BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 7 2010, pages 896–904
doi:10.1093/bioinformatics/btq051

Gene expression

Advance Access publication February 9, 2010

Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data

Antonio Reverter[1,*], Nicholas J. Hudson[1], Shivashankar H. Nagaraj[1], Miguel Pérez-Enciso[2] and Brian P. Dalrymple[1]

[1] Bioinformatics Group, CSIRO Livestock Industries, Queensland Bioscience Precinct, 306 Carmody Road, St. Lucia, Brisbane, Queensland 4067, Australia and [2] ICREA—Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

Associate Editor: David Rocke

ABSTRACT

Motivation: Although transcription factors (TF) play a central regulatory role, their detection from expression data is limited due to their low, and often sparse, expression. In order to fill this gap, we propose a regulatory impact factor (RIF) metric to identify critical TF from gene expression data.

Results: To substantiate the generality of RIF, we explore a set of experiments spanning a wide range of scenarios including breast cancer survival, fat, gonads and sex differentiation. We show that the strength of RIF lies in its ability to simultaneously integrate three sources of information into a single measure: (i) the change in correlation existing between the TF and the differentially expressed (DE) genes; (ii) the amount of differential expression of DE genes; and (iii) the abundance of DE genes. As a result, RIF analysis assigns an extreme score to those TF that are consistently most differentially co-expressed with the highly abundant and highly DE genes (RIF1), and to those TF with the most altered ability to predict the abundance of DE genes (RIF2). We show that RIF analysis alone recovers well-known experimentally validated TF for the processes studied. The TF identified confirm the importance of PPAR signaling in adipose development and the importance of transduction of estrogen signals in breast cancer survival and sexual differentiation. We argue that RIF has universal applicability, and advocate its use as a promising hypotheses generating tool for the systematic identification of novel TF not yet documented as critical.

Contact: tony.reverter-gomez@csiro.au

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

Received on October 16, 2009; revised on January 27, 2010; accepted on February 3, 2010

1 INTRODUCTION

Transcription factors (TF) play a central regulatory role in controlling gene expression. Previous studies demonstrate that TF are important in both normal and disease states (Vaquerizas *et al.*, 2009). However, low TF expression make their detection challenging and warrants alternative *in-silico* methods to facilitate the identification of critical TF from gene expression data.

In an attempt to derive more information from expression data, recent work has been devoted to inferring transcriptional regulation from expression data. By and large, these methods invoke the well-documented guilt-by-association heuristic by which groups of genes targeted by the same TF, and/or involved in the same biological pathways, have an expression profile that is more correlated than a randomly chosen group of genes (Wolfe *et al.*, 2005). Inspired by such heuristic, a rational approach for exploiting this co-expression phenomena and deciphering transcriptional regulation activity involves the reverse-engineering of gene regulatory networks using network inference algorithms such as (but not limited to) Bayesian networks (Friedman *et al.*, 2000), CLR (Faith *et al.*, 2007); ARACNe (Margolin *et al.*, 2008) and PCIT (Reverter and Chan 2008; Watson-Haigh *et al.*, 2010). Ergün *et al.* (2007) exploited the connectivity structure of a gene network to a test expression data and identified genetic drivers of prostate cancer using the so-called MNI algorithm (di Bernardo *et al.*, 2005). Other authors have undertaken a promoter sequence analysis of a correlated group of genes to identify sequence motifs corresponding to TF binding sites (Cowley *et al.*, 2009; Kerhornou and Guigó, 2007; Nagaraj *et al.*, 2008). An equally commendable strategy relies on assigning regulators to modules based on the co-expression between a candidate regulator and each of the members of the module. Examples of the latter approach include the learning module networks (LeMoNe) algorithm of Joshi *et al.* (2009) which generates a number of possible models explaining regulation activity and with every single model containing many regulators. An alternative method, initially introduced by Reverter *et al.* (2006a) and more recently implemented in Hudson *et al.* (2009a), is based on ranking TF by their absolute co-expression correlation averaged across all genes in a given module.

We recently described a regulatory impact factor (RIF) algorithm which correctly inferred myostatin as the gene containing the causal mutation from gene expression data alone, even though myostatin was not differentially expressed (DE) at any of ten developmental time points under surveillance (Hudson *et al.*, 2009b). This algorithm addresses an important biological issue because it better accounts for the functional activation of TF than does DE alone. For example, TF are activated following reversible phosphorylation, ligand binding, cellular localization, co-factor binding, missense mutations and 'receptiveness' of chromatin structure. Differential expression will overlook these vital changes in regulatory information, yet, a full interpretation of expression data clearly requires some means of quantification.

*To whom correspondence should be addressed.

In this article, we attempt to determine whether application of the RIF algorithm is generalizable. Is there a universal question one can ask of appropriately designed gene expression experiments to identify (i) causal regulators and (ii) the rewired transcriptional circuits through which they exert their phenotypic impact? In order to ascertain the generality of RIF, we explore the publicly available expression data from four experiments that cover a wide range of scenarios from *in-vitro* to *in-vivo* systems, from embryonic to adult stages, from developmental time-series to discrete perturbations.

Our study is organized in the following manner: we first provide an overview of how RIF operates. Next, we introduce the four datasets and put emphasis, not only in the relevance of the biological question each experiment addresses, but also in the design layout and the number of genes and TF included in each experiment, and where TF are obtained from the census provided by Vaquerizas *et al.* (2009). We then describe the normalization method and how DE genes are identified. Finally, we highlight the significance of computing two alternative measures of RIF (RIF1 and RIF2) and present the results in context of functional biology.

2 METHODS

2.1 An overview of RIF analysis

Figure 1 illustrates a schematic diagram of the process involved in RIF analysis. A microarray gene expression dataset spanning two biological conditions of interest (e.g. healthy and disease) is subjected to standard normalization techniques and significance analysis to identify the target genes whose expression is DE between the two conditions. Simultaneously, the collection of regulators (e.g. TF genes) included in the microarray data is mined from the literature (Vaquerizas *et al.*, 2009). Next, the co-expression correlation between each TF and the DE genes is computed for each of the two conditions. This allows for the computation of the differential wiring (DW) from the difference in co-expression correlation existing between a TF and a DE genes in the two conditions. As a result, RIF analysis assigns an extreme score to those TF that are consistently most differentially co-expressed with the highly abundant and highly DE genes (case of RIF1 score), and to those TF with the most altered ability to act as predictors of the abundance of DE genes (case of RIF2 score). Importantly, and as illustrated by the bottom right panel of Figure 1, a given TF may not show a change in expression profile between the two conditions to score highly by RIF as long as it shows a big change in co-expression with the DE genes. To this particular, the profile of the TF gene (triangle, solid line) is identical in both conditions (slightly downwards). Instead, the DE gene (circle, dashed line) is clearly over-expressed in condition B. Importantly, the expression of the TF and the DE gene shows a strong positive correlation in condition A, and a strong negative correlation in condition B.

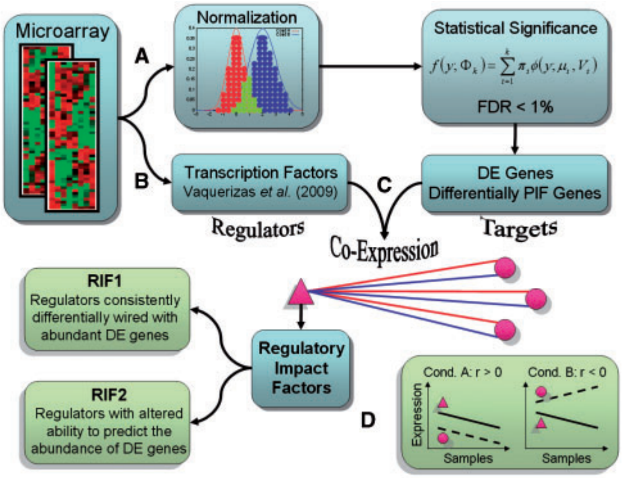
Microarray
A
B
Normalization
Statistical Significance
$f(y;\Phi_k)=\sum_{t=1}^{k}\pi_t\phi(y;\mu_t,V_t)$
FDR < 1%
Transcription Factors
Vaquerizas *et al.* (2009)
C
DE Genes
Differentially PIF Genes
Regulators
Co-Expression
Targets
RIF1
Regulators consistently differentially wired with abundant DE genes
RIF2
Regulators with altered ability to predict the abundance of DE genes
Regulatory Impact Factors
D
Cond. A: r > 0
Cond. B: r < 0
Expression
Samples
Samples

Fig. 1. A schematic diagram of the RIF analysis. (**A**) Microarray data is normalized and statistically assessed to identify differentially expressed (DE) genes and differentially PIF genes (represented by circles) which together are deemed as the Target genes; Simultaneously, (**B**) transcription factors (TF, represented by triangles) included in the microarray are collected and (**C**) their co-expression correlation with the target genes computed for each of the two conditions of interest; Finally, (**D**) the way in which TF and target genes are differentially co-expressed between the two conditions is used to compute the relevance of each TF according to RIF1 and RIF2.

2.2 Datasets

The first dataset is from the study of Timmons *et al.* (2007) who used *in-vitro* cell cultures to explore the mechanisms underlying brown and white adipocyte differentiation. A total of 24 hybridizations were performed using the RNA from two cell types (brown versus white adipocytes), cultured at two ages (4 and 7 days old), with five and six, and biological replicates for the brown and white adipocyte cultures, respectively. Using the MAS5 detection call utility, probes yielding an absent signal in all 24 hybridizations were removed. As a result, we retained 159 768 expression intensity readings from 5665 unique genes including 552 TF.

The second dataset is from the study of Small *et al.* (2005) who profiled gene expression during the differentiation and development of embryonic gonads in mice. The authors used 60 microarray chips each with 12 000 probe sets representing ∼8000 genes. The experimental design corresponds to a time course of gene expression in embryonic gonads (testes versus ovaries) at five time points post-coitum: 11.5 (indifferent gonads), 12.5, 14.5, 16.5 and 18.5 days (birth). Six biological replicates were available. After editing out probes with absent signal in all hybridizations, a total of 282 360 expression records from 9552 genes including 809 TF were used in the present study.

The third dataset is from Pérez-Enciso *et al.* (2009) who used 80 Porcine Affymetrix chips (each representing ∼15 000 genes) to survey the gene expression profile in a 4 (breeds) ×5 (tissues) ×2 (sexes) factorial design and two biological replicates. Using identical data editing criteria as in the previous datasets, the porcine dataset included 1 575 760 expression records (half for each sex) on 11 266 genes of which 912 were TF. The sex contrast, male versus female, was explored in the RIF analyses.

The last dataset belongs to the breast cancer survival study of Van't Veer *et al.* (2002) where 78 cDNA microarray chips were hybridized using the RNA samples from 34 and 44 patients with <5 and >5 years survival time, respectively. Log-ratios and associated *p*-values were downloaded from the original source and, for filtering purposes, log-ratios with associated *P*-values >0.9 were deemed as 'absent' and genes non-absent in more than eight samples (i.e. >10% of samples) were retained. As a result, the present study utilized 1 888 848 log-ratios on 22 635 genes including 892 TF.

2.3 Normalization and differential expression

As previously described (Reverter *et al.*, 2006b), a combination of ANOVA models and mixtures of distributions were employed to normalize expression signals and to identify DE genes, respectively. In detail, for each of the four datasets, data normalization was achieved by fitting a parsimonious mixed-effect ANOVA model with the following components:

$$\mathbf{y}_{ijk} = \mathrm{H}_i + \mathrm{G}_j + \mathrm{GC}_{jk} + \mathrm{e}_{ijk}, \tag{1}$$

where $\mathbf{y}_{ijk}$ is the vector of expression readings from the ith hybridization chip, on the jth gene at the kth condition; $\mathrm{H}_i$is the fixed effect of the ith hybridization and the fitting of which aims at normalizing the data by accounting for systematic non-genetic effects; G_j is the random effects of the average level of the jth gene; GC_{jk} is random interaction between the *j*th gene and the *k*th experimental condition and it captures differences from overall averages that are attributable to specific gene-condition combination; and $e_{ijk}$is the random residual error associated with $\mathbf{y}_{ijk}$.

Using standard statistical assumptions in mixed model theory, the effects of G_j, GC_{jk} and e_{ijk}were assumed to be independent realizations from a normal distribution with zero mean and between-gene, between-gene within-condition and within-gene components of variance, respectively. Restricted maximum likelihood estimates of variance components and solutions to model effects were obtained using the analytical gradients option of VCE6 software (ftp://ftp.tzv.fal.de/pub/vce6/). The solutions to the GC_{jk}effect were used as the normalized mean expression of each gene in each of the conditions under scrutiny. Finally, the difference between the normalized mean expression of a gene in the two conditions was computed as the measure of (possible) differential expression.

Following McLachlan *et al.* (2006), a two-component normal mixture model was fitted to identify DE genes.

$$f\left(\mathbf{d};\,\Phi\right)=\pi_0\phi_0\left(\mathbf{d};\,\mu_0,\,\sigma_0^2\right)+\pi_1\phi_1\left(\mathbf{d};\,\mu_1,\,\sigma_1^2\right), \qquad (2)$$

where **d** denotes the vector of DE measures for all the genes, and the two components in the mixtures correspond to: $\phi_0(\bullet)$ for the empirical null normal density with mean μ_0(not necessarily zero) and variance σ_0^2 (not necessarily one), encapsulating the non-DE genes; and $\phi_1(\bullet)$ for the non-null distribution corresponding to DE genes. Finally, the mixing proportions π_0 and π_1 are constrained to be non-negative and sum to unity.

Across the four datasets, parameters of the mixture model were estimated using the EMMIX-GENE software (McLachlan *et al.*, 2002) and an estimated experiment-wise false discovery rate (FDR) of <1% used as the threshold for determining which genes are DE.

2.4 Measures of RIF

RIF is a metric given to each TF that combines the change in co-expression between the TF and the DE genes (i.e. the potential targets). Two alternative measures of RIF are explored and computed as follows:

$$\begin{aligned}\mathrm{RIF1}_i&=\frac{1}{\mathrm{n}_{de}}\sum_{j=1}^{j=\mathrm{n}_{de}}\hat{a}_j\times\hat{d}_j\times\mathrm{DW}_{ij}^2 \qquad (3)\\ &=\frac{1}{\mathrm{n}_{de}}\sum_{j=1}^{j=\mathrm{n}_{de}}\mathrm{PIF}_j\times\mathrm{DW}_{ij}^2\end{aligned}$$

and

$$\mathrm{RIF2}_i=\frac{1}{\mathrm{n}_{de}}\sum_{j=1}^{j=\mathrm{n}_{de}}\left[\left(e1_j\times r1_{ij}\right)^2-\left(e2_j\times r2_{ij}\right)^2\right], \qquad (4)$$

where n_{de} is the number of DE genes; $\hat{a}_j$is the estimated average expression of the *j*th DE gene, averaged across the two conditions being contrasted; $\hat{d}_j$ is the estimated differential expression of the *j*th DE gene; and DW is the differential wiring between the *i*th TF and the *j*th DE gene, and computed from the difference between $r1_{ij}$ and $r2_{ij}$, the co-expression correlation between the *i*th TF and the *j*th DE gene in conditions 1 and 2, respectively (Hudson *et al.*, 2009b):

$$\mathrm{DW}_{ij}=r1_{ij}-r2_{ij}. \qquad (5)$$

The expression for $\mathrm{RIF1}_i$ in Equation (3) introduces the concept of phenotype impact factor (PIF) defined for each DE gene and computed from the product of its average expression and its differential expression. Decomposing its terms, PIF can be expressed as follows:

$$\mathrm{PIF}_j=\hat{a}_j\times\hat{d}_j=\frac{1}{2}\left(e1_j+e2_j\right)\left(e1_j-e2_j\right)=\frac{1}{2}\left(e1_j^2-e2_j^2\right) \qquad (6)$$

where $e1_j$ and $e2_j$ represent the expression of the *j*th DE gene in conditions 1 and 2, respectively. The definition of PIF_j in Equation (6) as the difference of squared expression allows for the alternative parameterization of RIF presented by $\mathrm{RIF2}_i$ in Equation (4), and where the difference of squared expression is weighted by the squared co-expression correlation between the TF and the DE genes in each of the two conditions. Recall that the squared correlation is equal to the coefficient of determination, a measure of goodness of fit representing the proportion of the variation in the response variable (i.e. the DE gene in our context) that is accounted for by the predictor (i.e. the TF gene in our context). Hence, this new definition of RIF shares the spirit of regression-based approaches to infer gene regulation [examples span from the NIR algorithm (Gardner *et al.*, 2003); to the very recent TILAR algorithm of (Hecker *et al.*, 2009)]. As first noted by Hudson *et al.* (2009b), RIF2 has the additional appeal of not being zeroed for self-regulated genes when a TF is also a DE gene, in which case DW = 0. In essence, while RIF1 captures those TF showing a large DW to those highly abundant highly DE genes (indeed the original question that gave rise to the discovery of RIF), RIF2 focuses on those TF showing evidence as predictors of the change in abundance of DE genes.

In order to allow comparing both measures of RIF between themselves and across datasets, RIF measures were transformed to a *z*-score by subtracting the mean and dividing by the standard deviation (SD).

Finally, we note that while a strong correlation is expected between DE and PIF, the latter places emphasis on the abundant genes that are not hugely DE (on the grounds that a relatively small change in expression of a very abundant transcript is predicted to have a relatively large impact on the molecular phenotype). On the other hand, a non-abundant gene will have to show a large DE in order to be differentially PIF. Similar to the way in which DE genes were determined, we will apply a two-component mixture model to identify genes that are differentially PIF at FDR <1%.

3 RESULTS

3.1 Mixture models, DE genes and differentially PIF genes

Table 1 presents the parameter estimates of the two-component mixture models in each dataset along with the number of DE genes and differentially PIF genes. While both components in the mixture have a mean close to zero, the larger variance estimated for the second component allowed it to more likely capture the extreme values of DE and PIF. Also, this second component was associated with the smaller of the two mixing proportions.

As expected, strong correlations were observed between DE and PIF. These equated to 0.89, 0.92, 0.91 and 0.94 for datasets 1–4,

Table 1. Parameter estimates for the 2-component mixture model and number (*N*) of DE and differentially PIF genes at FDR <1% in each of the four datasets under study

Data[a]		Parameters of the mixture						*N*
		π_0	μ_0	σ_0^2	π_1	μ_1	σ_1^2	
1	DE	0.87	−0.05	0.37	0.13	−0.60	4.52	219
	PIF	0.81	0.41	15.83	0.19	−1.99	141.9	226
2	DE	0.82	−0.05	0.21	0.18	0.30	2.91	545
	PIF	0.81	−0.34	11.56	0.19	1.96	105.5	393
3	DE	0.84	0.01	0.04	0.16	−0.01	0.50	517
	PIF	0.87	0.09	2.12	0.13	−0.23	21.16	306
4	DE	0.72	0.12	0.03	0.28	−0.01	0.07	328
	PIF	0.88	0.58	0.96	0.12	−0.73	5.73	439

[a] Data 1 = Brown *vs* White adipocytes; Data 2 = Testes *vs* Ovaries embryogeneis; Data 3 = Males *vs* Females pigs; Data 5 = more than 5 years *vs* less than 5 years survival to breast cancer.

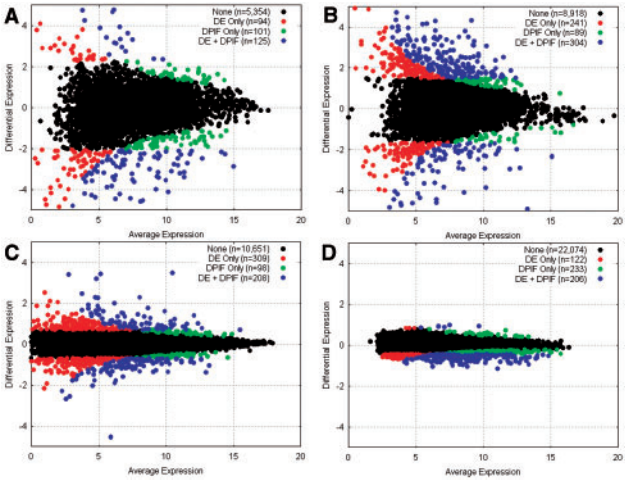

Fig. 2. Scatter of differential expression (y-axis) against the average expression (x-axis) for the four datasets: (**A**) Brown versus white adipocytes; (**B**) Testes versus ovaries differentiation; (**C**) Male versus female pigs; (**D**) >5 years versus <5 years survival to breast cancer. Color codes correspond to red, green and blue for DE genes only, differentially PIF genes only and DE and differentially PIF genes, respectively.

respectively. In consequence, we found a substantial number of genes overlapping by showing both DE and differential PIF. To this end, Figure 2 shows the scatter of DE (y-axis) over the average expression (x-axis). We use color codes in order to enhance distinguishing among genes that are DE only, differentially PIF only, and overlapping genes. As expected, application of PIF increases the representation of highly expressed genes among the key genes (Fig. 2). We use the same scales across the four panels to better appreciate the varying heterogeneity existing in the four datasets.

3.2 RIF analysis

Listed in Table 2 are the RIF1 and RIF2 z-scores for critical TF in each dataset along with their average and differential expression, while the same set of statistics for all TF and across the four datasets is given in the Supplementary Table.

Figure 3 shows the relationship between RIF1 and RIF2 for the four datasets, while the comparison of RIF with DE values is illustrated in Figure 4.

Contrary to our previous findings comparing two breeds of cattle (Hudson *et al.*, 2009b), no particular relationship was found between the two alternative measures of RIF in these datasets. The correlation coefficient (r) between RIF1 and RIF2 was estimated at approximately zero for datasets 1, 2 and 3, and moderately positive for dataset 4 ($r = 0.33$). These results suggest that both measures of RIF capture different, potentially equally valuable features when ranking TF (see 'RIF1 versus RIF2' section, later in this manuscript). In the remainder of this section, we will discuss the biological relevance of the predicted key TF.

3.3 Brown versus white adipocyte differentiation

Compared to white fat, brown fat contains a much higher number of mitochondria, more capillaries and is densely innervated (Nechad *et al.*, 1994).

Table 2. Average (A) and differential expression (DE) for critical TF identified by either RIF1 or RIF2 in each of the four datasets under study

TF	A	DE	RIF1	RIF2
	(1) Brown versus white adipocyte differentiation			
CREBBP	3.22	−2.47	−3.93	−0.43
CUTL1	7.72	−0.80	−0.86	−2.30
MYOG	5.28	2.32	−0.25	2.31
PPARBP	3.95	−0.59	−3.14	−0.22
RBL1	5.06	1.28	−0.16	2.31
	(2) Testes versus ovaries embryonic differentiation			
CBX5	8.10	0.22	2.34	−0.25
FIG1A	3.33	−3.02	−0.91	2.72
HOXC10	5.74	−0.06	−0.80	2.54
HOXD9	8.06	0.83	2.26	−0.49
NCOA3	6.74	−0.09	−0.94	2.26
POU4F1	3.73	−1.38	2.36	−0.36
	(3) Male versus female pre-pubertal pigs			
CHD9	12.12	−0.02	3.90	1.92
IRX3	5.47	−0.06	2.17	3.93
RNF14	9.72	0.36	4.40	1.89
SOX5	7.81	−0.37	3.10	1.20
TAF7L	5.34	1.60	−0.37	3.42
ZNF281	7.97	0.14	−1.99	2.94
	(4) Breast cancer: >5 years versus <5 years survival			
ABL1	7.22	0.20	−3.28	0.78
CARM1	10.20	−0.01	−4.28	−0.31
MAZ	13.17	−0.57	−1.27	−2.68
NFATC4	8.58	−0.28	−6.37	−1.48
NR2F1	8.58	0.09	−3.32	−0.19
PITX3	10.76	−0.40	−2.45	−2.99
RELA	9.14	0.11	−3.67	0.07
SMARCA2	8.65	−0.04	0.41	2.97
SMARCA4	11.42	−0.23	0.67	3.22

A number of TF were identified that positively or negatively regulate brown adipocyte development (Fig. 5), however, an equivalent list is not available for white adipocytes. RBL1 (p107) is awarded the third most positive RIF2 out of its 552 TF competitors (Table 2). In transgenic KO mice experiments, loss of this TF has been shown to culminate in a uniform replacement of white fat with brown fat (Scime *et al.*, 2005). However, RBL1 is not DE between the two tissues and therefore its central regulatory role cannot be inferred through conventional expression statistics. Rather, it is its huge change in network connectivity, in the absence of DE, which helps RIF analysis infer a major role for RBL1 in the brown and white adipocyte lineages.

Other TF of relevance captured by RIF analysis include CREBBP and PPARBP which have been show to play an important role in regulating adiposity and insulin resistance (Tsuchida *et al.*, 2005). CREBBP was found to be up-regulated in white adipocytes and interacts with CUTL1 which was identified in our analysis. CUTL1 interacts with RB1, which in turn regulates the expression of RBL1. PPARBP, which is firmly involved in PPAR signalling, was not DE and showed a very low average expression. Similarly, MYOG, the subject of the striking discovery in the original article of Timmons *et al.* (2007) was also found to be DE in the present analysis (2.32 in the $\log_2$-scale, or 5-fold increase in brown fat; Table 2), and given

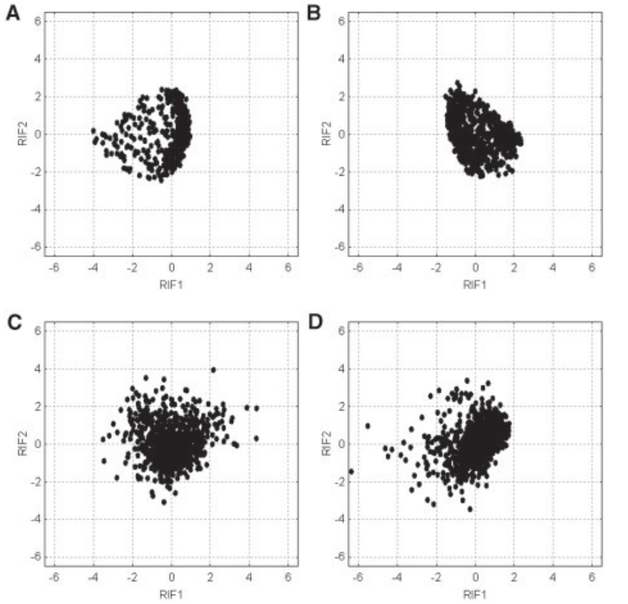

Fig. 3. Relationship between RIF1 and RIF2 for the TF included in each datasets: (**A**) Brown versus white fat; (**B**) Testes versus ovaries; (**C**) Male versus female pigs; (**D**) >5 years versus <5 years breast cancer survival.

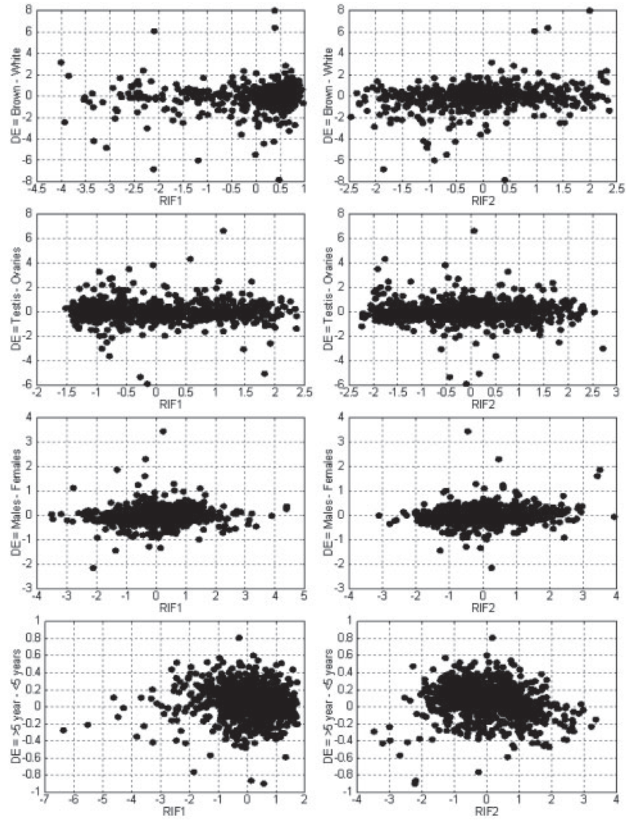

Fig. 4. Relationship between differential expression (DE; *y*-axis) and the two alternative measures of RIF (*x*-axis) for each of the four datasets.

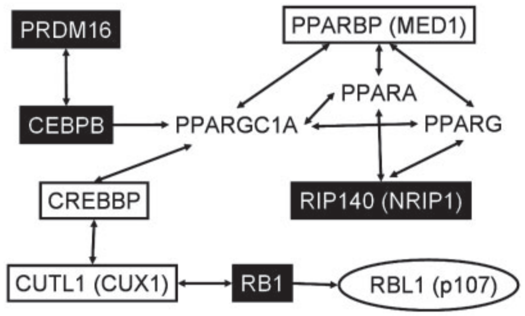

Fig. 5. Protein–protein and gene expression regulatory network involved in the specification of brown adipose tissue. Proteins with published roles are in white on black background. Proteins encoded by genes identified by the RIF analysis are in black on white background. Rectangles and ellipses indicate published and unpublished roles, respectively.

the fifth most extreme ranking according to RIF2. Overall, three of the TF identified by the RIF analysis point to a major role of the broader PPAR pathway in the differences between brown and white fat.

3.4 Testes versus ovaries embryonic differentiation

The gonads differentiation is better viewed as a number of transiently (i.e. often of relevance at one time point only) important TF operating in a successive regulatory cascade—first set in motion by the master regulator SRY (Wilhelm *et al.*, 2007). These various downstream TF may be awarded similar RIF rankings in our analysis, even though those that are highly DW at the earlier time points might be considered more fundamental or ‘causal’ from a biological perspective. For this reason the gonads data set is arguably the least amenable to RIF analysis. In spite of this limitation, RIF uncovered a number of well-documented TF involved in sex differentiation and gonad embryogenesis (Table 2): Figure 1A, a sex-specific marker gene (Scholz *et al.*, 2003); POU4F1 plays an important role during germ cell development (Budhram-Mahadeo *et al.*, 2001); CBX5 is involved in *de novo* methylation and its role in gonad development in mouse embryos has recently been established (Takada *et al.*, 2009); and NCOA3, a member of the steroid receptor co-activator family, contributes to the genetic control of androgenic hormone levels (Sheu *et al.*, 2006).

Finally, we note the ability of RIF analysis to identify two members of the homeobox family (HOXC10 and HOXD9). HOX genes encode evolutionarily conserved TFs which are important regulators of embryonic morphogenesis and tissue differentiation (Dessain *et al.*, 1992).

3.5 Male versus female pre-pubertal pigs

RNF14 showed the most extreme score according to RIF1 (4.40 SD units) and is known to interact with androgen receptor (AR) acting as a coactivator that induces AR-target gene transcription in prostate (Lan *et al.*, 2008). Similarly, IRX3 had the most extreme score according to RIF2 (3.93 SD units) and is a known candidate gene for sex determination (García-Ortiz *et al.*, 2009; Jorgensen and Gao 2005).

Two additional TF (CHD9 and ZNF281) were found to have an extreme score according to both RIF1 and RIF2 even though none of them was found to be DE (Table 2). While not much is known about CHD9, it has been implicated in the transcriptional regulation of osteoblast maturation (Shur *et al*., 2006) and the gender-specificity behind the molecular mechanisms underlying the regulation of ossification has long been established (Hong *et al*., 2009; and references therein). Also, very little literature exists on ZNF281, but it has very recently been shown to be itself regulated by SOX4, which is responsible for the precise differentiation and proliferation in multiple tissues (Scharer *et al*., 2009).

SOX5, which ranked highly according to RIF1 (3.10 SD units), is the co-activator of SOX9 (sex-determining region Y-type high mobility group box 9) and has recently been shown to play a role in sex reversal of the hermaphrodite red-spotted grouper (Huang *et al*., 2009). Arguably, the most sex-determining gene is SRY (see Wilhelm *et al*., 2007; and references therein) which in our analysis was not found to be either DE or to have an extreme RIF. This is most likely because SRY sets in motion the regulatory cascade earlier than the time period assayed in this experiment. We would predict SRY to be highly differentially wired to the highly DE genes at those earlier time points. However, the DE gene with the highest DW with SRY was parathyroid hormone-like hormone (PTHLH). Interestingly, an association has been reported between PTHLH and the number of functional and inverted teats in pigs (Tetzlaff *et al*., 2009).

A final examination of the highly-ranked TF revealed that, with the exception of TAFL7 which was over-expressed in the gonads relative to the other tissues, none of the remaining TF were DE in the original across-tissues analysis of Pérez-Enciso *et al*. (2009).

3.6 Breast cancer

A brief mining of the literature for our highly ranked TF (Table 2) revealed that a number of them are implicated in breast cancer either as oncogenes, tumor suppressors or as biomarkers. For instance, NFATC4 is a transcriptional coactivator of estrogen receptors in breast cancer cells (Zhang *et al*., 2007). Proto-oncogene ABL1, also with a significant RIF1 score, has long been established to be associated with breast cancer (Uhlen *et al*., 2005). In addition, PITX3 is a prognostic and diagnostic epigenetic biomarker for breast cancer (Dietrich *et al*., 2009).

Furthermore, we observed a number of TF with indirect association with breast cancer including: (i) CARM1, an essential co-activator for estrogen-induced breast cancer (Frietze *et al*., 2008); (ii) the role of chromatic remodelers (e.g. SMARCA2 and SMARCA4 in our case) has been well described [(see for instance the recent review by Reisman *et al*. (2009)], and in particular, SMARCA4 has been shown to interact with BRCA1 providing links between chromatin remodeling and breast cancer (Bochar *et al*., 2000); (iii) MYC-associated zing finger protein (MAZ) is responsible for the high expression of PPARG in breast cancer (Wang *et al*., 2008); (iv) the nuclear receptor NR2F1 interacts with estrogen receptor and regulates the expression of estradiol influencing the proliferation of breast cancer cells (Le Dily *et al*., 2008); and (v) RELA (nuclear factor kappa enhancer binding protein) regulates immunity, inflammation and apoptosis (see Skaug *et al*., 2009, for a review) and its association with breast cancer has long been documented (Neil *et al*., 2009).

We also compared our list of TF with a similar study using the same dataset from Cheng *et al*. (2009) in which gene expression data was integrated with transcription factor binding site (TFBS) information to identify potential TF associated with specific cancer type. Cheng *et al*. (2009) identified 26 TF at the 0.01 significance level ($Q < 0.01$) (six with positive correlationand 20 with negative correlation to DE genes) whereas our analyses revealed 71 TF with significant RIF scores. We also found some overlap in the TF family by comparing the two lists (PAX and GATA), but the majority of the TF that RIF identified was not identified by the position weight matrix approach of Cheng *et al*. (2009). However, whilst they did not find a strong signal associated with estrogen, the set of the top 20 TF identified by RIF (Table 2) contained six genes that interact with ESR1 (BAZ1B, NFATC4, NR2F1, PRDM2, SMARCA2 and SMARCA4). In fact, among the 892 TF included in the analyses, 71 had a RIF z-scores < -2 or >2. Of these, 12 are known to interact with ESR1 resulting in an over-representation hypergeometric test P-value of 8.73E−04. This is consistent with the demonstrated association between ESR1 status and prognosis for breast cancer.

3.7 Promoter sequence analyses

In order to obtain an independent evidence of the optimality of RIF, the results from applying the RIF algorithm to the breast cancer data were subjected to promoter sequence analysis to identify TF with TFBS in the promoter region of our target genes (i.e. DE and/or differentially PIF genes).

The MatInspector tool (Cartharius *et al*., 2005) within Genomatix suite (www.genomatix.de) was used to extract genome-wide TFBS for human (including 93 342 promoters in 31 883 loci). When cross-referencing our list of 561 targets against the human promoterome, we identified 12 TF with RIF z-scores < -2 or >2 and with TFBS in the promoter region of 242 target genes. These included MAZ and PITX3 already discussed (Table 2) with 191 and 5 TFBS, respectively. From the remaining 10 TF (GATA3, GFI1, HHEX, HOXC10, HOXC11, IRX4, LHX3, MSX1, PAX8 and RFX1), we highlight the following three for which their relevance in the context of breast cancer has been documented only recently: GATA3 with 160 TFBS and a RIF2 score of 3.28 inhibits breast cancer growth and metastasis (Dydensborg *et al*., 2009); LHX3 with 130 TFBS and a RIF1 score of −2.43 is an epigenetic biomarker for breast cancer (Dietrich *et al*., 2009); and PAX8 with 134 TFBS and a RIF1 score of −2.49 is a useful marker in distinguishing ovarian from mammary carcinomas (Nonaka *et al*., 2008).

3.8 RIF1 versus RIF2

As briefly mentioned earlier, we found no particular relationship between the two alternative measures of RIF in these datasets (Fig. 3), and this feature was attributed to both measures of RIF capturing different yet equally valuable features when ranking TF.

Numerically, the relationship between RIF1 and RIF2 can be explored from their expressions in Equations (3) and (4), respectively. Conditional on a given TF, the identity, abundance and differential expression of DE genes are fixed quantities. Hence, it suffices to explore RIF in the dynamic range of DW. Notably, DW^2 ranges from zero (case of identical co-expression correlation between the ith TF and the jth DE gene in both states; that is: $\mathrm{r1}_{ij} = \mathrm{r2}_{ij} = \mathrm{r}_{ij}$) to four (case of extreme ±1 and opposite $\mathrm{r1}_{ij}$ and $\mathrm{r2}_{ij}$). At $\mathrm{DW}^2 = 0$, then $\mathrm{RIF1}_i = 0$, while $\mathrm{RIF2}_i = 2 \times r_{ij}^2 \times \mathrm{PIF}_j$.

On the other extreme, at $\mathrm{DW}^2 = 4$, it follows that $\mathrm{RIF1}_i = 2 \times \mathrm{RIF2}_i = 4 \times \mathrm{PIF}_j$. Hence, the expectation is that RIF1 and RIF2 will produce similar ranking for a given TF when the selected set of DE genes are indeed targets of that TF and their expression is activated and/or inhibited in each experimental state by the TF under scrutiny. This also implies that, in situations where a TF is also a DE gene, RIF1 is unable to capture its relevance, while RIF2 assigns it a relevance in accordance to its own PIF.

In order to further explore this dichotomy, we selected two TF from the RIF analysis of the breast cancer data as representatives of extremely opposed RIF scores (Table 2): CARM1 (scoring highly negative according to RIF1, yet average according to RIF2) and SMARCA2 (highly positive by RIF2, yet average according to RIF1). Importantly, both TF are known to play a significant role in breast cancer (as discussed earlier) and have a moderate to high expression level, but neither is DE (i.e. while both could be easily detected they would not appear as relevant in an analysis based on expression only). Figure 6 shows the scatter plot of the co-expression relationship between CARM1 and SMARCA2 with the 561 target genes. Most co-expression correlations between CARM1 and the 561 target genes are above the diagonal, while most co-expression correlations involving SMARCA2 are below the diagonal. Notably, the target genes with an extreme DW (i.e. and hence away from the diagonal) with CARM1 have a near-zero DW with SMARCA2, and vice versa. This scenario was found for DE genes MSI1 and JAG2. Hence, it is DW driving the possible re-ranking between RIF1 and RIF2.

In an attempt to further illustrate how RIF1 captures those TF showing a large DW to those highly abundant highly DE genes, while RIF2 focuses on those TF showing evidence for a large change as predictors of the abundance of the DE genes in each condition, we selected four TF: the above-mentioned CARM1 and SMARCA2, as well as BAZ1B (very negative for RIF2 only) and NEUROG3 (very negative for both RIF1 and RIF2). For these four TF, Figure 7 illustrates the relationship between the PIF of each of the 561 target genes (either DE or differentially PIF genes) against DW^2 (left panels) and also against the change in predictive ability of the TF as measured by $(e1_j \times r1_{ij})^2 - (e1_j \times r1_{ij})^2$ (right panels). From the comparison of plots in Figure 7 with the RIF scores in Table 2, it becomes immediately apparent that while PIF drives the sign of RIF1, the sign of RIF2 is driven by the change, either positive or negative, in predictive ability of the TF. With most DE genes being down-regulated in our breast cancer contrast (Fig. 2D), all extreme RIF1 values are also negative. The magnitude of DW^2 dictates how extreme RIF1 values are likely to results. Using a nominal $\mathrm{DW}^2 > 0.7$ (i.e. the average of the y-axis in Fig. 7, left panels) it becomes apparent that, in terms of $|\mathrm{RIF1}|$, the ranking order is CARM1 > NEUROG3 > SMARCA2 > BAZ1B. On the

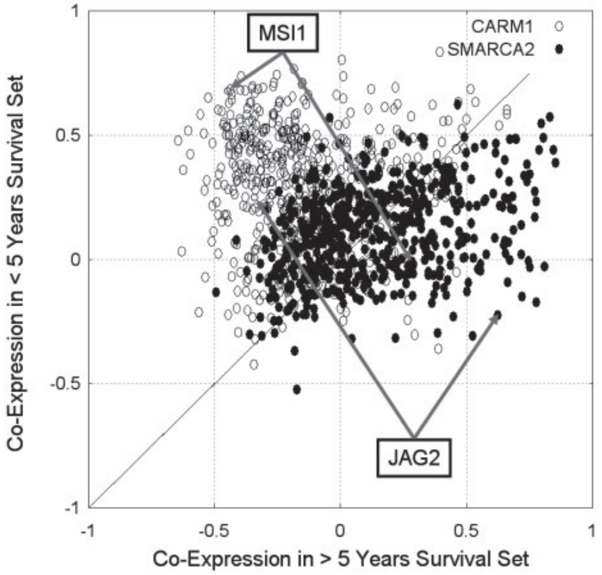

Fig. 6. Scatter plot of the co-expression relationships of CARM1 (open circles) and SMARCA2 (filled circles) with the 561 target genes (DE and/or differentially PIF genes) from the breast cancer dataset. These two TFs scored differently by either measure of RIF (Table 5). Also, the coordinates of two target genes (MSI1 and JAG2) with distinct and extreme differential wiring (DW) with either TF is highlighted by grey arrows.

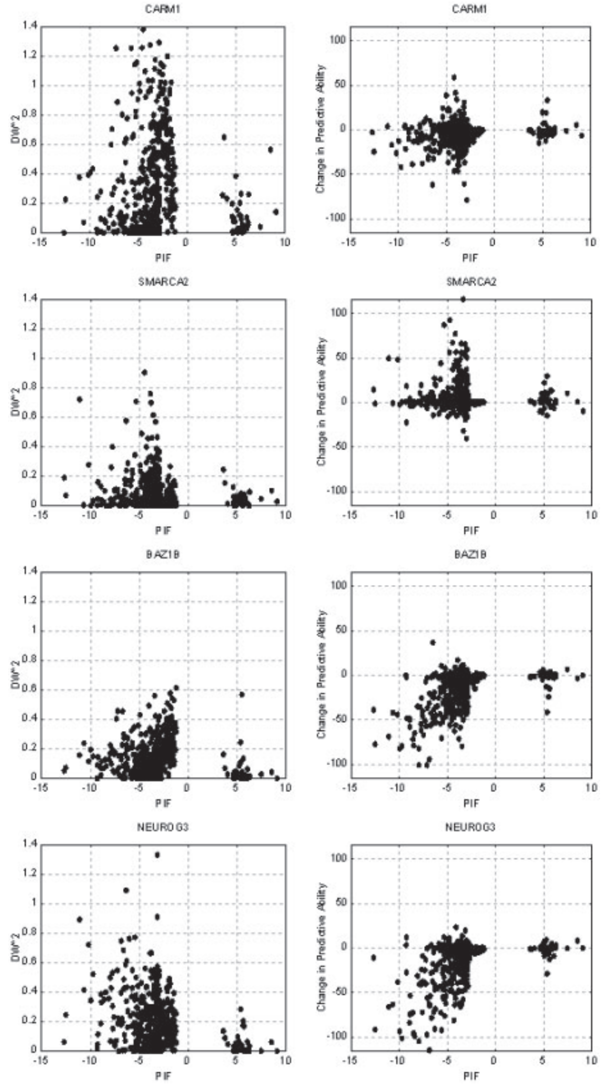

Fig. 7. For four key TF in the breast cancer dataset (top to bottom: CARM1, SMARCA2, BAZ1B and NEUROG3) with varying RIF scores according to RIF1 or RIF2 (Table 2), scatter plots showing the relationship between the PIF (x-axes) of each of the 561 target genes against DW^2 (left panels) and also against the change in predictive ability of the TF (right panels).

other hand, deviations from zero in predictive ability (i.e. zero being the average in the *y*-axis in Fig. 7, right panels) dictates both the sign and the magnitude of RIF2. CARM1, with most of the mass centered at zero, ranks poorly according to RIF2. SMARCA2, with most of the mass above zero, ranks highly positive; while BAZ1B and NEUROG3, with most of the mass below zero, rank highly negative.

4 CONCLUSIONS

In the last decade, the uptake of high-throughput gene expression microarray technology has been coupled with a substantial body of research devoted to the quantitative analysis of the resulting data, including issues of sequence annotation, platform sensitivity, transcriptome coverage, background correction, and normalization. As a result, large lists of DE genes have been reported and co-expression networks have been reversed-engineered. However, our understanding of the biological processes involved has not increased as much as might have been expected, especially in systems not studied in great detail by reductionist approaches.

The RIF algorithm was developed on a system where a single known mutation was largely responsible for a change in the phenotype and using data across a long-time course from 60 days post conception to 30 months of age. The dataset was derived from the same muscle type in the two breeds compared. In most respects, the differences between the expression of genes in the two datasets was very small. In this article, we have investigated the utility of the RIF algorithm from a range of differently structured datasets with increasing levels of diversity in origin and gene expression of the samples being compared. Importantly, the datasets ranged from a set of samples from the same tissue with different disease prognosis, through to a very complex comparison involving multiple tissues, from multiple breeds and both sexes. Although no single analysis can identify all the key TF involved in a process, it is clear that the combination of RIF1 and RIF2 identifies TF that are involved in key processes. In addition, the analysis also identifies the higher order drivers of the differences, although the success of this is likely to be dependent on some relevant a prior knowledge.

The three different analytical approaches (differential expression of TF, TF associated with DE genes, and TF that are differentially wired between the two datasets) can potentially identify distinct sets of genes with limited overlap. Since we do not know the true extent of the TF involved in the regulation of a complex trait under study, the observed differences cannot necessarily be attributed to high false positive or negative rates. Instead, the different approaches are identifying different sets of genes that may be involved in different parts of the process. One explanation for the lack of overlap between the first and the last two approaches is that DE analyses cannot by definition identify TF with activity modified in ways that do not involve a change in gene expression.

In this study, we have limited RIF analysis to identify key TF in two conditions or states only. In order to implement RIF to multi-condition arrangements, we could devise two possibilities: (i) compute RIF for every pair-wise condition and then apply a comparison of rankings for each TF (similar to meta-analysis strategies); and (ii) incorporate all pair-wise condition contrasts in the computation of RIF. However, we anticipate that this implementation to multi-condition experiments could result in the identification of TF that are minimally essential in each condition contrast and hence their use for understanding, and potentially manipulating, the design of complex phenotypes could be limited.

With the exception of the breast cancer dataset, we have applied RIF to experiments in which the number of replicates was balanced in each condition. In highly unbalanced designs, the co-expression correlations computed in each of the two conditions would have a vastly different standard errors associated with them. The condition with low replicates (or time points) would suffer from a large number of spurious correlations. In these cases, one should consider employing a significance analysis to only include those correlations that are deemed to be non-zero. This could be achieve by either using higher hard thresholds for the condition with the lower number of replicates, or soft weighted information-theory based threshold methods such as ARACNe (Margolin *et al*., 2008) or PCIT (Reverter and Chan, 2008).

It should also be noted that RIF is a function of expression data only. While this could be an advantage, one should not overlook that the quality of any expression-based metric is ultimately dependable on the quality of the original data, the processing algorithm to normalize it, and its effectiveness to account for systematic effects that can cause bias. Similarly, the analytical methods used here to detect DE and differentially PIF genes, while well-documented, assume a common residual variance for all genes. Because violations to this assumption could impact the outcome of RIF analyses, the relative advantage of joint versus gene-specific models should be considered a priori.

Finally, we note that the RIF algorithm is different to the MNI algorithm of di Bernardo *et al.* (2005), and recently applied by Ergün *et al.* (2007) to identify the mediator of prostate cancer, in that the MNI algorithm requires, as an initial phase, the re-construction of a gene regulatory network prior to applying its connectivity properties to a second (testing) dataset to identify the key regulators. Instead, RIF operates directly on the data at hand without having to rely in the availability, and subsequent processing, of existing data of similar, or preferably more comprehensive, biological characteristics than the one under scrutiny.

In conclusion, the RIF analysis appears to be a robust and valuable methodology to identify the regulators with the highest evidence of contributing to differential expression in two biological conditions, it shows potential to be applied to a wide range of gene expression data sets, and to significantly increase the biological knowledge that can be derived from such experiments.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Dr Aaron Ingham and Dr William Barendse for insightful suggestions in the preparation of this manuscript.

Conflict of Interest: none declared.

REFERENCES

Bochar,D.A. *et al*. (2000) BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell*, **102**, 257–265.

Budhram-Mahadeo,V. *et al*. (2001) The closely related POU family transcription factors Brn-3a and Brn-3b are expressed in distinct cell types in the testis. *Int. J. Biochem. Cell. Biol*., **33**, 1027–1039.

Cartharius,K. *et al.* (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.

Cheng,C. *et al.* (2009) Systematic identification of transcription factors associated with patient survival in cancers. *BMC Genomics*, **10**, 225.

Cowley,M.J. *et al.* (2009) Intra- and inter-individual genetic differences in gene expression. *Mamm. Genome*, **20**, 281–295.

Dessain,S. *et al.* (1992) Antp-type homeodomains have distinct DNA binding specificities that correlate with their different regulatory functions in embryos. *EMBO J.*, **11**, 991–1002.

di Bernardo,D. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, **23**, 377–383.

Dietrich,D. *et al.* (2009) Analysis of DNA methylation of multiple genes in microdissected cells from formalin-fixed and paraffin-embedded tissues. *J. Histochem. Cytochem.*, **57**, 477–489.

Dydensborg,A.B. *et al.* (2009) GATA3 inhibits breast cancer growth and pulmonary breast cancer metastasis. *Oncogene*, **28**, 2634–2642.

Ergün,A. *et al.* (2007) A network biology approach to prostate cancer. *Mol. Syst. Biol.*, **3**, 82.

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Frietze,S. *et al.* (2008) CARM1 regulates estrogen-stimulated breast cancer growth through up-regulation of E2F1. *Cancer Res.*, **68**, 301–306.

García-Ortiz,J.E. *et al.* (2009) Foxl2 functions in sex determination and histogenesis throughout mouse ovary development. *BMC Dev. Biol.*, **9**, 36.

Gardner,T.S. *et al.* (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Hecker,M. *et al.* (2009) Integrative modeling of transcriptional regulation in response to antirheumatic therapy. *BMC Bioinformatics*, **10**, 262.

Hong,L. *et al.* (2009) Steroid regulation of proliferation and osteogenic differentiation of bone marrow stromal cells: a gender difference. *J. Steroid. Biochem. Mol. Biol.*, **114**, 180–185.

Huang,W. *et al.* (2009) Expression pattern, cellular localization and promoter activity analysis of ovarian aromatase (Cyp19a1a) in protogynous hermaphrodite red-spotted grouper. *Mol. Cell. Endocrinol.*, **307**, 224–236.

Hudson,N.J. *et al.* (2009a) Inferring the transcriptional landscape of bovine skeletal muscle by integrating co-expression networks. *PLoS ONE*, **4**, e7249.

Hudson,N.J. *et al.* (2009b) A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.*, **5**, e1000382.

Jorgensen,J,S. and Gao,L. (2005) Irx3 is differentially up-regulated in female gonads during sex determination. *Gene Expr. Patterns*, **5**, 756–762.

Joshi,A. *et al.* (2009) Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, **25**, 490–496.

Kerhornou,A. and Guigó,R. (2007) BioMoby web services to support clustering of co-regulated genes based on similarity of promoter configurations. *Bioinformatics* **23**, 1831–1833.

Lan,K.C. *et al.* (2008) Expression of androgen receptor co-regulators in the testes of men with azoospermia. *Fertil. Steril.*, **89**, 1397–1405.

Le Dily,F. *et al.* (2008) COUP-TFI modulates estrogen signaling and influences proliferation, survival and migration of breast cancer cells. *Breast Cancer Res. Treat.*, **110**, 69-83.

Margolin,A.A. *et al.* (2008) ARANCE: an algorithm for the reconstruction of gene regulatory networks in mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

McLachlan,G.J. *et al.* (2002) A mixture model-based approach to the clustering of microarray data. *Bioinformatics*, **18**, 413–422.

McLachlan,G.J. *et al.* (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

Nagaraj,S.H. *et al.* (2008) Promoter sequence analysis of differentially expressed genes in sheep following a nematode parasite resistance challenge. *19th International Conference on Genome Informatics* (GIW2008), Gold Coast, Australia, 1–3 December 2008. Available at http://mlaa.com.au/giw2008/PDFposter/giw2008poster_submission_69.PDF

Nechad,M. *et al.* (1994). Production of nerve growth factor by brown fat in culture: relation with the in vivo developmental stage of the tissue. *Comp. Biochem. Physiol. Comp. Physiol.*, **107**, 381–388.

Neil,J.R. *et al.* (2009) X-linked inhibitor of apoptosis protein and its E3 ligase activity promote transforming growth factor-β-mediated nuclear factor-κB activation during breast cancer progression. *J. Biol. Chem.*, **284**, 21209–21217.

Nonaka,D. *et al.* (2008) Expression of pax8 as a useful marker in distinguishing ovarian carcinomas from mammary carcinomas. *Am. J. Surg. Pathol.*, **32**, 1566–1571.

Pérez-Enciso,M. *et al.* (2009) Impact of breed and sex on porcine endocrine transcriptome: a bayesian biometrical analysis. *BMC Genomics*, **10**, 89.

Reisman,D. *et al.* (2009) The SWI/SNF complex and cancer. *Oncogene*, **28**, 1653–1668.

Reverter,A. and Chan,E.K.F. (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, **24**, 2491–2497.

Reverter,A. *et al.* (2006a) A gene co-expression network for bovine skeletal muscle inferred from microarray data. *Physiol. Gen.*, **28**, 76–83.

Reverter,A. *et al.* (2006b) Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics*, **22**, 2396–2404.

Scharer,C.D. *et al.* (2009) Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer Res.*, **69**, 709–717.

Scholz,S. *et al.* (2003) Hormonal induction and stability of monosex populations in the medaka (Oryzias latipes): expression of sex-specific marker genes. *Biol. Reprod.*, **69**, 673–678.

Scime,A. *et al.* (2005). Rb and p107 regulate preadipocyte differentiation into white versus brown fat through repression of PGC-1alpha. *Cell Metab.*, **2**, 283–295.

Sheu,Y.T. *et al.* (2006) Nuclear receptor coactivator-3 alleles are associated with serum bioavailable testosterone, insulin-like growth factor-1, and vertebral bone mass in men. *J. Clin. Endocrinol. Metab.*, **91**, 307–312.

Shur,I. *et al.* (2006) Dynamic interactions of chromatin-related mesenchymal modulator, a chromodomain helicase-DNA-binding protein, with promoters in osteoprogenitors. *Stem Cells*, **24**, 1288–1293.

Skaug,B. *et al.* (2009) The role of ubiquitin in NF-kappaB regulatory pathways. *Annu. Rev. Biochem.*, **78**, 769–796.

Small,C.L. *et al.* (2004) Profiling gene expression during the differentiation and development of the murine embryonic gonad. *Biol. Reprod.*, **72**, 492–501.

Takada,S. *et al.* (2009) Potential role of miR-29b in modulation of Dnmt3a and Dnmt3b expression in primordial germ cells of female mouse embryos. *RNA*, **15**, 1507–1514.

Tetzlaff,S. *et al.* (2009) Association of parathyroid hormone-like hormone (PTHLH) and its receptor (PTHR1) with the number of functional and inverted teats in pigs. *J. Anim. Breed. Genet.*, **126**, 237–241.

Timmons,J.A. *et al.* (2007) Myogenic gene expression signature establishes that brown and white adipocytes originate from distinct cell lineages. *Proc. Natl Acad. Sci. USA*, **104**, 4401–4406.

Tsuchida,A. *et al.* (2005) Nuclear receptors as targets for drug development: molecular mechanisms for regulation of obesity and insulin resistance by peroxisome proliferator-activated receptor gamma, CREB-binding protein, and adiponectin. *J. Pharmacol. Sci.*, **97**, 164–170.

Uhlén,M. *et al.* (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics*, **4**, 1920–1932.

van't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 484–485.

Vaquerizas,J.M. *et al.* (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

Wang,X. *et al.* (2008) MAX drives tumor-specific expression of PPAR gamma 1 in breast cancer cells. *Breast Cancer Res. Treat.*, **111**, 103–111.

Watson-Haigh,N.S. *et al.* (2010) PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics*, **26**, 411–413.

Wilhelm,D. *et al.* (2007) Sex determination and gonadal development in mammals. *Physiol. Rev.*, **87**, 1–28.

Wolfe,C.J. *et al.* (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, **6**, 227.

Zhang,H. *et al.* (2007) Tissue type-specific modulation of ER transcriptional activity by NFAT3. *Biochem. Biophys. Res. Commun.*, **353**, 576–581.