Bioinformatics Advance Access published July 30, 2011

Penalized regression elucidates aberration hotspots mediating subtype-specific transcriptional responses in breast cancer

Yinyin Yuan [1,2*], Oscar M. Rueda [1,2], Christina Curtis [1,2,3], Florian Markowetz [1,2*]

[1] Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, [2] Department of Oncology, University of Cambridge, Cambridge CB2 0XZ, UK, [3] Current Affiliation: Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: Copy number alterations (CNAs) associated with cancer are known to contribute to genomic instability and gene deregulation. Integrating CNAs with gene expression helps to elucidate the mechanisms by which CNAs act and to identify the transcriptional downstream targets of CNAs. Such analyses can help to sort functional driver events from the many accompanying passenger alterations. However, the way CNAs affect gene expression can vary in different cellular contexts, for example between different subtypes of the same cancer. Thus it is important to develop computational approaches capable of inferring differential connectivity of regulatory networks in different cellular contexts.

Results: We propose a statistical deregulation model that integrates copy-number and expression data of different disease subtypes to jointly model common and differential regulatory relationships. Our model not only identifies copy-number alterations driving gene expression changes, but at the same time also predicts differences in regulation that distinguish one cancer subtype from the other. We implement our model in a penalized regression framework and demonstrate in a simulation study the feasibility and accuracy of our approach. Subsequently, we show that this model can identify both known and novel aspects of cross-talk between the ER and NOTCH pathways in ER-negative-specific deregulations, when compared with ER-positive breast cancer. This flexible model can be applied on other modalities such as methylation or microRNA and expression to disentangle cancer signaling pathways

Availability: The Bioconductor-compliant R package DANCE is available from www.markowetzlab.org/software/

Contact: yinyin.yuan@cancer.org.uk, florian.markowetz@cancer.org.uk

1 INTRODUCTION

Somatic copy number alterations (CNAs) are known to be associated with cancer (Pollack *et al.*, 1999, 2002). They are particularly important for tumourigenesis, contributing to genomic instability and gene deregulation. Array comparative genomic hybridization (aCGH) has been used extensively to assess genome-wide copy number states in cancer, and statistical methods can be used to identify recurrent alterations in a particular disease state or subtype (Pollack *et al.*, 1999, 2002; Chin *et al.*, 2006, 2007). However, while some CNAs are driver mutations that are functionally important and impact tumour progression, they are often accompanied by numerous passenger events that confer no selective growth advantage (Pollack *et al.*, 2002; Greenman *et al.*, 2007). One of the challenges in cancer genomics is to identify functional driver events amidst many passenger alterations.

To better understand the mechanisms by which CNAs influence disease progression, it is necessary to integrate copy-number data with an intermediate phenotype like gene expression (Chen *et al.*, 2008). Integrating CNA data and RNA expression data can discover the primary aberrations that lead to downstream changes (Chin *et al.*, 2006). However, this is a challenging task for several reasons: First, CNAs can influence the expression of the proximal genes within a several Mb window (*cis*-acting), but can also exert effects elsewhere throughout the genome (*trans*-acting). Second, CNAs can span several Mb and thus make it difficult to distinguish between driver genes and passenger genes in this region. Thus, there is a need for new efficient computational approaches to refine the location of drivers.

While describing the regulatory role of CNAs for gene expression is already a difficult task, it is only the first step needed in understanding how tumours differ from normal tissue and how regulatory relationships vary between different disease subtypes. Cancer is triggered by collaborating factors and deregulated genes acting through signaling pathways are relevant for tumour growth and survival (Vogelstein and Kinzler, 2004). Several known examples of pathway deregulation result in aberrant signaling, the inhibition of apoptosis, and increased cell proliferation (Adjei and Hidalgo, 2005). Components in the deregulation network of signaling pathways represent attractive targets for therapeutic purpose (Watters and Roberts, 2006). A better understanding of the impact of CNAs on pathway activity and its variation in different cancer subtypes would be a major step forward in molecular medicine.

Many experimental and computational approaches have been proposed to identify the deregulation of cellular components, such as tumour suppressor silencing or oncogene activation, that contribute to tumour development, based on gene expression (Alizadeh *et al.*, 2000; Golub *et al.*, 1999; West *et al.*, 2001; Huang *et al.*, 2003; Rhodes *et al.*, 2004; Segal *et al.*, 2004; Hummel *et al.*, 2006; Bild *et al.*, 2006; Furge *et al.*, 2007; Slavov and Dawson, 2009). All of these works are important steps in this field, however, many different regulatory events are reflected in gene expression, including the activity of transcription factors, small non-coding RNAs, as well as gene dosage. From microarray expression data alone it is very hard to decide which of these resulted in the observed expression change.

Investigating changes in regulatory relationships between copy-number alterations and gene expression in different cancer subtypes has seldom been explored. While there is an increasing number of methods that integrate DNA copy number data and RNA expression data (Chin *et al.*, 2006, 2007; The Cancer Genome Atlas Research Network, 2008; Akavia *et al.*, 2010), most methods that compare cancer subtypes are geared towards supervised classification of tumour samples (Daemen *et al.*, 2009; Horlings *et al.*, 2010). Only few integration methods address the topic of differential regulation: One example is a bivariate approach by Schäfer *et al.* (2009) to search for abnormalities jointly at the DNA/RNA level. The abnormalities represent strong deviations from the reference, e.g. the amplification of a genomic region and over-expression of the proximal gene. Another example (that we compare against on real data) is DRI (Salari *et al.*, 2009), which can discover joint aberrations between two sample classes in paired copy-number and expression data. Both methods are computationally efficient, but are restricted to paired expression and CNA profiles on individual genes. In other words, these two methods can only infer *cis*-effects.

Here, we present **DANCE** (Deregulation Analysis in Networks of Copy-number driven Expressions) a systematic approach to decipher both common and differential regulatory mechanisms between disease subtypes. We propose a statistical deregulation model that integrates copy-number and expression data to jointly model common and differential regulatory relationships. Our model not only identifies CNAs driving gene expression changes, but at the same time also predicts differences in regulation that distinguish one cancer subtype from the other. Using one subtype as a reference we summarize the differential regulatory relationships of the other subtype in a *deregulation network*. To our knowledge, it is the first model to integrate both copy number and expression data to discover changes in the regulatory architecture between cancer subtypes.

The next section introduces our model, which, in a simulation study over a range of different experimental settings, outperforms alternative models and competing methods. Following, we use DANCE to study the deregulation between Estrogen Receptor (ER) positive and negative breast cancer (Chin *et al.*, 2006), which represent relatively good and poor outcome groups, respectively. While DANCE is able to tackle high-dimensional problems, to add interpretability we focus our analysis on signaling pathways. In particular, we are interested in the cross-talk between NOTCH and ER pathways, which may guide decisions about patients likely to benefit from Notch inhibitors. The results show that DANCE uncovers and extends known aspects of differential ER-Notch cross-talk in ER-positive versus negative disease. Thus, it can be a useful approach to infer critical aspects of pathway deregulation amongst cancer subtypes.

2 A JOINT MODEL FOR DEREGULATION

Let us assume that we have copy number data for p independent probes or regions $X = \{x_1, x_2, ..., x_p\}$ and mRNA expression data for q probes or genes $Y = \{y_1, y_2, ..., y_q\}$, where each vector x_i, y_j is a collection of n observations from n different individuals or samples. Let us also suppose that the n individuals can be divided into two groups, n_1 samples from sample class 1 and n_2 samples from sample class 2. We denote as $X_1(n_1 \times p$ matrix) and $Y_1(n_1 \times q)$ the observations from class 1, the deregulated sample class and as $X_2(n_2 \times p)$ and $Y_2(n_2 \times q)$ the observations from sample class 2; that is the reference sample class.

Our goal is to integrate the two data types, X_1, X_2 and Y_1, Y_2 into a model where the predictors are the copy number data (X_1, X_2) and the responses are the expression data (Y_1, Y_2). We assume that the expression of a particular gene can be affected by the copy number of an arbitrary set of copy number regions, but the impact of this influence can differ between the two groups. We can summarize the relationship with the following equations:

$$\begin{aligned} Y_1 &= X_1 B^r + X_1 B^d + \epsilon_1 \\ Y_2 &= X_2 B^r + \epsilon_2 \end{aligned} \tag{1}$$

where $B^r(p \times q)$ and $B^d(p \times q)$ denotes the reference and differential structure in the data respectively, and ϵ_1 and ϵ_2 are Gaussian noise with distribution of $N(0, I)$. We assume, without loss of generality, that both copy number and expression have been previously centered around zero.

Parallel estimation model We can estimate the two networks independently and obtain the deregulation $\hat{B}^d$ subtracting one from the other. That is,

$$\begin{aligned} Y_1 &= X_1 B_1 + \epsilon_1, \\ Y_2 &= X_2 B_2 + \epsilon_2, \\ \hat{B}^d &= \hat{B}_1 - \hat{B}_2 \end{aligned} \tag{2}$$

A similar model using only expression data has been used by Zhang *et al.* (2008) to study changes in transcriptional networks between two experimental conditions.

Sequential estimation model Alternatively, we can first infer the reference network and secondly use this estimate to infer the deregulation matrix B^d based on equation 2:

$$\begin{aligned} Y_2 &= X_2 B_2 + \epsilon_2, \\ Y_1 &= X_1 \hat{B}_2 + X_1 B_1 + \epsilon_1, \\ \hat{B}^d &= \hat{B}_1. \end{aligned} \tag{3}$$

Joint estimation model Eq.1 can be combined into one equation:

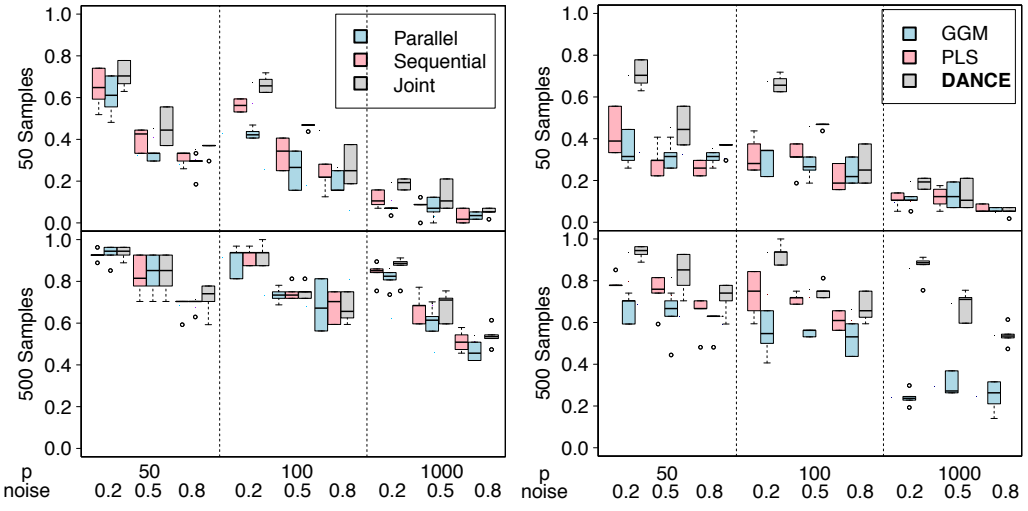
$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_1 \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} B^r \\ B^d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \tag{4}$$

which can be written as

$$\tilde{Y} = \tilde{X}\tilde{B} + \tilde{\epsilon}, \tag{5}$$

if we denote $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ as $\tilde{Y}$, $\begin{bmatrix} X_1 & X_1 \\ X_2 & 0 \end{bmatrix}$ as $\tilde{X}$, $\begin{bmatrix} B^r \\ B^d \end{bmatrix}$ as $\tilde{B}$ and $\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$ as $\tilde{\epsilon} \sim N(0, \Sigma_q)$. That is, the variance for each gene can be different. This estimation scheme improves over the parallel and the sequential methods, because it uses all the data for the estimation of all of the parameters and does not propagate errors in subsequent steps, as the sequential method does. The simulations in section (3.1) also reveal this advantage.

The matrix of coefficients $\tilde{B}$ has a simple biological interpretation: each term $\tilde{b_{ij}}$ indicates whether a predictor $\tilde{x_i}$ in $\tilde{X}$ is interacting with a response variable $\tilde{y_j}$ in $\tilde{Y}$, that is, if the copy number of a particular probe or region has an influence on the expression of a particular gene. The sign of this coefficient indicates if this effect is positive (gains produce up-regulation) or negative (gains produce down-regulation).



(a) Precision (b) Precision

Fig. 1. Simulation results for learning the non-zero coefficients/edges in the deregulation network over 100 runs. For simulated data sets of different sample sizes, number of copy number/expression probes (p) and noise levels, all methods were given prior knowledge of number of true edges in the networks hence recall curves are not necessary. (a) Comparing joint, parallel, and sequential models using L_1-regression. The joint model (DANCE, grey) performs best with only 50 samples, although the performances of three models are close for the 500 sample sets. (b) Comparing different network inference methods. DANCE (grey) has the best overall performance.

In a typical dataset, the number of variables will be much larger than the number of individuals, therefore the three models defined above can be solved with any regression method able to cope with such a scenario. In the next section we compare two widely used methods, namely, the Gaussian Graphical Model (GGM) (Schäfer and Strimmer, 2005b) and the Partial Least Squares approach (PLS) (Pihur *et al.*, 2008). GGM employs a shrinkage estimator of the covariance matrix to infer partial correlations among hundreds of variables. This approach is designed to recover large networks from datasets with small sample size by multiple testing of the edges based on the local false discovery rate to detect those significant in the network. PLS is also a powerful tool in inferring relations between many variables with high efficiency.

Alternatively, $\tilde{B}$ can be obtained by providing a sparse solution for every $\tilde{\mathbf{b_j}}$ through regression on every response variable $\tilde{y_j}$ with L_1 regularization (Tibshirani, 1994), also known as lasso. L_1-regression has been applied to genome-wide association studies to delineate causal SNPs in disease (Wu *et al.*, 2009; Shi *et al.*, 2007), and is particularly useful in situations where the number of predictors far exceeds the number of samples (Tibshirani, 1997). The lasso estimator for our model is given by optimizing the following objective function:

$$\underset{\tilde{\mathbf{b_j}}}{\operatorname{argmin}} \left\| \tilde{y_j} - \tilde{X}\tilde{\mathbf{b_j}} \right\|_2 + \lambda_j \left\|\tilde{\mathbf{b_j}}\right\|_1, \qquad (6)$$

where λ_j is the regularization parameter controlling the sparsity and strength of regularization. When λ_j is increased the number of non-zero values in $\tilde{\mathbf{b_j}}$ is reduced, so the matrix of coefficients is more sparse. It is a difficult to select *a priori* a value for λ_j, therefore we use cross-validation within a range of possible values for the parameter using the penalized package (Goeman, 2009) with a matrix wrapper in the lol package in R, on which DANCE is based. Subsequently, we fit an ordinary least square regression with the non-zero coefficients, and select only the coefficients with p-value lower than 0.05. This is to exclude non-zero coefficients that are not significantly different from zero.

After choosing an optimal λ_j value for every expression response, we compute the lasso solution by solving Eq.6 and obtain the coefficient matrix $\tilde{B}$. Because of the L_1-constraint in lasso, this matrix will be sparsely populated and many coefficients will be zero. The sub-matrix B^d of $\tilde{B}$ represents the deregulated interactions between the copy number predictors and transcriptional responses. The joint model with and sparse solution that we propose in DANCE is subsequently validated with both simulated and biological data.

3 EXPERIMENTS ON SIMULATED DATA

We use simulated data to show that the joint model outperforms the sequential and parallel models, and that L_1 inference is more accurate than other alternatives.

We generated simulated datasets with the number of predictors p varying from 50 to 1000 to observe the effects of predictor numbers on accuracy. The number of responses q was fixed at 10, which would not affect the inference results as the inference for each response is carried out independently. The sample size n for both the reference and deregulated sample group was tested at 50 and 500. Although for a single cohort $n = 50$ is closer to the real scenario, with combined data sets from multiple studies $n = 500$ is also possible. We also added Gaussian noise $N(0, \sigma^2)$ to the simulated data with noise level σ varied from 0.2 to 0.8, as appears in Fig.1. A sparse reference network B^r was randomly generated with $p^{-4}\times q\times 3$ edges/non-zero coefficients. This is to ensure each response has a reasonable number of regulators. Then, the deregulation network B^d was generated with a set of non-zero coefficients composed by $p^{-4}\times$ q edges. All coefficients were randomly sampled from $N(0, I)$. By allowing overlaps between B^r and B^d, we assume that deregulations occur as a result of changes in the strength of interactions in the reference network as well as due to new interactions. Let X_1, X_2

follow multivariate Gaussian $\sim N(0, \Sigma)$ each with p variables of n samples, where Σ is the covariance matrix where $\Sigma_{ij} = 0.2^{|i-j|}$ to introduce similarities to adjacent variables – a feature of copy number data. Let Y_2 and Y_1 then be defined as $Y_2 = X_2 B^r$ and $Y_1 = X_2(B^r + B^d)$. Finally, for each combination of parameters, 100 replications of the simulations were run incorporating a Gaussian noise to Y_1, Y_2 with different noise levels.

3.1 One-step model versus two-step models

Fig. 1(a) shows the results of L_1-regression applied to the joint, parallel and sequential models. The performance of each of them is measured by precision with respect to the non-zero coefficients estimated in the deregulation network B^d. First, we found that with small sample size ($n = 50$), the joint model (Eq.5) yields consistently more accurate results than the parallel (Eq.3) and sequential models (Eq.4) which learn two structures separately. Not surprisingly, we observed low accuracies for all models with 1000 predictors because of the small sample size. With large sample size 500, all models perform reasonably well, which is also expected.

3.2 Deregulated structure learning with simulated data

Now we turn to the problem of comparing different network inference methods for estimating the deregulation network using the joint model. We selected two existing methods that are applicable to the small sample and large scale problem: a Gaussian Graphical Model (GGM) and the Partial Least Squares approach (PLS), and compared them with L_1 regularization. Precision curves for the inference results are given in Fig.1(b), which correspond to the accuracy of each method in detecting non-zero coefficients in the true deregulation network B^d. Again, recall curves are not necessary as the number of true edges are given to all methods as prior knowledge. The results for PLS when $p = 1000$ and $n = 500$ are missing, because for PLS these data sets are too computationally intensive. Again, we expect low performance for all methods when the sample size (n=50) is far smaller than the number of predictors (p=1000), which suggests that in this scenario network inference is unlikely to yield reliable results. In all other cases, DANCE outperforms the other two methods.

4 ER DEREGULATION IN BREAST CANCER

In this section we demonstrate the effectiveness of the proposed DANCE method on real data using Chin *et al.* (2006) breast cancer dataset, which consists of 89 samples assayed for both mRNA expression and DNA CNAs. The dataset includes 55 ER positive and 34 ER negative samples and thus facilitates a study of deregulation between these two major subclasses of breast cancer. It is known that ER negative breast cancer has worse prognosis than ER positive breast cancer in the early stage, and they appear to be different diseases (Chin *et al.*, 2007). Therefore it is important to understand which molecular components contribute to the deregulation between ER positive and ER negative status.

4.1 DANCE points to ER-NOTCH cross-talk

First, we applied both DANCE and DRI to this dataset to compare their utility to integrate genomic and transcriptomic data. Interestingly, notch signaling has been implicated in adenocarcinoma development in the mouse mammary gland following pathway activation and the diminished expression of NUMB, a negative regulator of the Notch pathway, in as many as 50% of breast cancer samples (Stylianou *et al.*, 2006; Rizzo *et al.*, 2008). Here, we examine the results of DANCE and DRI as applied to these two important pathways in breast cancer.

Technical details We employed a set of 50 genes involved in the ER (based on GO) and Notch (based on KEGG) signaling pathways to assess the ability of DRI and DANCE to uncover meaningful properties of this pathway in the Chin dataset.

As noted in the Introduction, DRI is an integrative tool that aims to identify expression changes decoupled by DNA aberrations under two conditions based on a paired t-test, when both expression and copy number data are available for the same sample. The output is a list of genes altered between two conditions, with the associated confidence given by a significance score. Since it requires pair-wise comparison, the 50 genes were mapped to 48 unique copy number BAC probes and 107 Affymetrix expression probes. As input to the DRI R package, 48 BAC copy number probes were paired with 107 expression probes, resulting in 107 paired data points.

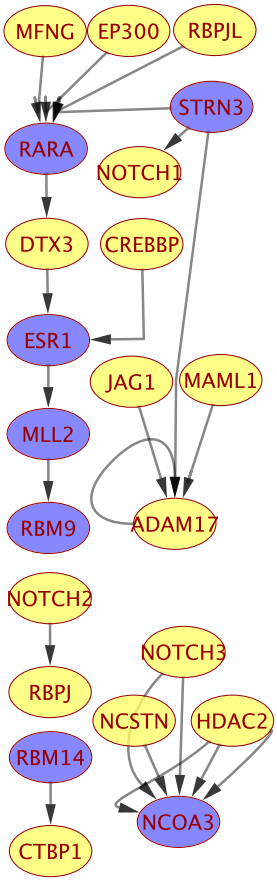
In contrast, DANCE is based on L_1-regression of 48 BAC probes on all 107 expression probes. L_1-regression is a prominent method in high-dimensional studies due to its efficiency in performing sparse statistical inference in the small-sample (small n), but large-scale (large p) setting (Tibshirani, 1997). Note that for DANCE the reference sample class was taken to be the set of ER positive samples, whereas the deregulated class consisted of ER negative samples.

Inference results For the DRI output, we observed a gap between an FDR (FAlse Discovery Rate) cutoff of 0.3 and 0.35 in that few genes were identified with FDR < 0.3, but a significant increase was seen for an FDR > 0.3 and < 0.35. Hence an FDR of 0.3 was selected for this study. Using an FDR cutoff of 0.3, 9 probes (6 genes, *NCOA3*, *ESR2*, *EGLN2*, *PSENEN*, *ESR1*, *DVL1*) were found to be significantly altered between ER positive and ER negative conditions at both the DNA and RNA level.

Instead of merely a list of genes, DANCE is capable of inferring the deregulation network in the ER-Notch pathways as shown in Fig. 2. The deregulation network represents an aberrant situation where copy number and expression are coordinately deregulated in a particular context, such as between ER positive and ER negative disease. Here, an edge traverses from a node representing the influence of copy number of a particular gene to a node representing the expression of another gene to indicate potential regulatory interactions. The genes denoted under the *Copy number* heading correspond to copy number nodes in the deregulation network and represent the source node, whereas genes under the *Expression* heading, represent the expression nodes and are the indirect or direct downstream targets of copy number events. The count indicates the number of times a gene appears in the network, which may result either from multiple probes targeting the gene or else from multiple interactions involving that gene.

Comparing the two sets of genes, it is reassuring that ER-related genes such as *ESR1* and *NCOA3* are identified by both methods. However, the result from DANCE include several important genes implicated in tumourigenesis that are not obtained using DRI (Fig. 2).

Not surprisingly, a central component of the deregulated network in ER negative versus ER positive breast cancer is the estrogen receptor alpha, *ESR1*. More specifically, the deregulation network



Copy-Number	Count	Coef	Chr	Band	Probe	Description
HDAC2	3	0.493	6	q22.1	RP11-51N24	histone deacetylase 2
STRN3	3	0.393	14	q12	RP11-48L1	nuclear autoantigen
NOTCH3	2	0.287	19	p13.12	RP11-107O2	Notch homolog 3
DTX3	1	0.181	12		VYS12P2692	ESTs, Highly similar to Deltex3
ESR1	1	0.173	6	q25.1	RP11-20H19	estrogen receptor 1
ADAM17	1	0.139	2	p25.1	RP11-185P14	a disintegrin and metalloproteinase domain 17 (tumor necrosis factor, alpha, converting enzyme)
MLL2	1	0.132	3	q12	CTD-2014B13	trinucleotide repeat containing 21
EP300	1	0.131	22	q13.2	RP11-206B19	E1A binding protein p300
NOTCH2	1	0.128	1	p12	RP11-29M22	Notch homolog 2
MAML1	1	0.121	5	q35.3	RP11-51D11	Histone methyltransferase. Methylates 'Lys-4' of histone H3; Acts as a coactivator for estrogen receptor by being recruited by ESR1, thereby activating transcription
CREBBP	1	0.121	16	p13.3	RP11-380F1	encodes a gene with intrinsic histone acetyltransferase activity and also acts as a scaffold to stabilize additional protein interactions with the transcription complex
RBPJL	1	0.118	20	q13.12	RMC20P100	recombining binding protein suppressor of hairless (Drosophila)-like
JAG1	1	0.118	20	p12.2	CTC-334G22	Ligand for the NOTCH1 receptor
NCSTN	1	0.113	1	q23.2	RP11-97G24	nicastrin
RBM14	1	0.11	11	q13.1	RP11-527H7	RNA binding motif protein 4
RARA	1	0.109	17	q21.2	RMC17P034	retinoic acid receptor, alpha
MFNG	1	0.108	22	q13.1	RP11-35I10	manic fringe homolog
Expression	**Count**	**Coef**	**Chr**	**Band**	**Probe**	**Description**
ADAM17	3	0.379	2	p25.1	205745_x_at	a disintegrin and metalloproteinase domain 17 (tumor necrosis factor, alpha, converting enzyme)
NCOA3	2	0.317	20	q13.12	211352_s_at	nuclear receptor coactivator 3
NCOA3	2	0.315	20	q13.12	209061_at	nuclear receptor coactivator 3
ESR1	2	0.302	6	q25.1	205225_at	estrogen receptor 1
RARA	2	0.295	17	q21.2	211605_s_at	retinoic acid receptor, alpha
RARA	2	0.227	17	q21.2	203749_s_at	retinoic acid receptor, alpha
MLL2	1	0.173	3	q12	216845_x_at	trinucleotide repeat containing 21
NCOA3	1	0.136	20	q13.12	209060_x_at	nuclear receptor coactivator 3
RBM9	1	0.132	22	q13.1	216215_s_at	Repressor of tamoxifen transcriptional activity
RBPJ	1	0.128	11	p13	211974_x_at	Transcriptional regulator of the Notch pathway; suppressor of hairless homolog
NOTCH1	1	0.125	9	q34.3	218902_at	Notch homolog 1
NCOA3	1	0.125	20	q13.12	207700_s_at	nuclear receptor coactivator 3
CTBP1	1	0.11	4	p16.3	213980_s_at	encodes a protein that binds to the C-terminus of adenovirus E1A proteins. This phosphoprotein is a transcriptional repressor and may play a role during cellular proliferation
DTX3	1	0.109	12	q13.3	221835_at	ESTs, Highly similar to Deltex3
ADAM17	1	0.104	2	p25.1	213532_at	a disintegrin and metalloproteinase domain 17 (tumor necrosis factor, alpha, converting enzyme)

Fig. 2. DANCE infers a deregulation network showing the cross-talks between ER and NOTCH pathways. Blue nodes correspond to genes in the ER pathway while yellow denotes NOTCH pathway. Directed edges origin from the copy number instances of the genes and points to the expression instances of the genes.

indicates the differential regulation of *ESR1* by *CREBBP*, *DTX3*, *RARA*. It is significant that this relationship is recovered since it is well known that estrogen receptor (ER)-positive breast cancer cells are hormonally regulated and are inhibited by retinoids, whereas ER-negative breast cancer cells are generally not (Rosenauer *et al.*, 1998). The retinoic acid receptors (RARs), including *RARA*, are members of the steroidthyroid hormone receptor gene family and are ligand-dependent transcription factors, which exhibit growth inhibitory activity against breast cancer cells (Zarubin *et al.*, 2005).

Interestingly, increased *RBPJ*-dependent Notch signaling has been shown to result in the transformation of normal breast epithelium via the inhibition of apoptosis (Stylianou *et al.*, 2006). Here we show that *RBPJ* is deregulated in ER negative disease as a result of aberrations in *NOTCH2* copy number as are several other downstream components of the Notch signaling pathway, including *NOTCH3* which coordinately influences *NCOA3* expression levels along with *HDAC2* and *NCSTN*. Moreover, *NOTCH3* has been shown to play a crucial role in the proliferation of ErbB2-negative human breast cancers, which may represent a subset of cases. One means by which this might be effected is through alterations in *NCOA3* expression levels since siRNA depletion of *NCOA3* has been shown to increase apoptosis and reduce *ESR1* transcriptional activity in the MCF-7 breast cancer cell line (Karmakar *et al.*, 2009).

Finally, recent studies have identified cross-talk between the estrogen receptor and notch signaling pathways, and suggest novel therapeutic approaches (Rizzo *et al.*, 2008). These studies indicate that estrogen inhibits *NOTCH1* activity by altering its cellular distribution and suggests that estrogen affects a step of Notch activation distal to ligand binding and directly or indirectly inhibits Notch cleavage. In agreement with these findings, DANCE reveals that although *JAG1* is frequently over-expressed in cancer (Reedijk *et al.*, 2005) potentially as a result of amplification, it does not influence *NOTCH1* expression. Rather, it may act downstream to modulate the expression of the matrix metalloproteinase, *ADAM17*, which in turn modulates Notch cleavage. Thus DANCE is able to infer known aspects of cross-talk deregulation in ER-positive versus negative disease in the absence of prior information, indicating its utility in recovering relevant aspects of differential pathway activation in subtypes of disease.

To further examine several of the key pathway components highlighted above, we compared the correlation between the expression profiles of these genes and the copy number states of their putative upstream effectors (based either on pathway structure or the output of DANCE). The expression level of each of these genes was also examined for the ER positive (black) and ER negative (red) subgroups Fig.3.

Here we observe that *STRN3* copy number is significantly correlated with *NOTCH1* expression in ER negative cases only. Note that while the method does not distinguish between direct and indirect effects, the two genes are located on different chromosomes

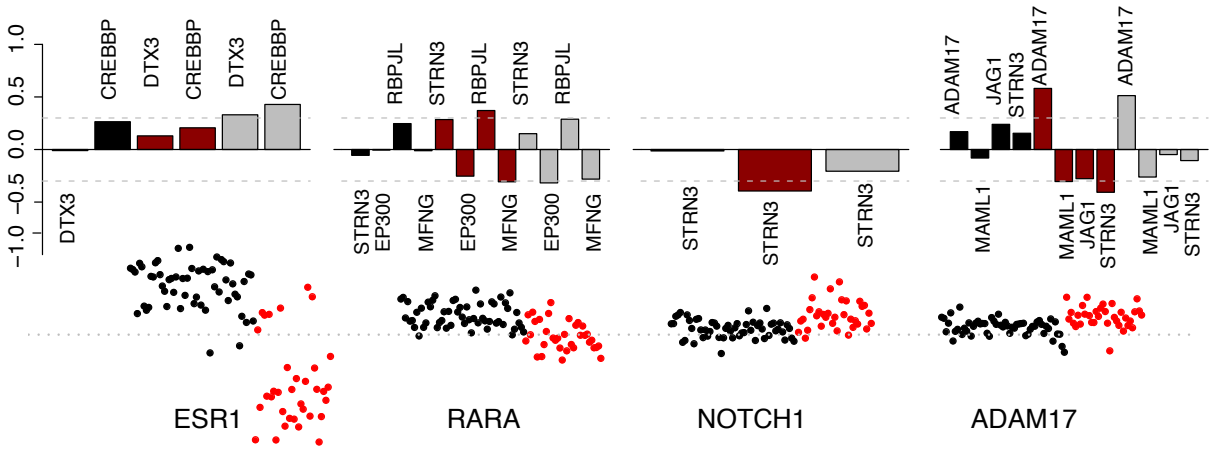


Fig. 3. Expression profiles of selected expression targets and their changes in correlations with their copy number regulators, as inferred by DANCE. The expression profiles of the genes denoted are also plotted separately for the ER positive cases (black) and ER negative cases (red). The correlation between the expression profiles of the genes denoted and the copy number profiles of upstream regulators are shown as barplots for the ER positive cases (black), ER negative cases (red), and all cases (grey). The expression profiles show that these genes should be ER-modulated, while the barplots suggest that the differential regulations may exist between ER subtypes.

and hence represents *trans*-regulation. We also observe that *ESR1* expression levels are substantially lower in ER negative disease, and are positively correlated with *STRN3* copy number and negatively correlated with *EP300* and *MFNG* copy number in ER negative, whereas ER positive samples do not exhibit significant correlation with any of the putative regulators, suggesting differential regulatory mechanisms are at play. We note that several of these target genes were not identified as being deregulated using the DRI approach, which only considers *cis*-effects. Despite the fact that the *trans*-effect of CNAs are often ignored, they represent an important mode of regulation. In summary, we not only confirm some of our observations with the literature, but also observe additional relationships of potential relevance.

5 DISCUSSION

The last few years have yielded an increase in the number of datasets which include multiple genomic measurements on the same sample and this will likely be a trend for sometime to come. However, there are presently limited tools that facilitate the integration of diverse data types. Furthermore, we are not aware of any method that attempts to infer patterns of deregulation resulting from copy number aberrations or expression alterations between groups of samples. The proposed DANCE model is based on a simple, yet powerful framework that benefits from a sparse solution for inference of network deregulation and with several key strengths. First, gene expression changes are interpreted in the context of both *cis*- and *trans*- copy number effects. Additionally, coordinated deregulation of expression and copy number is not restricted to pairs but is open to global search, making it possible to detect concomitant changes in signaling cascades. Finally, the model facilitates a sparse solution resulting in clearly interpretable deregulation network structure, thus minimizing the inference error.

Recent approaches have focussed on transcriptional deregulation visible as changes in the co-expression of groups (mostly pairs) of genes (Mentzen *et al.*, 2009; Mo *et al.*, 2009; Slavov and Dawson, 2009; Xu *et al.*, 2008; Kostka and Spang, 2004). Examining sets of differently expressed genes and successive functional analysis of these sets (e.g. by finding enriched pathways) help point to putative deregulated pathways across several types of cancer (Bild *et al.*, 2006; Edelman *et al.*, 2008; Liu and Ringnér, 2007). As mentioned in the introduction, relying on single data source restricts our view of the complicated biological processes underlying cancer. We emphasize that integrating different data sources helps to circumvent this problem and allows the identification of candidate drivers of gene expression changes.

We have set the deregulation inference in the context of pathways because there is often redundancy between signaling components, and an understanding of pathway regulation is essential for the development of targeted therapeutics. DANCE identifies pathway components that are differentially regulated in specific cellular contexts such as tumour subgroups, as demonstrated above. Such an approach is also extensible to experiments in which a particular pathway component has been perturbed via either knockdown or over-expression, making it a useful method in hypothesis driven studies. In summary, DANCE is computationally efficient and represents a promising tool for the integration of large-scale multi-dimensional genomic datasets.

ACKNOWLEDGEMENTS

We acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited.

REFERENCES

Adjei, A. A. and Hidalgo, M. (2005). Intracellular signal transduction pathway proteins as targets for cancer therapy. *J Clin Oncol*, **23**(23), 5386–5403.

Akavia, U. D. *et al.* (2010). An integrated approach to uncover drivers of cancer. *Cell*, **143**(6), 1005–1017.

Alizadeh, A. A. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–511.

Bild, A. H. *et al.* (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**(7074), 353–357.

Chen, Y. *et al.* (2008). Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.

Cheung, V. G. and Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature reviews. Genetics*, **10**(9), 595–604.

Chin, K. *et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**(6), 529–541.

Chin, S. F. *et al.*(2007). High-resolution array-cgh and expression profiling identifies a novel genomic subtype of er negative breast cancer. *Genome Biology*, **8**, R215+.

Daemen, A. *et al.* (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*, **1**(4), 39.

Edelman, E. J. *et al.* (2008). Modeling cancer progression via pathway dependencies. *PLoS Comput Biol*, **4**(2).

Furge, K. A. *et al.* (2007). Identification of deregulated oncogenic pathways in renal cell carcinoma: an integrated oncogenomic approach based on gene expression profiling. *Oncogene*, **26**(9), 1346–1350.

Goeman, J. J. (2009). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*.

Golub, T. R. *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Greenman, C. *et al.* (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, **446**(7132), 153–158.

Horlings, H. M. *et al.* (2010). Integration of DNA copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clin Cancer Res*, **16**(2), 651–663.

Huang, E. *et al.* (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet*, **34**(2), 226–230.

Hummel, M. *et al.* (2006). A biologic definition of burkitt's lymphoma from transcriptional and genomic profiling. *The New England journal of medicine.*, **354**, 2419–2430.

Karmakar, S. *et al.* (2009). Unique roles of p160 coactivators for regulation of breast cancer cell proliferation and estrogen receptor-alpha transcriptional activity. *Endocrinology*, **150**(4), 1588–1596.

Kostka, D. and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, **20 Suppl 1**, i194–i199.

Liu, Y. and Ringnér, M. (2007). Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis. *Genome Biol*, **8**(5).

Mentzen, W. I. *et al.* (2009). Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumour. *BMC Genomics*, **10**, 601.

Mo, W. J. *et al.* (2009). A stochastic model for identifying differential gene pair co-expression patterns in prostate cancer progression. *BMC Genomics*, **10**, 340.

Pihur, V., Datta, S., and Datta, S. (2008). Reconstruction of genetic association networks from microarray data: A partial least squares approach. *Bioinformatics*. January.

Pollack J.R. *et al.* (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet.* 1999 Sep;23(1):41-6.

Pollack, J. *et al.* (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumours. *Proc Natl Acad Sci USA*, **99**, 12963–12968.

Reedijk, M. *et al.*(2005). High-level coexpression of JAG1 and NOTCH is observed in human breast cancer and is associated with poor overall survival. *Cancer Res*, **65**(18), 8530–8537.

Rhodes, D. R. *et al.* (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, **101**(25), 9309–9314.

Rizzo, P. *et al.* (2008). Cross-talk between NOTCH and the estrogen receptor in breast cancer suggests novel therapeutic approaches. *Cancer Res*, **68**(13), 5226–5235.

Rosenauer, A. *et al.* (1998). Estrogen receptor expression activates the transcriptional and growth-inhibitory response to retinoids without enhanced retinoic acid receptor alpha expression. *Cancer Res*, **58**(22), 5110–5116.

Salari, K. *et al.* (2009). DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics*, page btp702.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist Appl Genet Mol Biol*, **4**, 32.

Schäfer, M. *et al.* (2009). Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, **25**(24), 3228–3235.

Segal, E. *et al.* (2004). A module map showing conditional activity of expression modules in cancer. *Nat Genet*, **36**(10), 1090–1098.

Shi, W. *et al.* (2007). Detecting disease-causing genes by Lasso-patternsearch algorithm. *BMC Proceedings*, **1**(Suppl 1), S60+.

Slavov, N. and Dawson, K. (2009). Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proceedings of the National Academy of Sciences*, **106**, 4079.

Stylianou, S. *et al.* (2006). Aberrant activation of NOTCH signaling in human breast cancer. *Cancer Res*, **66**(3), 1517–1525.

The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**(7216), 1061–1068.

Tibshirani, R. (1994). Regression shrinkage and selection via the Lasso.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**(4), 385–395.

van de Wiel, M.A. and van Wieringen, W.N. (2007). CGHregions: Dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, (2).

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine*, **10**(8), 789–799.

Watters, J. W. and Roberts, C. J. (2006). Developing gene expression signatures of pathway deregulation in tumours. *Molecular Cancer Therapeutics*, **5**(10), 2444–2449.

West, M. *et al.* (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, **98**(20), 11462–11467.

Wu, T. T. *et al.* (2009). Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics*, **25**(6), 714–721.

Xu, M. *et al.* (2008). An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC genomics*, **9 Suppl 1**, S12.

Zarubin, T. *et al.*(2005). Identification of eight genes that are potentially involved in tamoxifen sensitivity in breast cancer cells. *Cell Res*, **15**(6), 439–446.

Zhang, B. *et al.* (2008). Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*.