediating ER-related activity and perhaps the cellular responses to drugs including Tamoxifen. Further investigation of this together with genes most highly scoring on factor 25 may well prove informative.

³http://www.broad.mit.edu/gsea/msigdb/cards/BRCA_ER_NEG.html and http://www.broad.mit.edu/gsea/msigdb/cards/BRCA_ER_POS.html

⁴http://www.broad.mit.edu/gsea/msigdb/cards/FRASOR_ER_UP.html and http://www.broad.mit.edu/gsea/msigdb/cards/BECKER_TAMOXIFEN_RESISTANT_DN.html

Appendix

Models and Computations: Bayesian Latent Factor Regression Models

MCMC in BFRM

The details of MCMC analysis of Bayesian latent factor regression models and the implementation in the BFRM software have been well documented (West 2003; Lucas et al. 2006; Wang et al. 2007; Carvalho et al. 2008) and illustrated in a number of related applications (Chang et al. 2007; Chen et al. 2007; Seo et al. 2007; Lucas et al. 2008; Lucas et al. 2008; Shen and West 2008). The MCMC uses a Gibbs sampling format as fully detailed in Carvalho et al. (2008) and to which readers interested in implementation details should refer; the BFRM software used is fully documented with examples and freely available⁵. In summary, the component conditional distributions sequenced through are as follows, in each case conditional on the data and all other model quantities denoted by the “ $-$ ” in conditionings.

- Resample latent factors from conditional posteriors induced by the Dirichlet process prior. For each sample i , $p(\Lambda_{:,i}|-)$ is sampled based on the two-step configuration sampler of Dirichlet process mixture models (West et al. 1994; MacEachern and Müller 1998). After each sweep through the n factor vectors, there will typically be a reduction to some smaller number of unique realised factor vectors $\Lambda_{:,i}$ induced by the inherent clustering mechanisms of the Dirichlet process model and reflecting the non-normality of the underlying latent factor distribution.
- Resample independently from the set of inverse gamma complete conditionals $p(\tau_j|-)$ for the variances of regression parameters and factor loadings.
- Resample independently from the set of inverse gamma complete conditionals $p(\psi_g|-)$ for residual variances.
- For each j in turn, resample the $\alpha_{g,j}$ parameters (denoting both regression parameters and factor loadings with no loss of generality) from complete conditional posteriors under the structured sparsity priors. This involves a custom simulation step that samples, independently for $g = 1 : p$, the set of p distributions $p(\alpha_{g,j}, \pi_{g,j}|-)$ via composition. A detail is that with s latent factors the first s loading parameters are constrained to be non-negative, and this is incorporated in the algorithm. See Lucas et al. (2006) and the appendix in Carvalho et al. (2008).
- Resample the $\rho_{g,j}$ hyperparameters of the sparsity priors independently from implied beta complete conditionals.

Evolutionary Factor Models Search

A key element of the strategy for biological pathway exploration as developed in this case study is the use of evolutionary stochastic model search to expand the model gene set and number of factors to refine an initial focus on a specified set of pathway-specific genes. This involves evolutionary search, fully detailed in Carvalho et al. (2008), summarised as follows:

- In any “current” model based on p genes and k factors, compute approximate variable inclusion probabilities $Pr(\beta_{g,k} \neq 0|X)$ for all genes g not in the current model.

⁵<http://www.stat.duke.edu/research/software/west/bfrm/>

- Rank and select genes with highest inclusion probabilities subject to exceeding a specified probability threshold. Stop if no additional genes are so significant.
- Refit the expanded model, also increasing to $k + 1$ factors with an initial, random choice of the founder of the new factor (the gene listed at $g = k + 1$). From this analysis identify the gene with highest estimated $Pr(\beta_{g,k+1} \neq 0|X)$ and refit the model with that gene now at $g = k + 1$.
- Cut back to fewer than $k + 1$ factors by dropping any factor j such that fewer than some small pre-specified number of genes have $Pr(\beta_{g,j} \neq 0|X)$ exceeding a high threshold. Otherwise, accept the expanded model and continue to iterate the model evolutionary search.
- Stop if the above process does not include additional genes or factors, or if the numbers exceed some pre-specified targets on the number of genes included in the model and/or the number of factors.

In our case study here, all MCMC analyses used summaries from runs of length 10,000 following 1,000 discarded as burn-in. Each iterate of the evolutionary model search brought in at most 10 new genes and required a threshold of at least 0.95 on gene \times factor inclusion probabilities to add such genes to the current model. In considering expanding to add a further factor, we required that at least 5 genes have gene \times factor inclusion probabilities of at least 0.85. Model search was allowed to proceed until as many as 500 genes and 50 factors have been added to the model, with a control restricting maximum number of genes per factor to at most 30. This means that once the number of genes with nonzero loadings on a given factor reached 30, then no further genes were incorporated into the latent factor model on the basis of that factor. Note that the factor analysis of the breast cancer data terminated on reaching 500 genes, at which point the analysis had evolved from the initial 9 factors to a final total of 33.

Models and Computations: Bayesian Weibull Survival Regression Models

Model search and analysis in the Weibull regression models used the SSS software (Hans et al. 2007; Hans et al. 2007) that is fully documented with examples and freely available⁶. Under normal priors on the regression parameter vector γ and a gamma prior on the Weibull index a in any specified regression model, this analysis computes posterior modes for (a, γ) and estimates the corresponding model marginal likelihood via Laplace approximation. With a specified prior probability of covariate inclusion – treating covariates independently and each with the same prior inclusion probability – this allows for the computation of approximate posterior probabilities across any specified set of models. Shotgun stochastic search for such regression models (Hans et al. 2007) moves around the large space of subsets of candidate covariates visiting many possible regression models. The algorithm is designed to seek out models in regions of high posterior probability. SSS is very efficient and, in terms of large numbers of models searched, very aggressive, with the ability to rapidly explore and catalogue many, many models. SSS is also parallelisable and the code has both serial and parallel versions.

In the survival analysis of the case study here, SSS-based exploration of the space of Weibull survival models was allowed to proceed for 10,000 model-search iterations, and recorded the top 1,000 models in terms of approximate posterior probability. The resulting approximate posterior predictive distribution for a new patient sample – the distribution required for the summary comparisons reported – is then constructed as the mixture, weighted by approximate posterior model

⁶<http://www.stat.duke.edu/research/software/west/sss/>

probabilities, of the 1,000 Weibull models with parameters (a, γ) set at the computed posterior modes. The prior probability of covariate inclusion, that defines an implicit penalty on model dimension, was set so that the prior expected number of covariates in the model is 3. Model search was initiated by evaluating a randomly selected model with 2 covariates.

Models and Computation: Bayesian Probabilistic Pathway Annotation

A biological pathway is represented by an unordered list of, typically, a small number of genes (10s to several 100s) in typical pathway data bases. In essentials, PROPA analysis of the BFRM outputs for any factor k scores each pathway gene list according to the relative enrichment of genes g with high values of $Pr(\beta_{g,k} \neq 0 | X^{vivo})$, and also for the relative paucity of genes g with low values, in the list. Some pathway gene lists also carry information about the sign of changes in expression of genes in the list as a result of intervention on the pathway, so that the estimated signs of the $\beta_{g,k}$ for genes in the list are also then relevant in assessing the results of PROPA.

For any one of the factors, say factor k , in the BFRM analysis of the breast data, the PROPA framework idealises an underlying, true pathway \mathcal{F} related to the factor; \mathcal{F} , the *factor pathway*, is simply an unknown list of genes that are changed in expression with this factor, i.e., a set of genes with non-zero loadings on the factor. The estimated values $Pr(\beta_{g,k} \neq 0 | X^{vivo})$ over all (40,000+) probes provide data for PROPA analysis to make inference on the true gene \times factor associations. Write $\Pi = \{Pr(\beta_{g,k} \neq 0 | X^{vivo}), g = 1 : p\}$ and $A_{1:m} = (A_1, \dots, A_m)$ for the set of annotated gene lists in the biological data base. Each A_j is a list of genes that are known to be related to an underlying biological pathway \mathcal{A}_j ; note that A_j is typically incomplete, as there will usually be additional genes linked to \mathcal{A}_j that are not yet listed in A_j ; similarly, A_j may contain genes that are false positives, and may in future be removed from the list as additional biological experiments arise. PROPA addresses these issues within the analysis. PROPA represents the problem of matching the pathway \mathcal{F} underlying the factor with the annotated pathways \mathcal{A}_j via posterior probabilities $Pr(\mathcal{F} = \mathcal{A}_j | \Pi, A_{1:m}) \propto Pr(\mathcal{F} = \mathcal{A}_j | A_{1:m}) p(\Pi | A_{1:m}, \mathcal{F} = \mathcal{A}_j)$ and focuses on the likelihood terms $p(\Pi | A_{1:m}, \mathcal{F} = \mathcal{A}_j)$ as j moves across all the pathways; these are the overall measures that underlies pathway assessment, and can be applied whatever the chosen values of the priors $Pr(\mathcal{F} = \mathcal{A}_j | A_{1:m})$. Full details of the PROPA analysis, within which the terms $p(\Pi | A_{1:m}, \mathcal{F} = \mathcal{A}_j)$ are marginal likelihoods from a class of statistical models, are developed in [Shen and West \(2008\)](#) and rely on computations using both MCMC and variational methods, the latter for evaluation of marginal likelihoods based on the outputs of within-model MCMC. The resulting marginal likelihood values are the *pathway scores* that PROPA delivers and that are used to rank the $j = 1 : m \approx 1,000$ pathways on each factor.

References

- Badache, A. and A. Gonçalves (2006). The ERB-B2 signaling network as a target for breast cancer therapy. *Journal of Mammary Gland Biology and Neoplasia* 11, 13–25.
- Bertucci, F., N. Borie, C. Ginestier, A. Groulet, E. Charafe-Jauffret, J. Adélaïde, J. Geneix, L. Bachelart, P. Finetti, A. Koki, F. Hermitte, J. Hassoun, S. Debono, P. Viens, V. Fert, J. Jacquemier, and D. Birnbaum (2004). Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene* 23(14), 2564–2575.
- Bild, A. H., G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. Nevins (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357.
- Carvalho, C., J. Lucas, Q. Wang, J. Chang, J. Nevins, and M. West (2008). High-dimensional sparse factor modelling - Applications in gene expression genomics. *Journal of American Statistical Association (to appear)*.
- Chang, H., J. Sneddon, A. Alizadeh, R. Sood, R. West, K. Montgomery, J.-T. Chi, M. van de Rijn, D. Botstein, and P. Brown (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biology* 2, E7 2004 Jan, <http://dx.doi.org/10.1371/journal.pbio.0020007>.
- Chang, J., C. Carvalho, S. Mori, A. Bild, Q. Wang, M. West, and J. Nevins (2007). A genomic strategy to elucidate fundamental units of cellular networks. *Discussion Paper #07-09, Department of Statistical Science, Duke University*; <http://ftp.stat.duke.edu/WorkingPapers/07-09.html>.
- Chen, J. L.-Y., J. Lucas, T. Schroeder, S. Mori, J. Wu, J. Nevins, M. Dewhirst, M. West, and J.-T. Chi (2007). The genomic analysis of lactic acidosis response in human cancers. *PLoS Genetics (submitted for publication)*.
- Chi, J.-T., Z. Wang, D. Nuyten, E. Rodriguez, M. Schaner, A. Salim, Y. Wang, G. Kristensen, A. Helland, A. Borresen-Dale, A. Giaccia, M. Longaker, T. Hastie, G. Yang, M. van de Vijver, and P. Brown (2006). Gene expression programs in response to hypoxia: Cell type specificity and prognostic significance in human cancers. *PLoS Medicine* 3, E47 2006 Mar.
- Dressman, H. K., C. Hans, A. Bild, J. Olsen, E. Rosen, P. K. Marcom, V. Liotcheva, E. Jones, Z. Vujaskovic, J. R. Marks, M. W. Dewhirst, M. West, J. R. Nevins, and K. Blackwell (2006). Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant therapy. *Clinical Cancer Research* 12, 819–216.
- Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search in regression with many predictors. *Journal of the American Statistical Association* 102, 507–516.
- Hans, C., Q. Wang, A. Dobra, and M. West (2007). SSS: High-dimensional Bayesian regression model search. *Bulletin of the International Society for Bayesian Analysis* 14, 8–9, <http://www.stat.duke.edu/research/software/west/sss/>.
- Huang, E., S. Chen, H. K. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang (2003). Gene expression predictors of breast cancer outcomes. *The Lancet* 361, 1590–1596.
- Huang, E., S. Ishida, J. Pittman, H. Dressman, A. Bild, M. D’Amico, R. Pestell, M. West, and J. Nevins (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genetics* 34, 226–230.

- Huang, E., M. West, and J. R. Nevins (2002). Gene expression profiles and predicting clinical characteristics of breast cancer. *Hormone Research* 58, 55–73.
- Lucas, J., C. Carvalho, D. Merl, and M. West (2008). In-vitro to In-vivo factor profiling in expression genomics. In D. Dey, S. Ghosh, and B. Mallick (Eds.), *Bayesian Modeling in Bioinformatics*. Taylor Francis.
- Lucas, J., C. Carvalho, Q. Wang, A. Bild, J. R. Nevins, and M. West (2006). Sparse statistical modelling in gene expression genomics. In P. Müller, K. Do, and M. Vannucci (Eds.), *Bayesian Inference for Gene Expression and Proteomics*, pp. 155–176. Cambridge University Press.
- Lucas, J. E., C. M. Carvalho, L. Chen, J.-T. Chi, and M. West (2008). Bench-to-bedside and cross-study projections of genomic biomarkers: An evaluation in breast cancer genomics. *Discussion Paper #07-24, Department of Statistical Science, Duke University*; <http://ftp.stat.duke.edu/WorkingPapers/07-24.html> (Submitted for publication).
- MacEachern, S. N. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* 7(2), 223–238.
- Miller, L. D., J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences* 102, 13550–13555.
- MSigDB (2008). *Molecular Signatures Data Base*. Broad Institute, <http://www.broad.mit.edu/gsea/msigdb/>.
- Nevins, J. R., E. S. Huang, H. Dressman, J. Pittman, A. T. Huang, and M. West (2003). Towards integrated clinico-genomic models for personalized medicine: Combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics* 12, 153–157.
- Pittman, J., E. Huang, H. Dressman, C. F. Horng, S. H. Cheng, M. H. Tsou, C. M. Chen, A. Bild, E. S. Iversen, A. T. Huang, J. R. Nevins, and M. West (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences* 101, 8431–8436.
- Rich, J., B. Jones, C. Hans, E. Iversen, R. McClendon, A. Rasheed, D. Bigner, A. Dobra, H. Dressman, J. Nevins, and M. West (2005). Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Research* 65, 4051–4058.
- Semenza, G. (2001). Hypoxia-inducible factor 1: Oxygen homeostasis and disease pathophysiology. *Trends in Molecular Medicine* 7, 345–50.
- Seo, D. M., P. J. Goldschmidt-Clermont, and M. West (2007). Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Annals of Applied Statistics* 1, 152–178.
- Shen, H. (2007). *Bayesian Analysis in Cancer Pathway Studies and Probabilistic Pathway Annotation*. PhD Thesis, Duke University, <http://www.stat.duke.edu/people/theses/ShenH.html>.
- Shen, H. and M. West (2008). Bayesian modeling for biological pathway annotation of genomic signatures. *Discussion Paper #08-13, Department of Statistical Science, Duke University*; <http://ftp.stat.duke.edu/WorkingPapers/08-13.html> (Submitted for publication).
- Sørli, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein,

- P. Lønning, and A. Børresen-Dale (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 98, 10869–10874.
- van't Veer, L., H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.
- Wang, Q., C. Carvalho, J. Lucas, and M. West (2007). BFRM: Bayesian factor regression modelling. *Bulletin of the International Society for Bayesian Analysis* 14, 4–5, <http://www.stat.duke.edu/research/software/west/bfrm/>.
- West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 7*, pp. 723–732. Oxford University Press.
- West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. R. Marks, and J. R. Nevins (2001). Predicting the clinical status of human breast cancer utilizing gene expression profiles. *Proceedings of the National Academy of Sciences* 98, 11462–11467.
- West, M., A. T. Huang, G. Ginsberg, and J. R. Nevins (2006). Embracing the complexity of genomic data for personalized medicine. *Genome Research* 16, 559–566.
- West, M., P. Müller, and M. D. Escobar (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In A. Smith and P. Freeman (Eds.), *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pp. 363–386. London: Wiley.