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

Antonio Reverter1,∗, Nicholas J. Hudson1, Shivashankar H. Nagaraj1, Miguel Pérez-Enciso2 and Brian P. Dalrymple1

1Bioinformatics Group, CSIRO Livestock Industries, Queensland Bioscience Precinct, 306 Carmody Road, St. Lucia, Brisbane, Queensland 4067, Australia and 2CREA—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 (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.
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 whether TF were selected from the genes set provided by Vaqueras 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 (Vaqueras 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 gene 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.

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 Perez-Enciso et al. (2009) who used 80 Porcine Affymetrix chips (each representing ~15 000 genes) to survey the gene expression profile in 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 1575 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:

\[ y_{ijk} = H_i + G_j + GC_{ijk} + e_{ijk}, \]

where \( y_{ijk} \) is the vector of expression readings from the \( n \)th hybridization chip, on the \( j \)th gene at the \( k \)th condition; \( H_i \) is the fixed effect of the \( n \)th 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 \( j \)th gene; \( GC_{ijk} \) is random interaction between the
jth gene and the ith experimental condition and it captures differences from overall averages that are attributable to specific gene-condition combination; and \( q_{ij} \) is the random residual error associated with \( y_{ij} \).

Using standard statistical assumptions in mixed model theory, the effects of \( G_1 \), \( G_2 \) and \( e_0 \) 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.tux.dal.ca/pub/vce6/vce6/) . The solutions to the \( G_2 \alpha \) effect was 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.

\[
 f(d, \Phi) = \theta_0 \phi(d; \mu_0, \sigma_0^2) + \theta_1 \phi(d; \mu_1, \sigma_1^2),
\]

where \( d \) denotes the vector of DE measures for all the genes, and the two components in the mixures correspond to \( \Phi(\cdot) \) for the empirical null normal density with mean \( \mu_0 \) (not necessarily zero) and variance \( \sigma_0^2 \) (not necessarily one), encapsulating the non-DE genes; and \( \Phi(\cdot) \) for the non-null distribution corresponding to DE genes. Finally, the mixing proportions \( \theta_0 \) and \( \theta_1 \) 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:

\[
\text{RIF}_1 = \frac{\sum_{i=1}^{N} \hat{\alpha}_i \hat{\delta}_i \times \hat{\delta}_i \times \text{DW}^2_i}{\sum_{i=1}^{N} \text{PIF}_i \times \text{DW}^2_i},
\]

(3)

and

\[
\text{RIF}_2 = \frac{1}{\sum_{i=1}^{N} \left[ \left( \hat{\delta}_i \times \hat{\delta}_i \times \hat{\delta}_i \right) - \left( e_2 \times e_2 \times e_2 \right) \right]}.
\]

(4)

where \( N \) is the number of DE genes, \( \hat{\alpha}_i \) is the estimated average expression of the jth DE gene, averaged across the two conditions being contrasted, \( \hat{\delta}_i \) is the estimated differential expression of the jth DE gene; and DW is the differential wiring between the ith TF and the jth DE gene, and computed from the difference between \( e_1 \) and \( e_2 \); the co-expression correlation between the ith TF and the jth DE gene in conditions 1 and 2, respectively (McLachlan et al., 2009b):

\[
\text{DW}_i = r_{e_1} - r_{e_2},
\]

(5)

The expression for RIF1 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:

\[
\text{PIF}_i = \hat{\delta}_i \times \hat{\delta}_i = \frac{1}{2} \left( e_1 \times e_2 \right) \left( e_1 \times e_2 \right) - \frac{1}{2} \left( e_1 \times e_1 \right) \left( e_1 \times e_1 \right),
\]

(6)

where \( e_1 \) and \( e_2 \) represent the expression of the jth DE gene in conditions 1 and 2, respectively. The definition of PIF in Equation (6) as the difference of squared expression allows for the alternative parameterization of RIF presented by RIF2, 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>
<thead>
<tr>
<th>Dataset</th>
<th>Parameters of the mixture ( \pi_0, \pi_1, \sigma_0^2, \sigma_1^2 )</th>
<th>DE survival vs breast cancer.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \pi_0 = 0.96 ), ( \pi_1 = 0.04 ) ( \sigma_0^2 = 0.25 ), ( \sigma_1^2 = 0.75 )</td>
<td>0.96 vs 0.04, 0.25 vs 0.75</td>
</tr>
<tr>
<td>2</td>
<td>( \pi_0 = 0.99 ), ( \pi_1 = 0.01 ) ( \sigma_0^2 = 0.15 ), ( \sigma_1^2 = 0.85 )</td>
<td>0.99 vs 0.01, 0.15 vs 0.85</td>
</tr>
<tr>
<td>3</td>
<td>( \pi_0 = 0.80 ), ( \pi_1 = 0.20 ) ( \sigma_0^2 = 0.30 ), ( \sigma_1^2 = 0.70 )</td>
<td>0.80 vs 0.20, 0.30 vs 0.70</td>
</tr>
<tr>
<td>4</td>
<td>( \pi_0 = 0.75 ), ( \pi_1 = 0.25 ) ( \sigma_0^2 = 0.40 ), ( \sigma_1^2 = 0.60 )</td>
<td>0.75 vs 0.25, 0.40 vs 0.60</td>
</tr>
</tbody>
</table>

* Data 1 = Brown vs White adipocytes; Data 2 = Testes vs Ovaries embryos; Data 3 = Males vs Females pigs; Data 4 = more than 5 years vs less than 5 years survival to breast cancer.
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, 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).

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 and differential PIF
genes only and DE and differentially PIF genes, respectively.

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

<table>
<thead>
<tr>
<th>TF</th>
<th>A</th>
<th>DE</th>
<th>RIF1</th>
<th>RIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CREBBP</td>
<td>3.22</td>
<td>−2.47</td>
<td>−3.93</td>
<td>−0.43</td>
</tr>
<tr>
<td>CUTL1</td>
<td>7.72</td>
<td>−0.80</td>
<td>−0.86</td>
<td>−2.30</td>
</tr>
<tr>
<td>MYOG</td>
<td>5.28</td>
<td>2.32</td>
<td>−0.25</td>
<td>2.31</td>
</tr>
<tr>
<td>PPARBP</td>
<td>3.95</td>
<td>−0.59</td>
<td>−3.14</td>
<td>−0.22</td>
</tr>
<tr>
<td>RBL1</td>
<td>5.06</td>
<td>1.28</td>
<td>−0.16</td>
<td>2.31</td>
</tr>
</tbody>
</table>

(1) Brown versus white adipocyte differentiation

(2) Testes versus ovaries embryonic differentiation

(3) Male versus female pre-pubertal pigs

(4) Breast cancer: >5 years versus <5 years survival

<table>
<thead>
<tr>
<th>TF</th>
<th>A</th>
<th>DE</th>
<th>RIF1</th>
<th>RIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD9</td>
<td>12.12</td>
<td>0.02</td>
<td>3.90</td>
<td>1.92</td>
</tr>
<tr>
<td>IRX3</td>
<td>5.47</td>
<td>−0.06</td>
<td>2.17</td>
<td>3.93</td>
</tr>
<tr>
<td>RNF14</td>
<td>9.72</td>
<td>0.36</td>
<td>4.40</td>
<td>1.89</td>
</tr>
<tr>
<td>SOX5</td>
<td>7.81</td>
<td>−0.37</td>
<td>3.10</td>
<td>1.20</td>
</tr>
<tr>
<td>TAF7L</td>
<td>5.34</td>
<td>1.60</td>
<td>−0.37</td>
<td>3.42</td>
</tr>
<tr>
<td>ZNF281</td>
<td>7.97</td>
<td>0.14</td>
<td>−1.99</td>
<td>2.94</td>
</tr>
</tbody>
</table>

(5) Male versus female pre-pubertal pigs

(6) Breast cancer: >5 years versus <5 years survival

<table>
<thead>
<tr>
<th>TF</th>
<th>A</th>
<th>DE</th>
<th>RIF1</th>
<th>RIF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL1</td>
<td>7.22</td>
<td>0.20</td>
<td>−3.28</td>
<td>0.78</td>
</tr>
<tr>
<td>CARM1</td>
<td>10.20</td>
<td>0.01</td>
<td>−4.28</td>
<td>−0.31</td>
</tr>
<tr>
<td>MAZ</td>
<td>13.17</td>
<td>−0.57</td>
<td>−1.27</td>
<td>−2.68</td>
</tr>
<tr>
<td>NFATC4</td>
<td>8.58</td>
<td>−0.28</td>
<td>−6.37</td>
<td>−1.48</td>
</tr>
<tr>
<td>NR2F1</td>
<td>8.58</td>
<td>0.09</td>
<td>−3.32</td>
<td>−0.19</td>
</tr>
<tr>
<td>PITX3</td>
<td>10.76</td>
<td>−0.40</td>
<td>−2.45</td>
<td>−2.99</td>
</tr>
<tr>
<td>RELA</td>
<td>9.14</td>
<td>0.11</td>
<td>−3.67</td>
<td>0.07</td>
</tr>
<tr>
<td>SMARCA2</td>
<td>8.65</td>
<td>−0.04</td>
<td>0.41</td>
<td>2.97</td>
</tr>
<tr>
<td>SMARCA4</td>
<td>11.42</td>
<td>−0.23</td>
<td>0.67</td>
<td>3.22</td>
</tr>
</tbody>
</table>
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 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.
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 TAF7 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, PFFX3 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 chromatin remodelers (e.g. SMARC2 and SMARCA4 in our case) has been well described (see for instance the recent review by Reisman et al., 2009); and in particular, SMARC4 has been shown to interact with BRCA1 providing links between chromatin remodeling and breast cancer (Bochar et al., 2003); (iii) MYC-associated zinc finger protein (MAZ) is responsible for the high expression of PPARG in breast cancer (Wang et al., 2008); (iv) the nuclear receptor NR2FI interacts with estrogen receptor and regulates the expression of estradiol influencing the proliferation of breast cancer cells (Le Dity 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 (Neal 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 correlation and 20 with negative correlation to DE genes) whereas our analyses revealed 71 TF with significant RIF scores. We compared our list of TF with those found 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 closely matched those found recently by using PPI as a criterion (see for instance Cheng et al., 2009). In fact, among the 892 TF included in the analyses, 71 had a RIF z-scores $<-2 < 2$. 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 MalInspector tool (Catharius 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 $<-2 < 2$ 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, HHIEX, HOXC10, HOXC11, IRX4, LHX3, MSX1, PAX8 and RXF1), we highlight the following three for which their relevance in the context of breast cancer has been documented recently: GATA3 with 160 TFBS and a RIF2 score of 3.28 inhibits breast cancer growth and metastasis (Dyedensburg 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: $r_{ij} = r_{ij} = r_{ij}$) to four (case of extreme $d_1$ and opposite $d_2$). At DW$^2 = 0$, then RIF1$ = 0$, while RIF2$ = 2 \times 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 (x-axes) of each of the 561 target genes against DW2 (left panels) and also against the change in predictive ability of the TF (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 DW2 dictates how extreme RIF1 values are likely to results. Using a nominal DW2 > 0.7 (i.e. the average of the y-axis in Fig. 7, left panels) it becomes apparent that, in terms of [RIF1], the ranking order is CARM1 > NEUROG3 > SMARCA2 > BAZ1B. On the
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 RIF. CARM1, with most of the mass centered at zero, ranks poorly according to RIF. 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 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 prognoses, 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 priori 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.

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


The closely related PIWI family transcription factors Bm-sb and Bm-3b are expressed in distinct cell types in the testis. Int. J. Biochem. Cell. Biol., 33, 1027–1039.
Hecker, M. et al.

Gardner, T. S.

García-Ortiz, J. E.

Friedmann, N.

Ergün, A.

et al.

Nagaraj, S. H.

Cheng, C.

A. Reverter et al.

Cowley, M. J.

Dessain, S.

et al.

1608–1615.

Bioinformatics

regulated genes based on similarity of promoter configurations.

men with azoospermia.

prioritization of model predictions.

of action via expression profiling.

breast cancer metastasis.

5

J. Comput. Biol

spotted grouper.

analysis of ovarian aromatase (Cyp19a1a) in protogynous hermaphrodite red-

International Conference on Genome Informatics

reverse-engineered gene networks.

to antirheumatic therapy.

(2005) Inferring genetic networks and identifying compound mode

of action via expression profiling.

Science

301–103.5

5

3

J. Comput. Biol

Bioinformatics

specificities that correlate with their different regulatory functions in embryos.

reverse-engineered gene networks.

5

3


97

164–170.


1507–1514.

Wang, X.

et al.

Vaquerizas, J. M.

et al.

Wang, X.

et al.

Uhlén, M.

et al.

Neil, J. R.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.