Published online 24 November 2011

Nucleic Acids Research, 2012, Vol. 40, No. 6 **2377–2398**
doi:10.1093/nar/gkr902

Gene network inference and visualization tools for biologists: application to new human transcriptome datasets

Daniel Hurley[1,2,3], Hiromitsu Araki[2,4], Yoshinori Tamada[4,5], Ben Dunmore[4,6], Deborah Sanders[4,6], Sally Humphreys[4,6], Muna Affara[6], Seiya Imoto[5], Kaori Yasuda[4,7], Yuki Tomiyasu[4,7], Kosuke Tashiro[7], Christopher Savoie[4], Vicky Cho[3], Stephen Smith[6], Satoru Kuhara[7], Satoru Miyano[5], D. Stephen Charnock-Jones[6,8,*], Edmund J. Crampin[1,9,*] and Cristin G. Print[2,3,*]

[1]Auckland Bioengineering Institute, [2]Department of Molecular Medicine and Pathology, School of Medical Sciences, Faculty of Medical and Health Sciences, [3]Bioinformatics Institute, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand, [4]GNI Ltd, Shinjuku Park Tower N, 30th Floor, 3-7-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 163-1030, [5]Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan, [6]Departments of Pathology and Obstetrics & Gynaecology, The Rosie Hospital, The University of Cambridge, Cambridge CB2 0SW, UK, [7]Department of Molecular Biosciences, Faculty of Agriculture, Kyushu University, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan, [8]National Institute for Health Research, Cambridge Comprehensive Biomedical Centre, UK and [9]Department of Engineering Science, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Received August 15, 2011; Revised October 2, 2011; Accepted October 5, 2011

ABSTRACT

Gene regulatory networks inferred from RNA abundance data have generated significant interest, but despite this, gene network approaches are used infrequently and often require input from bioinformaticians. We have assembled a suite of tools for analysing regulatory networks, and we illustrate their use with microarray datasets generated in human endothelial cells. We infer a range of regulatory networks, and based on this analysis discuss the strengths and limitations of network inference from RNA abundance data. We welcome contact from researchers interested in using our inference and visualization tools to answer biological questions.

INTRODUCTION

Traditional methods for analysing transcriptome data [for instance, clustering algorithms (1), principal component analysis (2) and the use of linear models to detect differential expression (3)] have made a significant contribution to biology; for instance, they have identified RNA transcripts that are regulated by drugs and during development, and they have provided clinically useful tumour classifications. However, these methods do not effectively identify how thousands of different RNAs in a cell operate synergistically in pathways and networks.

Gene regulatory networks attempt to address this issue. These can be described as circuit diagrams showing putative co-expression and in some cases directional cause-and-effect relationships between RNAs. Gene regulatory networks can be constructed using any type of transcriptome data, such as data gathered from microarray or RNAseq experiments. In a gene regulatory network, RNA transcripts are represented as nodes in a graph, each node corresponding to one or more RNAs. Links between nodes are represented as edges on the graph, which indicate putative relationships between RNAs, where the abundance of one RNA can affect the abundance of a second RNA. This regulation can be simple (e.g. *RNA A* encodes a transcription factor protein that promotes the transcription of *RNA B*) or complex (e.g. through multiple molecular steps involving protein signalling cascades or

*To whom correspondence should be addressed. Email: dscj1@cam.ac.uk
Correspondence may also be addressed to Edmund J. Crampin. Email: e.crampin@auckland.ac.nz
Correspondence may also be addressed to Cristin G. Print. Email: c.print@auckland.ac.nz

metabolites). Therefore, although proteins and metabolites are not explicitly shown in gene regulatory network graphs, they may contribute to the functional relationships encapsulated by the edges. Some inference methods infer 'directed' networks, in which a putative causal influence of one RNA upon the abundance of another is modelled, while other methods are 'undirected' and do not specify a direction of interaction. The number of published approaches to gene network inference has grown quickly in the last 5 years to encompass many sophisticated approaches (4,5), and gene regulatory networks have contributed to significant biological findings in several species ranging from simple organisms [for example, *Escherichia coli* (6–8), *Salmonella enterica* (9) and *Halobacterium salinarium* (10)] to humans (11,12). However, gene regulatory network inference currently faces several barriers to adoption as a technique commonly used by experimentally focused researchers in biology and medicine.

The first barrier to the common use of gene regulatory networks is related to limitations of the available data. Traditional systems identification techniques assume that the number of variables in a system under investigation is considerably fewer than the number of 'observations' or measurements of those variables (13). Many of the 'simulated' datasets used for benchmarking gene network inference approaches have no more variables than observations; examples of this issue include many of the simulated datasets produced for the DREAM gene network inference competition (14,15), and the small number of experimental datasets commonly used for benchmarking network inference algorithms [e.g. the SOS pathway knockdown dataset (16) and the *Drosophila melanogaster* developmental timecourse data from the FlyEx database (17)].

However, due to financial and experimental constraints, many of the transcriptome datasets to which biologists would like to apply gene network analysis have many more variables than observations. One solution has been to increase the number of observations by assembling 'compendium' datasets made up of a variety of smaller datasets generated from cells or tissues in different states. Basso *et al.* (11) successfully used this approach in an investigation of c-Myc relationships from a compendium of healthy and malignant human B-cell microarray data. However, format incompatibilities and experimental differences can make it hard to combine data from different microarray platforms and studies, and the reproducibility of microarray results from different laboratories has been problematic in the past (18). More fundamentally, if compendium datasets are derived from a large and varied population of cells in a mixture of different states, then a 'regression to the mean' effect may take place, in which so many different influences are active across the cell population that those influences only active at certain times, or in certain cellular states, are not represented strongly enough to be significant in the network model. This explanation was proposed for the poor performance of the entries in the DREAM2 genome-scale gene network challenge (19). To respond to the lack of appropriately dimensioned datasets that places a barrier to gene network use, in this article, we publish a large new dataset containing microarray data for ~20 000 variables (probes) and 400 observations (400 separate targeted siRNA disruptions in a single cell type—primary endothelial cells). We hope that the large number of observations in this dataset will allow researchers to analyse several hundred RNAs at a time for the rigorous assessment and optimization of gene network inference methods, and for assessment of the sensitivity of these methods to data dimensionality. We also hope this dataset will reveal new insights into endothelial cell/vascular biology and pathology.

A second barrier to adoption by biologists of gene network inference as a common technique is the degree of programming skill that is required to use network inference approaches. Currently, the way in which inference methods are implemented varies significantly from method to method and between research groups. Different algorithms, developed using different programming languages, require different formats for input data and are operated using different commands. Table 1 shows some examples of the broad range of gene network inference algorithms available, and compares the technologies they use. Researchers wanting to use these inference algorithms must be proficient in the languages used to implement each one, must reformat their data separately for each one, and must rearrange the networks inferred by each one into a common format if they wish to compare them. We have found no published algorithms that take a range of datasets as input and infer networks from each one. In an attempt to address this gap, in this article, we report a software framework to simplify the use of a range of common network inference algorithms by experimentally-focused researchers and bioinformaticians without requiring specialist programming knowledge.

A third barrier to adoption by biologists of gene network inference as a common technique is related to the complexity of evaluating different approaches to network inference. For a novel network inference method to be deemed successful, it must at least demonstrate a biologically meaningful or statistically significant result in terms of some evaluation technique. Table 2 shows examples of commonly used methods for evaluation of gene networks. Researchers frequently use the same 'inference methods' as other groups, but they very rarely use the same 'evaluation techniques'. Although researchers usually make the data sets used and the source code for their inference method available, tools for evaluation of networks are only rarely included in the packages released for each algorithm. Therefore, replicating previously published evaluations can be very difficult. To make progress here, a set of separate evaluation tools is required that can be applied in the same manner across a variety of network inference methods with minimal effort. Responding to this problem, we have assembled a suite of methods for

Table 1. Examples of gene network inference algorithms

Algorithm type	Examples	Gene network type	Technologies	Input/output formats
Mutual information	ARACNE (11), CLR (20)	Non-linear relationships, undirected edges	Compiled C/C++ executable, MATLAB scripts	Plain-text, input expression data in .EXP format, output network in .ADJ format (ARACNE)
Bayesian inference	BANJO (21), SiGN-BN (22–24)	Non-linear relationships, directed edges	Java, C	Plain-text, input expression data in unnamed tab-delimited format, output network embedded in a report (BANJO), CSML markup format (SiGN-BN)
Correlation	Pearson product-moment correlation, Spearman rank correlation, implemented in most scientific languages or platforms	Linear relationships, undirected edges	Various	Various
Dynamical systems	NIR (16), MIKANA (25,26), TSNI (27)	Non-linear relationships, directed edges	MATLAB	MATLAB data structures (NIR)

Table 2. Common evaluation techniques for network inference

Evaluation method	Description	Element used	Examples
Overlap with reference networks	Create a reference network from literature, KEGG, experimental data or simulated data. Calculate precision, area under an ROC or PRC plot, etc.	Relationships	DREAM conference assessment, many competitive studies (28–33)
Annotation enrichment	Look for annotations in common for children of a hub or hubs (e.g. GO path, TF binding site)	Genes	(34)
Clinical outcome prediction	Using a clinical dataset, show that features are meaningful predictors of outcome (e.g. survival in Kaplan-Meier analysis or other clinical metric (e.g. tumour grade)	Relationships or Genes	(35)
Functional prediction	Using some functional data, show that features predict functional behaviour (e.g. apoptosis, cell-cycle changes)	Relationships	(36,37)

evaluating any kinds of gene regulatory network, which has been integrated into a single unified software framework.

In general bioinformatics, workflow engines such as GenePattern (38), Taverna (39), Galaxy (40) and geWorkbench (41) are under active development to make bioinformatics functions available to researchers who lack specialist programming knowledge, and to make standard analyses easily repeatable and reliable. These platforms are rich in functions for general bioinformatics, but gene network inference algorithms and analysis methods are not well-represented in them.

In this article, we report the assembly of gene network inference and evaluation methods described earlier as a set of modules (Figure 1) in the GenePattern biologist-focused bioinformatics web environment. These modules have been submitted to the Broad Institute GenePattern module repository, and can alternatively be used in the MATLAB environment if preferred. In subsequent sections, we illustrate the use of these tools on gene networks inferred from our endothelial cell microarray datasets.

MATERIALS AND METHODS

Creating the siRNA disruptant dataset

Cell culture, siRNA transfection and tumour necrosis factor treatment. Umbilical cords were collected after written informed consent was given and approval of the study received from the Cambridge Research Ethics Committee. Human umbilical vein endothelial cells (HUVECs) were isolated from umbilical cords by collagenase digestion and cultured at 37°C/5% CO_2 in basal culture medium supplemented with a proprietary mixture of heparin, hydrocortisone, epidermal growth factor, fibroblast growth factor, 2% fetal calf serum (FCS) (EGM-2, Cambrex, Workingham, UK). Equal numbers of HUVECs from 10 individuals were pooled, plated at 2.5×10^5 cells/well in six-well plates and

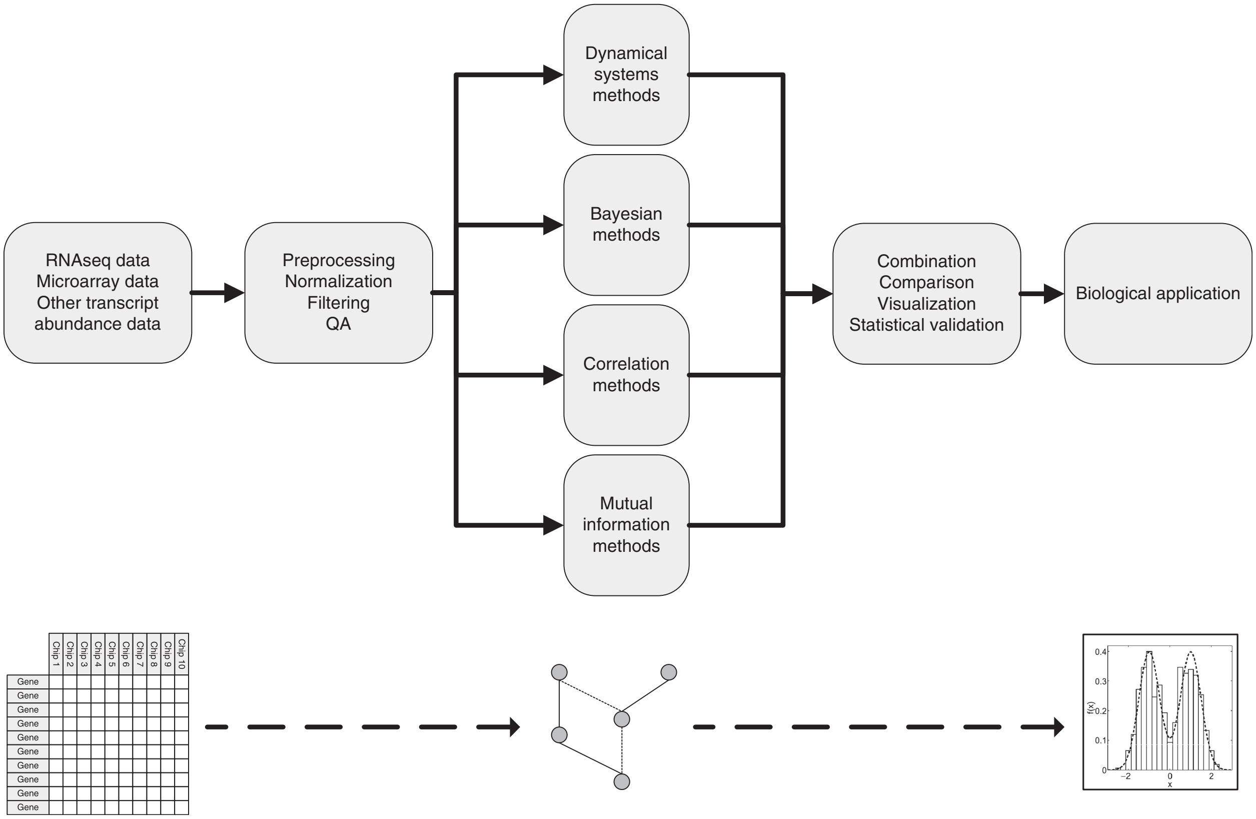

Figure 1. Schematic of the network inference framework. From left to right, transcriptome data are passed through pre-processing and normalization functions, then used as input for the range of network inference algorithms described in Table 1. Networks inferred by each of the methods are output in a standard format, in which they can be compared against each other and against literature relationships using methods described in Table 2. Finally, conclusions from the network comparison and analyses are used to inform experimental decisions.

allowed to recover for 24 h, at which time they were ~70% confluent.

To support the choice of pooling 10 biological isolates to minimize individual variation in our dataset, we carried out an in silico pooling experiment. Using microarray expression data from 15 different HUVEC isolates, we assessed how individual variance in the dataset decreased when the mean expression value for an increasing number of isolates was calculated. Mean expression values were calculated for all combinations of the 15 different HUVEC isolates on a gene by gene basis. The variance between all combinations of the mean values was then calculated and the ratios between all variance values above an arbitrarily set minimum threshold value calculated for all combinations of the mean. The mean value of the variance ratios for each 'meaned' group of isolates was then plotted against the mean number of isolates (Supplementary File S1). It can be seen from the curve that even when a more stringent variance threshold of 1.5 is used, increasing the number of isolates pooled beyond 10 individuals will not result in an additional marked decrease in the individual variability in the dataset.

Four hundred transcription factors, signalling molecules, receptors and ligands were selected as siRNA targets based on their relevance to endothelial biology and pathology. siRNA 'smartpools' from Dharmacon Inc. (Lafayette, CO, USA) were transfected into the cells using the siFectamine transfection reagent (ICVEC, London, UK) used according to the manufacturer's instructions.

To generate a time course dataset related to inflammatory conditions, pools of 70% confluent HUVECs isolated and cultured as above were treated for 24 h with 10 ng/ml tumour necrosis factor (TNF-α).

RNA preparation and gene array analysis. Total RNA was prepared using Trizol reagent (Invitrogen, London, UK) and assessed using an Agilent 2100 bioanalyser. Biotin-labelled complex cRNAs were prepared and hybridized to CodeLink UniSet Human 20K Bioarray microarrays according to the manufacturer's protocols (GE Healthcare, Amersham, UK). The quality of the expression data from all chips was confirmed using CodeLink Expression Analysis Software (version 4.1). To ensure that expression levels were comparable between the arrays the data was normalized using the cyclic Loess method (42).

Data processing and descriptive statistics. siRNA gene array data were $\log_2$-transformed and log ratios between each observation (siRNA knockdown treatment) and the median of all 400 arrays were calculated on a gene-by-gene basis. To quantify the effects of the siRNA knockdowns on individual RNA transcripts, from the microarray data we calculated log ratios as $\log_2$ (abundance of each transcript/median abundance of the transcript across all 412 microarrays). Then, within each of the 400 microarrays, these log ratios were transformed to Z-scores, i.e. expressed as multiples of the standard deviation, which indicates the degree of siRNA-induced deviation from the median expression of each transcript relative to the siRNA-induced deviations from median expression of all other transcripts on any particular microarray. This analysis showed that 70% of the genes were knocked-down to ≤40% of their median value.

Genes were removed from the dataset if >10% of the measurements for them across all 400 chips were marked 'absent' based on CodeLink Expression Analysis Software. For the chips and genes with <10% 'absent' markings, the missing data were imputed using the LSImpute missing value estimation method (43) as described in our previous report (36)

Identifying genes interacting with the Rel/NFκB family

Since the computational methods used were not able to model relationships between all ≈20 000 RNA transcripts analysed by the microarrays, we chose a subset of transcripts for network analysis. To identify a subset of genes and relationships likely to be relevant to *Rel/NFκB* transcription factors, the Ingenuity Pathways Analysis (Ingenuity Systems, CA, USA) database and the Biobase BKL TRANSPATH database (44,45) were queried.

Using the Ingenuity Pathways Analysis database, a list of genes linked to *Rel/NFκB* family members either up- or down-stream, directly or indirectly, by the terms 'expression', 'trans-activation', 'DNA-binding' and 'transcription' were identified by Official Gene Symbol (OGS). All probeIDs on the CodeLink microarrays that mapped to the OGSs identified were included, giving a final list of 379 genes identified by probeID. Data for this list of genes across all 400 chips was extracted, and this formed the source dataset for the following sections.

Constructing a reference network from literature-based datasets. Systems biology databases like IPA and TRANSFAC contain gene–gene relationships that have been demonstrated in a wide variety of cell types, using many different experimental methods. To investigate how these relationships are represented in this endothelial cell dataset, we constructed a literature-derived network for comparison with networks inferred from the data using the different methods implemented in the framework.

Relationships between the list of 379 genes were extracted from IPA and TRANSFAC, and a literature-derived network was generated to describe these relationships. A total number of 2607 edges were identified between all genes identified by OGS. Literature-derived relationships between genes in this list that did not involve Rel/NFκB family members were included since they may represent direct or influences on the target genes.

Because the experimental dataset to be used in the network inference was specified by CodeLink probeID, an equivalent version of this reference network specified by probeIDs was also created. For each edge between OGSs in the reference network identified by IPA/TRANSFAC, edges were assumed to exist between each of the probeIDs mapping to any of the OGSs, creating a total of 4524 edges in the 'probeID-equivalent' reference network.

Inferring gene regulatory networks

Using common network inference methods implemented within the framework, we generated a set of gene regulatory networks from the two microarray datasets, and compared the relationships present in each of these inferred gene networks to those present in the reference networks described earlier and in the 'Results' section. For the siRNA disruptant dataset, the methods used were ARACNE (11), BANJO (21), MIKANA (25,26) and SiGN-BN (22–24). For the TNF timecourse dataset, the methods used were BANJO, MIKANA and SiGN-BN. All methods were used as per their published instructions.

RESULTS

siRNA disruptant microarray dataset

We generated two novel microarray datasets in human umbilical vein endothelial cells (HUVECs) as source data to evaluate the biological relevance of various gene network inference techniques. In the first dataset, HUVECs were perturbed under standardized conditions using a panel of siRNAs directed against 400 different transcription factors and signalling molecules chosen for their relevance to endothelial cell biology and pathology. The global changes in transcript abundance that result directly and indirectly from reducing the abundance of the target RNAs were measured using CodeLink UniSet Human 20K Bioarray microarrays (19 881 probes corresponding to 16 911 distinct Entrez IDs). To verify the effectiveness of the knockdown procedure we determined the distribution of the ratios of the expression of each target mRNA in the particular experiment in which it was targeted, to the median of that RNA's expression across all 400 microarrays (Figure 2A). The siRNA-mediated knockdowns were relatively effective; 70% of the target RNAs were knocked-down to ≤40% of their median expression value. These data have been deposited in the Gene Expression Omnibus (GEO) database with accession number GSE27869.

TNF timecourse microarray dataset

Inflammation is the body's response to tissue injury, in which blood vessels dilate and recruit leukocytes, which enter injured, infected or neoplastically transformed tissues. This critical physiological process is coordinated by endothelial cells, and in particular by their response to the pro-inflammatory growth factor TNF. In the second dataset, HUVECs were treated with TNF and samples were harvested at eight timepoints after treatment (0, 1, 1.5, 2, 3, 4, 5 and 6 h) in triplicate (24 microarrays in total). Transcript abundance at each timepoint in each of the three replicates was measured using Codelink microarray

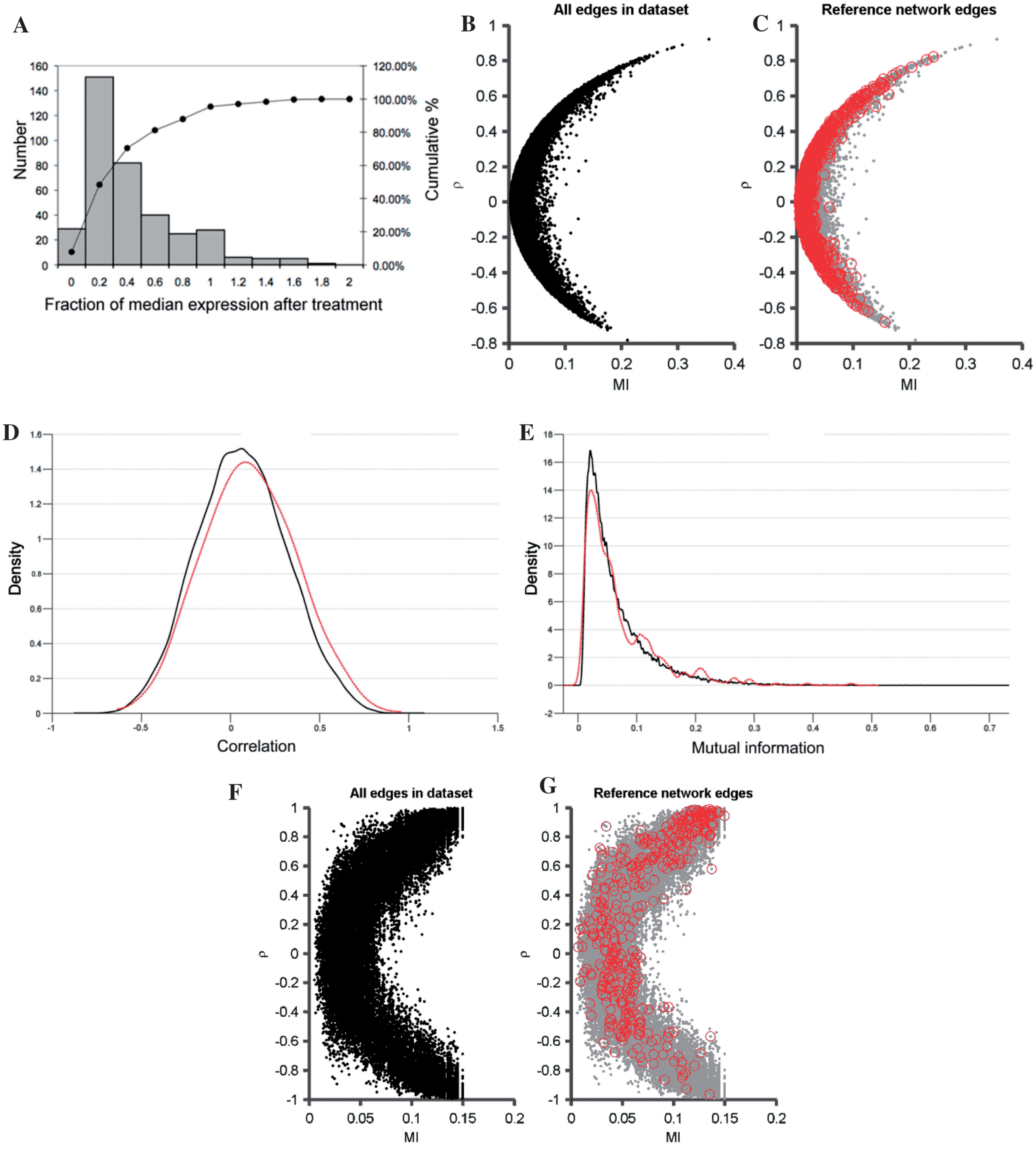

Figure 2. Results from siRNA-mediated perturbation of human endothelial cells. (**A**) Histogram of knockdown effectiveness for 400 siRNA-mediated perturbations in endothelial cells as reported by microarray (a value of 1 on the *x*-axis corresponds to no change to the array signal for the target RNA after siRNA knockdown, while a value of 0.1 corresponds to a 90% reduction in the array signal for the target RNA, so that only 10% of the median signal remains). (**B**) Distribution of Spearman correlation (ρ) and MI for all possible transcript-pairs (~140 000) between the 379 Rel/NFκB-related transcripts in the siRNA-mediated knockdown dataset. (**C**) Correlation and MI for all possible pairings from (A) are shown in grey, overlaid by the pairings found in the Rel/NFκB-related reference network in red. (**D**) Distribution of correlation for all possible pairs between the 379 Rel/NFκB-related transcripts in the siRNA-mediated knockdown dataset (black line) compared to correlation for pairs in the Rel/NFκB-related reference network (red line). (**E**) Distribution of mutual information for all possible pairs between the 379 Rel/NFκB-related transcripts in the siRNA-mediated knockdown dataset (black line) compared to mutual information for pairs in the Rel/NFκB-related reference network (red line). (**F**) Distribution of Pearson correlation and mutual information for all possible pairs (~67 000) between the 260 TNF-related transcripts in the TNF timecourse dataset. (**G**) Distribution of Pearson correlation and mutual information for all possible pairs from panel F shown in grey, overlaid by the pairs found in the TNF-related reference network (in red).

chips as described earlier and in the 'Materials and Methods' section. These data have been deposited in the GEO database with accession number GSE27870. We believe that a full analysis of these datasets will provide significant insights into endothelial cell biology, as well as providing a platform for further development of gene network inference techniques.

Reference network to compare to the siRNA disruptant microarray dataset

As discussed earlier, for both theoretical reasons (ratio of variables to observations) and practical reasons (computer hardware limits) many current gene network inference methods cannot operate on datasets containing more than a few hundred RNAs. Therefore, we selected a subset of the RNA transcripts measured by these gene arrays for network analysis. For the siRNA dataset, we focused on the Rel/NFκB transcription factor family, which is biologically relevant to endothelial cell biology, especially to the process of inflammation (46) and which contains very complex regulatory pathways that should challenge gene network inference (47). To identify the RNAs most likely to be relevant to Rel/NFκB transcription factors, the Ingenuity Pathways Analysis database (Ingenuity Systems, CA, USA) and the Biobase TRANSPATH database (44,45) were queried, identifying 379 RNAs either upstream or downstream of Rel/NFκB transcription factor activity. Relationships between these 379 RNAs were extracted from both databases and pooled with additional relationships we had manually identified from the published literature to produce an *Rel/NFκB*-related 'reference network' of 1250 previously experimentally identified relationships between any of the 379 RNAs where the abundance of one RNA may influence the abundance of another RNA. This reference network is available as Supplementary File S3.

Reference network to compare to the TNF timecourse microarray dataset

To identify RNAs and RNA-to-RNA relationships active in the TNF-treated timecourse dataset, a linear model was fitted to triplicated eight timepoint microarray data using the *lm* function in the statistical framework *R*. Next, the subset of transcripts that were differentially expressed between any two of the time points with $P < 0.001$ were extracted from the model. Two hundred and sixty transcripts were identified as being differentially expressed in this manner. As with the siRNA data set above, Ingenuity Pathways Analysis, TRANSPATH and manual searching were used to extract previously experimentally identified relationships between any of these 260 RNAs, such that the abundance of one RNA may influence the abundance of another RNA—699 relationships of this type were identified. The reference network for TNF-regulated RNAs is available as Supplementary File S4. The separate Rel/NFκB and TNF reference networks are used when evaluating gene regulatory networks generated from the siRNA treated and TNF-stimulated HUVEC microarray datasets respectively, below.

Types of Rel/NFκB-associated relationships present in the siRNA disruptant microarray dataset

A parameter that can be used to summarize the strength of 'linear' relationship between two RNAs across a data set is Pearson's correlation coefficient (ρ). A corresponding parameter for summarizing the strength of any relationship—'either linear and non-linear'—between two RNAs is mutual information (MI). The ratio ρ/MI can be used as a simple marker of relationship linearity. As an initial analysis of the characteristics of the pairwise RNA-to-RNA relationships in the siRNA data set, we examined the distribution of correlation versus MI for 'all possible' pairwise RNA-to-RNA relationships between the 379 *Rel/NFκB*-associated RNAs identified above (Figure 2B). A similar distribution to that shown in Figure 2B of correlation versus mutual information for RNA–RNA pairs has previously been observed for transcript abundance datasets (48). The subset of these pairwise relationships that are also present in the *Rel/NFκB*-related literature-based reference network are highlighted in red in Figure 2C. While some relationships found in the NFκB-related literature-based reference network had high mutual information and high correlation across our siRNA dataset, the relationships present in the reference network are generally relatively low-MI (90th percentile = 0.1420) and relatively low absolute value of correlation (90th percentile = 0.4520) relationships. Conversely, many of the strong RNA-to-RNA relationships apparent in our endothelial cell siRNA dataset have not been previously reported in the literature. Figure 2D and E show that the distributions of correlation and mutual information plotted separately as kernel density graphs. We had initially expected the reference network relationships to be more enriched for high correlation and high mutual information edges in the dataset than is apparent here. This issue is discussed below.

Types of TNF-associated relationships present in the TNF time course microarray dataset

The pairwise RNA-to-RNA relationships in the TNF-treated time course dataset were analysed in the same way. Figure 2F shows the distribution of correlation versus mutual information for 'all possible' pairwise relationships between the 260 TNF-regulated RNAs. Figure 2G shows these same pairwise comparisons in grey, with those that are also present in the reference network for TNF-regulated RNAs highlighted in red. In this case, the RNA-to-RNA relationships that are identified in the reference network generally have relatively high MI (50th percentile = 0.1315) and relatively high absolute correlation (50th percentile = 0.5932). This suggests that, unlike the *Rel/NFκB*-associated RNA-to-RNA relationships identified in the siRNA data set, many strong RNA-to-RNA relationships from the published literature were represented in our TNF timecourse data. However, there are still many strong RNA-to-RNA relationships apparent in our endothelial cell TNF timecourse data that have not been previously reported.

Coregulation reference networks

Not all RNA-to-RNA relationships identified across the siRNA or TNF timecourse microarray datasets will represent one RNA directly regulating the other. A significant proportion of observed RNA-to-RNA relationships may represent both of the RNAs being co-regulated by a third RNA, as has been previously described (49). For example, this could occur if the third RNA encoded a transcription factor that regulated both of the other RNAs by binding to promoter elements in their encoding genes. To investigate this possibility, an additional 'coregulation reference network' was prepared from each of the reference networks described above, in which edges are non-directional and represent RNA pairs that have previously been reported to share common upstream regulators (that is, the coregulation reference networks were generated by forming edges from any pairs of transcripts that shared a common regulator in the reference networks described above). The coregulation reference networks for the 379 *Rel/NFκB*-related RNAs and for the 260 TNF-regulated RNAs are provided in Supplementary Files S5 and S6, respectively.

The reference networks used in the analysis discussed below are the most extensive and accurate that we could assemble using current resources. However, systems biology databases are continually improving, and we envisage that future improvements in the content and accessibility of systems biology databases will allow more accurate and extensive analysis using reference networks. Nevertheless, the reference networks we assembled above were useful for focusing our analysis in this article, as described below.

An illustrative analysis: recovering previously known relationships associated with the Rel/NFκB family of transcription factors from the siRNA and TNF datasets

We investigated how the RNA-to-RNA relationships making up the *Rel/NFκB*-based reference network were represented in gene regulatory networks inferred from the siRNA microarray dataset. First, using common network inference methods, we generated a set of gene regulatory networks from the 379 *Rel/NFκB*-related RNAs across the siRNA dataset, and compared the relationships present in each of these inferred gene networks to those present in our NFκB reference network.

In a similar fashion, using several inference methods commonly applied to timecourse data, we generated gene networks using the 260 *TNF*-related RNAs across the *TNF* timecourse dataset. Table 3 summarizes the inferred networks produced for each of the two data sets.

All of the methods used above are included in the software framework available with this article except the SiGN-BN Bayesian network methods since they require a massively parallel supercomputer, and were conducted in the Human Genome Center, Institute of Medical Science, University of Tokyo.

To determine whether these inference methods were recovering more experimentally verified relationships than would be expected by chance, we compared the number of edges in the intersection between each inferred and reference network, with the distribution of the number of edges in the intersection between 1000 randomly relabelled inferred networks and the reference network. Figure 3 shows the results for each network type.

Figure 3 suggests that the number of *Rel/NFκB* reference network edges that were present in each of the five networks (ARACNE, MIKANA, BANJO, correlation and SiGN-BN) inferred from the 400 array/379 NFκB-associated mRNA dataset (red vertical line) was greater than the number expected to be present due to chance.

Figure 4 indicates that the number of *TNF* reference network edges that are present in the networks inferred from the *TNF* time course data set (red line) is also greater than the number expected to be present due to chance for the timecourse correlation, MIKANA (DS) and SiGN-BN networks.

Networks inferred using different methods recover similar types of RNA-to-RNA relationships

Next, we looked at the types of experimentally verified relationships that were recovered by the different types

Table 3. Summary of inferred networks and reference networks showing the number of relationships and density of each network

	# edges	Network density (edges/nodes)
NFKB-associated perturbation data		
ARACNE (mutual information)	964	3.03
BANJO (Bayesian inference)	456	1.29
MIKANA (dynamical systems)	2272	5.99
SiGN-BN (Bayesian inference)	2498	6.71
Spearman's correlation	2226	10.12
NFκB-associated reference network	*1260*	*5.12*
NFκB-associated 'coregulation' reference network	*11 582*	*54.89*
TNF-associated timecourse data		
BANJO (Bayesian inference)	1599	4.36
MIKANA (Dynamical systems)	533	1.45
SiGN-BN (Bayesian inference)	1969	8.41
Time-delay Pearson's correlation	2715	7.69
TNF-associated reference network	*699*	*5.68*
TNF-associated 'coregulation' reference network	*5599*	*49.99*

Values in italics indicates literature-derived reference networks.

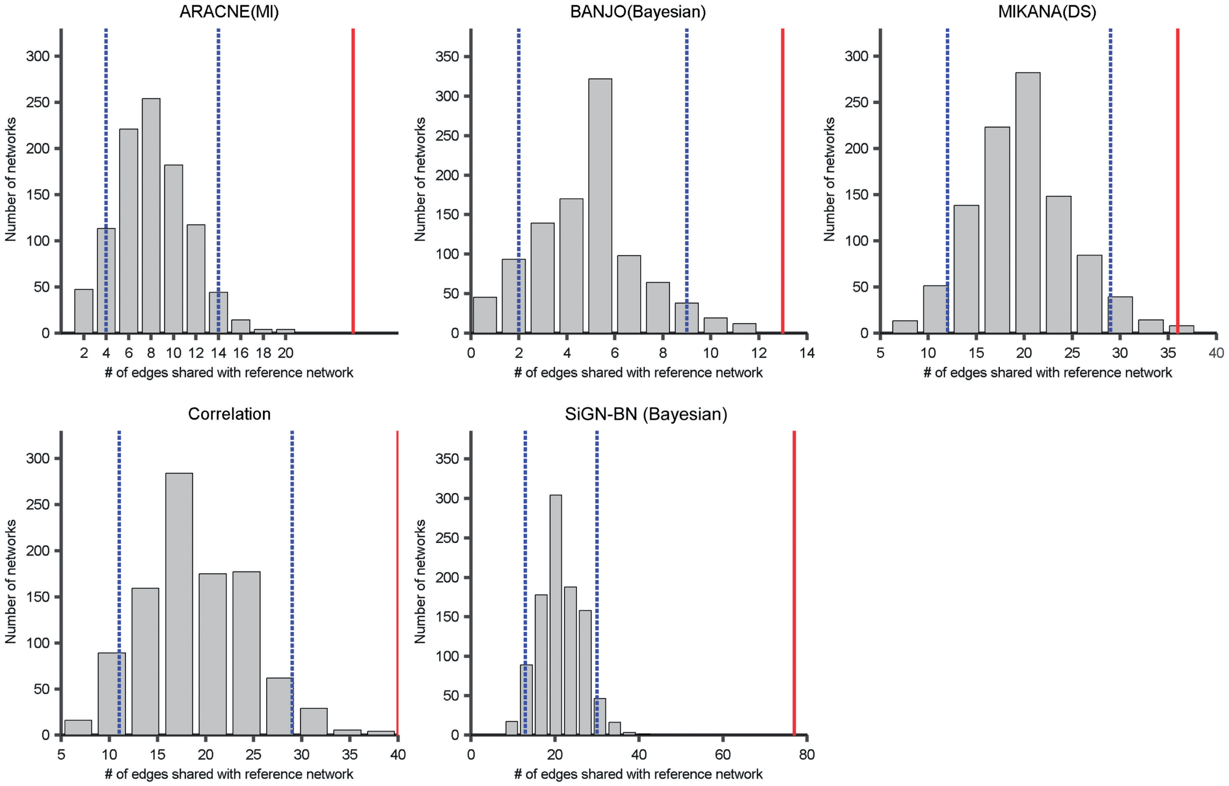

Figure 3. Number of edges (x-axis) in the Rel/NFκB reference network present in each inferred gene network generated from the siRNA dataset using five methods. The grey histogram represents the distribution of shared edges in the randomly relabelled network and the reference network. The red line indicates on each x-axis the number of reference network edges present in each inferred network, while the blue lines indicate the 95% confidence interval for the distribution of reference network edges present in the 1000 randomly relabelled networks.

of network inferred from the 379 *Rel/NFκB*-associated mRNA/400 siRNA microarray data set, and we found that different networks showed a surprising degree of coherence between the types of relationships they recovered.

Figure 5A shows the distribution of mutual information and correlation for the reference network edges present in each of the inferred networks; most of the relationships identified in each inferred network type have $|\rho| > 0.5$. This point is also illustrated by Figure 5B, which shows that most reference network relationships identified by any of the network inference methods tend to be between transcripts with high absolute value of Pearson's correlation. Restricting the inferred networks to a subset of edges with the strongest support from the data or strongest interaction strength (as determined by the particular network inference method used), tends to select for edges with high absolute value of correlation. Since these networks are generated using different inference methods based on different mathematical approaches, it is noteworthy that the most strongly-supported edges in each method tend to be edges with the highest absolute value of correlation.

Figure 5C shows the overlap between the *Rel/NFκB* reference network edges present in each inferred network as a Venn diagram. The ARACNE (MI) network shares most of its recovered reference network edges (24/26) with the SiGN-BN (Bayesian) network, and nearly half (12/26) with the MIKANA network. In the MIKANA and SiGN-BN networks, there are a set of relationships not shared with any of the other networks, but there is also a large cohort of relationships in common (e.g. 25/36 of the *Rel/NFκB* reference network edges identified by the MIKANA network were also identified by at least one other network inference method).

As discussed earlier, it is possible that some RNA-to-RNA relationships (edges) identified across the siRNA data set by the five gene network inference methods were 'coregulated' transcripts—that is, pairs of transcripts that are regulated by the same 'parent' transcript, which may or may not be a node in the network under study. To investigate this possibility, we created 'coregulation' reference networks as described earlier. We measured the overlap between these coregulation networks and the networks inferred using the different inference methods, and compared the number of overlapping relationships to a distribution obtained by chance from randomly relabelled inferred networks. Figure 6 shows the result for the coregulation network created for the *Rel/NFκB*-centred knockdown dataset.

Figure 6 suggests that the ARACNE (MI), MIKANA (DS), correlation and SiGN-BN inference methods using the siRNA data set recovered more coregulation reference edges from the 379 *Rel/NFκB*-associated mRNA/400 siRNA data than we would expect by chance alone.

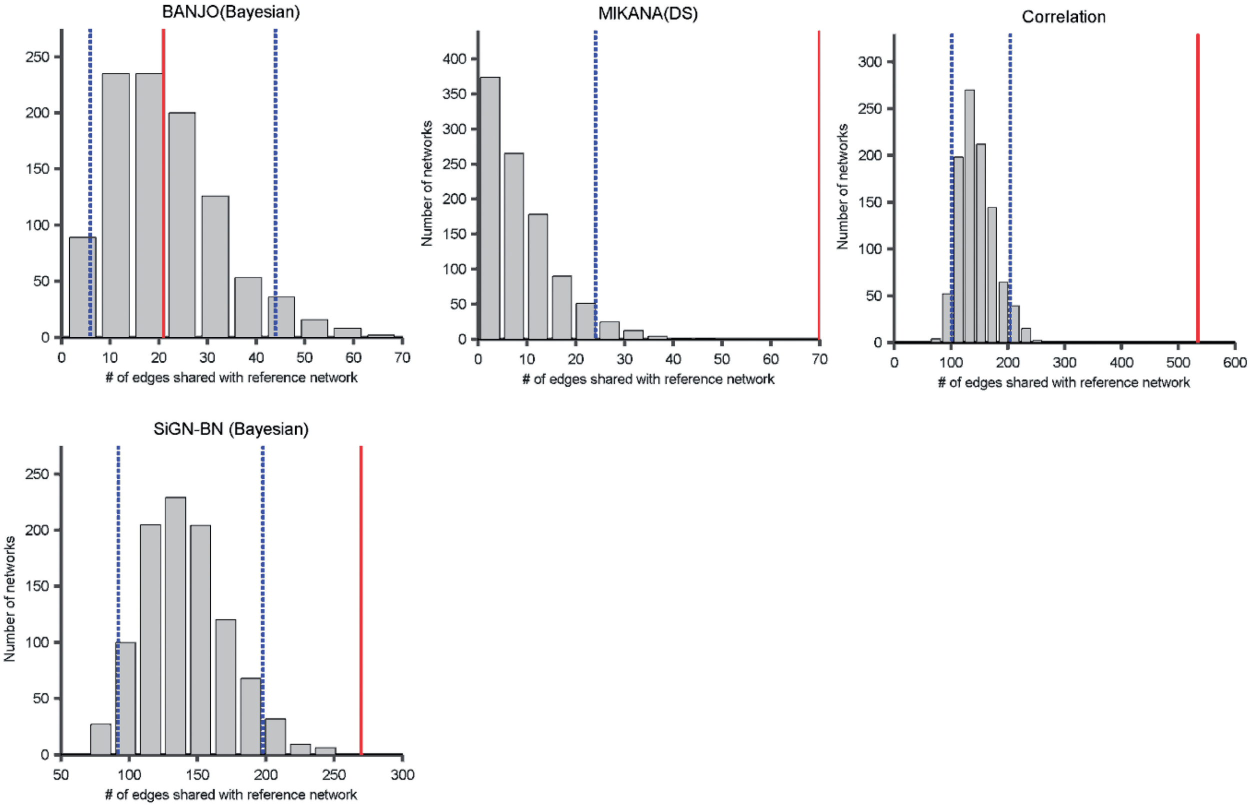

Figure 4. Number of edges (x-axis) in the TNF-based reference network present in each inferred gene network generated from the TNF timecourse dataset using five methods. The grey histogram represents the distribution of shared edges in the randomly relabelled network and the reference network. The red line indicates on each x-axis the number of reference network edges present in each inferred network, while the blue lines indicate the 95% confidence interval for the distribution of reference network edges present in the 1000 randomly relabelled networks.

We then performed a similar analysis on the *TNF* timecourse dataset (Figure 7).

Figure 7 indicates that the number of reference coregulatory network relationships that are recovered by the inferred networks (red line) from the TNF timecourse dataset is outside the 95% confidence interval for the distribution of relationships recovered at random (blue lines) for the timecourse correlation, MIKANA and SiGN-BN networks, suggesting that these methods recover more coregulatory relationships from the data than we would expect by chance alone.

The identification of statistically significant numbers of co-regulatory reference network relationships by both disruptant and timecourse network inference methods is interesting, since coregulatory relationships are both biologically relevant and common; for example, molecules within the same pathway or functional group appear to be more often co-regulated than expected by chance (50). Therefore, we suggest that the identification of coregulatory relationships is an important function of gene network analysis, and should be included in the evaluation of gene network inference methods.

Networks inferred using different methods have different structure

Results in the previous section suggest a high degree of similarity between the pairwise edges recovered by the different network inference methods (Figure 5C). However, gene regulatory networks can also be examined at a level above that of direct pairwise relationships. A common feature of interest in a gene network is the presence of highly connected nodes with many relationships, sometimes referred to as network 'hubs'. For example, a network hub with many downstream nodes associated with a specific cellular process may in theory be a master-regulator of that cellular processes. Potentially, hubs may be more likely to be biologically valid than single edges, since the information supporting the identification of each hub (a large number of paired observations) is far greater than the information supporting the identification of each edge (a single paired observation). Regulatory hubs are familiar to most biologists in the context of transcription factors that act as 'master regulators' of development (51–53). Identifying the regulatory hubs associated with human disease has become a major scientific focus, since these hubs may provide critical insights into pathology as well as providing new clinical biomarkers and drug targets (54,55).

For undirected networks, a hub can be any RNA involved in many relationships, but for directed networks, a hub can be either an RNA which influences many other RNAs ('out-degree' hub), or one which itself is influenced by many other RNAs ('in-degree hub').

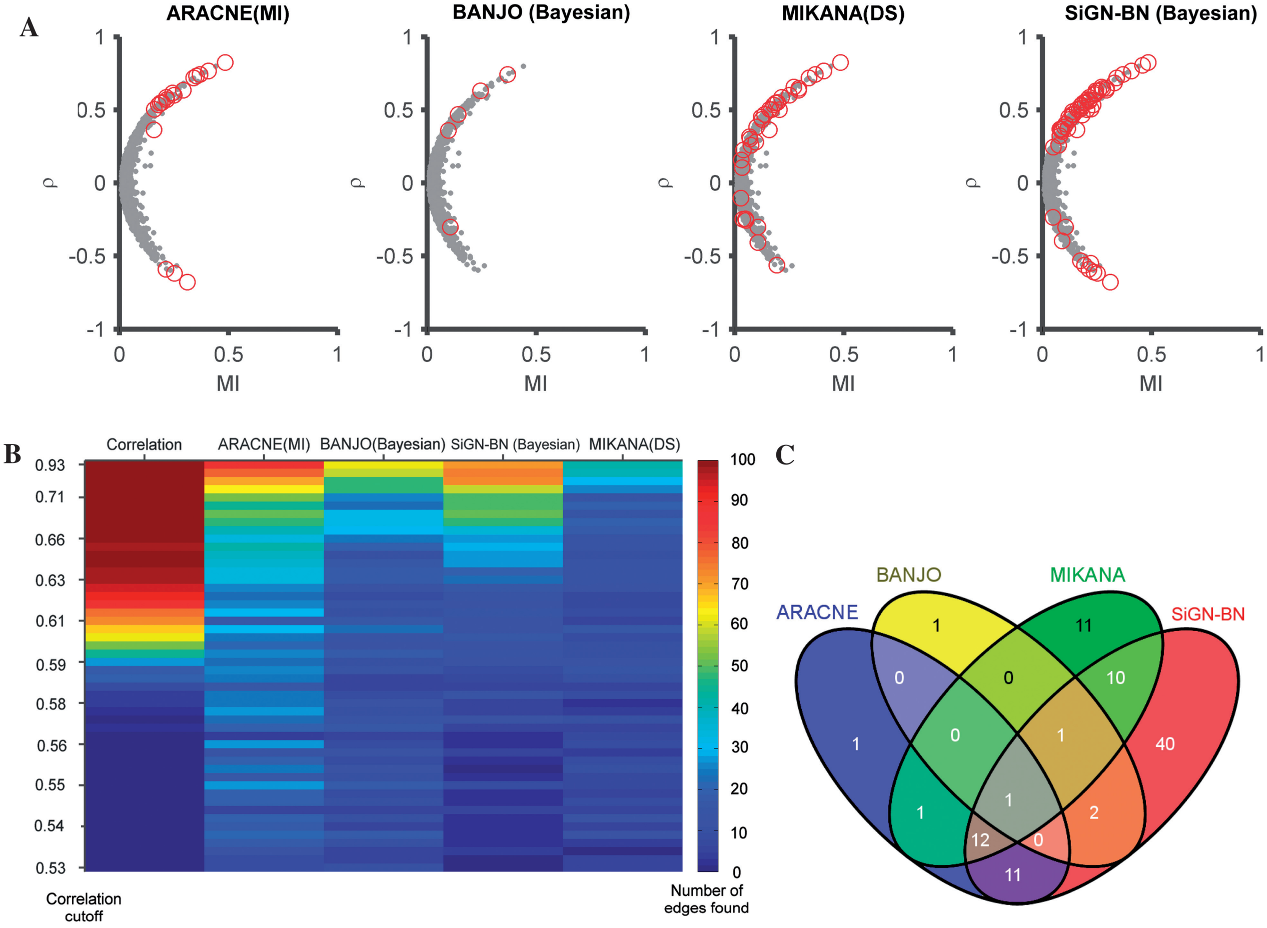

Figure 5. Comparison of the types of reference network edges recovered by different inferred networks. (**A**) MI versus correlation graphs for the edges recovered by each of the inferred networks. Red circles indicate the MI and correlation for the reference network edges, while grey dots indicate the distribution of all possible edges in the 379 NFκB-associated mRNA/400 siRNA microarray data set. (**B**) Heat-map of NFκB reference network edges sorted into bins in descending order of absolute value of Pearson's correlation. Only the 5000 most correlated edges are included, so the *y*-axis bins include only reference network edges with absolute value of Pearson's correlation coefficient between 0.53 and 0.93. Band colour represents the fraction of edges in each bin that were identified by each inferred network, as defined by the key at the right of the heat map. (**C**) Venn diagram showing the reference network edges present in four of the inferred networks.

In addition, some researchers have identified biologically important hubs as those network nodes through which the greatest amount of information passes (56). 'In-degree' and 'out-degree' hubs also play different roles in non-biological systems; a comparison between network structures in transcriptional regulatory networks and in computer operating systems (56) suggests that 'out-degree' hubs are uncommon in operating system structure, but that 'in-degree' hubs represent highly reused functional modules.

For the methods that infer directed networks, (MIKANA, and the BANJO and SiGN-BN methods), we chose to focus on out-degree hubs, or nodes with many 'children'. The reason is that we expect out-degree hubs to be more likely to be involved in the regulation of many other RNAs, and consequently to be more biologically relevant. In-degree hubs possibly represent transcripts for which the 'explanation' in a network sense is complex, and consequently more likely to be affected by noise in experimental data. Marbach et al. (15) found that relationships for in-degree hubs were generally more difficult to recover from simulated data than relationships for transcripts with few parents, and we expect this trend to be the same in our experimental data.

To visualize the relationship between connectivity of nodes (in terms of the number of relationships in which they are involved), we plotted the connectivity of each node in one network versus its connectivity in another network. Additionally, we represented the fraction of relationships each node has in common between the two networks on a colour scale, from blue (few common relationships) to red (many common relationships; see 'Materials and Methods' section). To illustrate this approach, we show comparisons between two pairs of gene networks generated from the 379 *Rel/NFκB*-associated mRNA/400 siRNA dataset in Figure 8.

Figure 8A shows a general trend of correlation between connectivity in the ARACNE (MI) and SiGN-BN (Bayesian) networks. In general, nodes that were highly connected hubs in one network were also highly connected hubs in the other network, and most nodes with ≤10

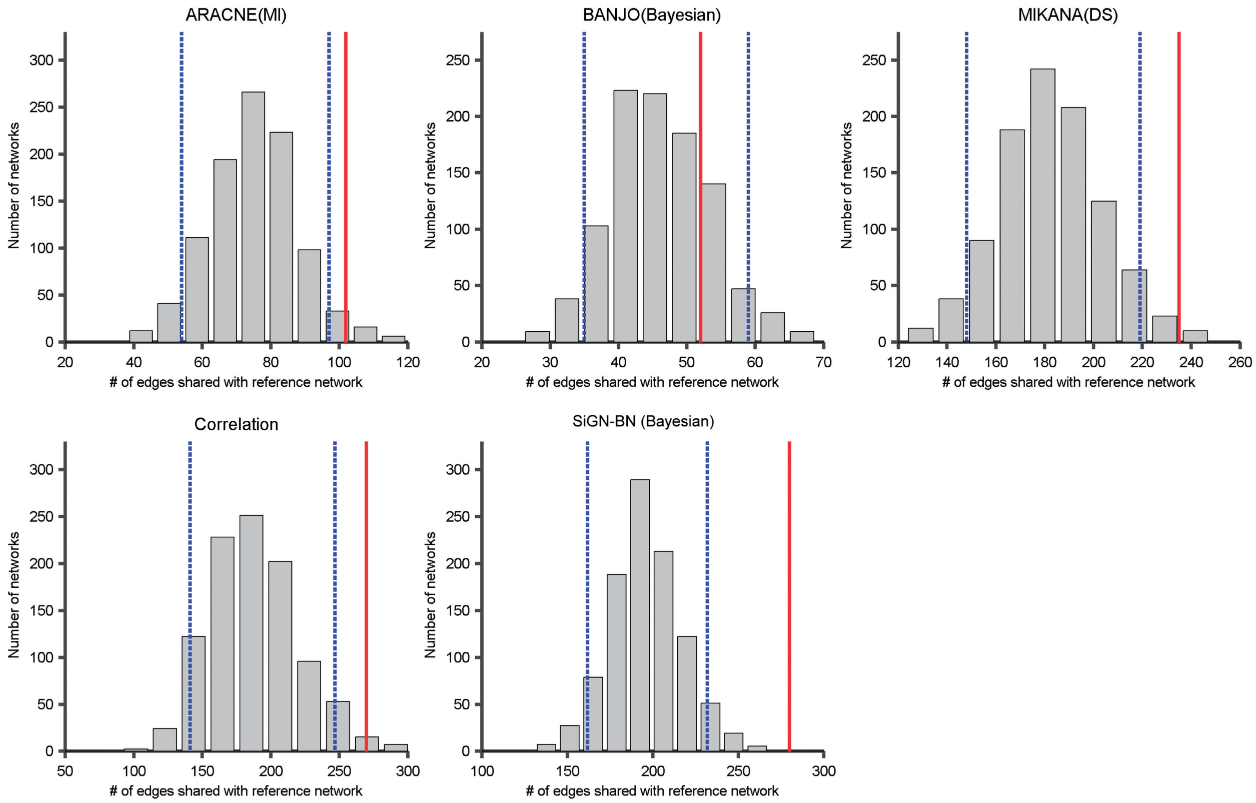

Figure 6. Coregulatory relationships in the Rel/NFκB-based reference network present in networks inferred using five different methods, compared to coregulation network relationships recovered at random. The red line indicates the number of coregulatory reference network relationships present in each inferred network, while the blue lines indicate the 95% confidence interval for the distribution of edges recovered from 100 randomly relabelled inferred networks.

connections have >50% of their children in common between the two networks. However, the most highly connected transcripts do not have many connections in common; some of the genes that are connected to ≤10 transcripts in both networks share only around 50% of these connections (the points toward the top right corner). Figure 8B compares the ARACNE (MI) and MIKANA networks and shows that despite the similar types of edge recovered by the two network methods (Figure 5), we observed little relationship between connectivity of the individual nodes in the two networks. Of the nodes that are well-connected in both networks (having more than 10 connections in both networks), many have <50% of their children in common. The presence of hubs in two networks that are defined by different sets of edges is interesting; this may reflect the different biological meaning of edges and hubs in different network inference methods, or the different criteria applied by each inference method for identifying network edges. The software framework we provide will allow more extensive future analysis of this issue across several network inference methods.

The siRNA data reveals indirect relationships between the upstream regulators of NFKB1 and NFKB1 targets

Many previously identified relationships between *NFKB1* and its known targets were not observed as edges in the inferred gene networks described earlier. A possible reason for this observation is that the *NFKB1* transcript has only low correlation or mutual information across the siRNA data set with its known targets. This would not be surprising, given that the activity of the NFκB1 transcription factor family is not only regulated by the transcription of the genes encoding its protein subunits, but is strongly regulated by post-transcriptional mechanisms including: post-translational modification (phosphorylation/acetylation) by upstream proteins (57,58), complex formation with other transcription factors (59), cytoplasmic sequestration (60) and dimer exchange (61). This means that, although the abundance of targets of the *Rel*/*NFκB* family is determined by the activity of this family, the abundance of targets of the *Rel*/*NFκB* family may not be associated with the abundance of the RNAs that encodes the *Rel*/*NFκB* family members. In other words, we do not expect *Rel*/*NFκB* activity to be strictly proportional to *Rel*/*NFκB* subunit mRNA abundance.

Figure 9 summarizes the key regulators of the *Rel*/*NFκB* family and describes the basic types of relationships between them.

Figure 9 suggests a 'generation skip' hypothesis, in which *NFKB1*-to-*NFKB1* target relationships may not be apparent in gene networks inferred from RNA data, due to strong effects on *Rel*/*NFκB* activity of post-transcriptional mechanisms, swamping the effects on

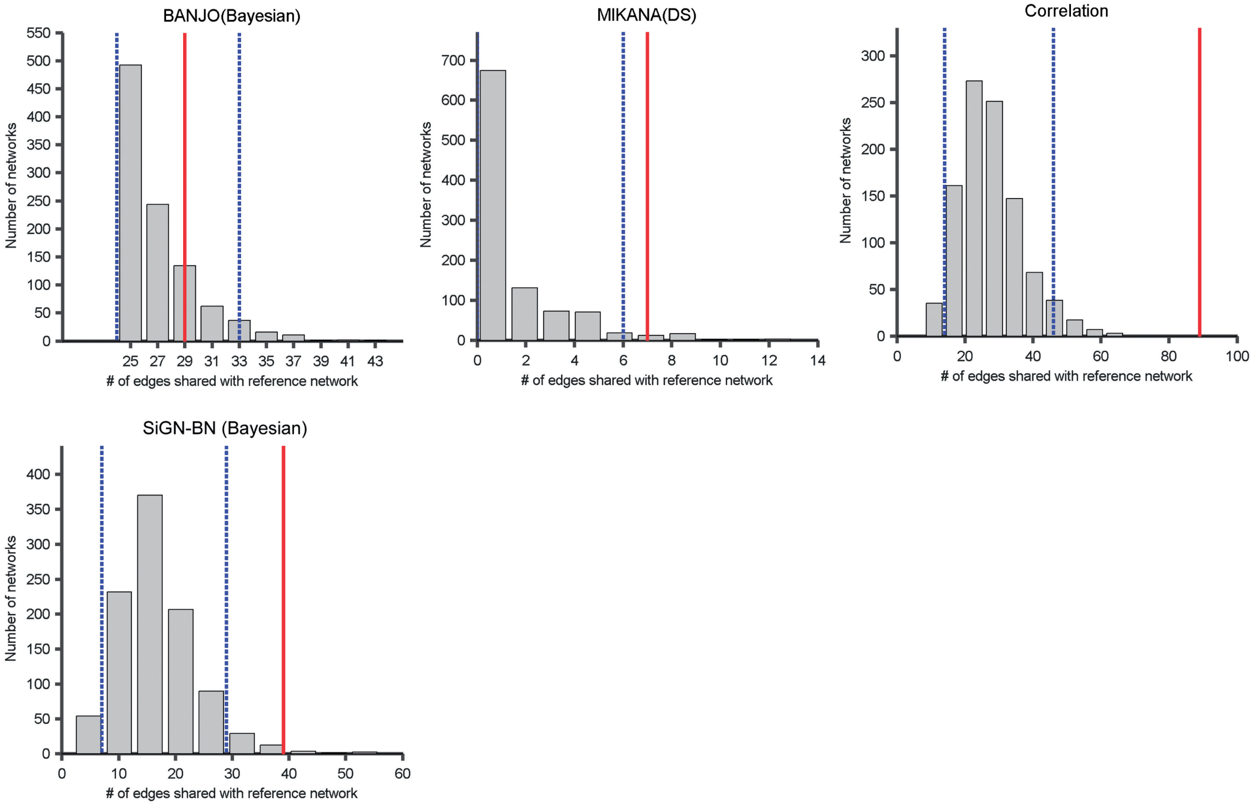

Figure 7. Coregulatory reference network relationships present in the TNF-based reference network inferred using three different methods, compared to coregulatory reference network relationships recovered at random. The red line indicates the number of relationships recovered by each inferred network, while the blue lines indicate the 95% confidence interval for the distribution of edges recovered from 100 randomly relabelled networks.

Rel/*NFκB* activity of *Rel*/*NFκB* subunit RNA abundance. If this was the case, the gene networks may be expected nevertheless to contain information about relationships between upstream regulators of *NFKB1* and *NFKB1* targets. To investigate this hypothesis, we compared the Spearman correlation and mutual information across our siRNA dataset of *NFKB1*-to-*NFKB1* targets (Figure 10A red curves), to the correlation and mutual information across our siRNA dataset of the upstream regulators of *NFKB1*-to-*NFKB1* targets (Figure 10A black curves). We observed, on average, stronger correlation and MI between *NFKB1* regulators and *NFKB1* targets, than between *NFKB1* itself and its targets. This observation is consistent with our ‘generation skip’ hypothesis outlined above, and is further illustrated using specific examples of relationships between upstream regulators of *NFKB1*, *NFKB1* itself and the targets of *NFKB1* (Figure 10C and Supplementary File S2). This now requires more detailed investigation using other transcription factor examples.

This interesting observation illustrates two key points about gene network analysis. First, as described in the Introduction, the nature of gene network edges is complex. Edges may represent indirect relationships between RNAs involving one or more intervening protein-based steps (as may be the case for the relationships in the siRNA dataset between *NFKB1* upstream regulators and *NFKB1* targets). Second, gene networks cannot identify relationships that are not evident in the data set used to infer the networks. In this example, the dataset contains very little information about relationships between *NFKB1* RNA abundance and the abundance of its target RNAs, since the transcription factor activity of NFKB1 is regulated to such a large extent by localization and phosphorylation that the abundance of the RNA encoding NFKB1 bears little relationship to the activity of this transcription factor, or to the abundance of its targets. Similar phenomena are known to occur for several transcription factors (62). We found that in contrast, relationships between RNAs encoding upstream regulators of NFKB1 transcription factor activity and RNAs encoding NFKB1 targets were evident in the data and can be inferred in gene networks—leading to an apparent generation skip in the inferred networks. While beyond the scope of this article, in the future it will be interesting to use databases of transcription factor post-translational modifications (63) to identify instances of this phenomena and to use this information as a prior when generating gene regulatory network models.

Pathway-level comparisons between gene networks

Finally, we developed a method to compare relationships in one network against ‘pathways’, made up of multiple

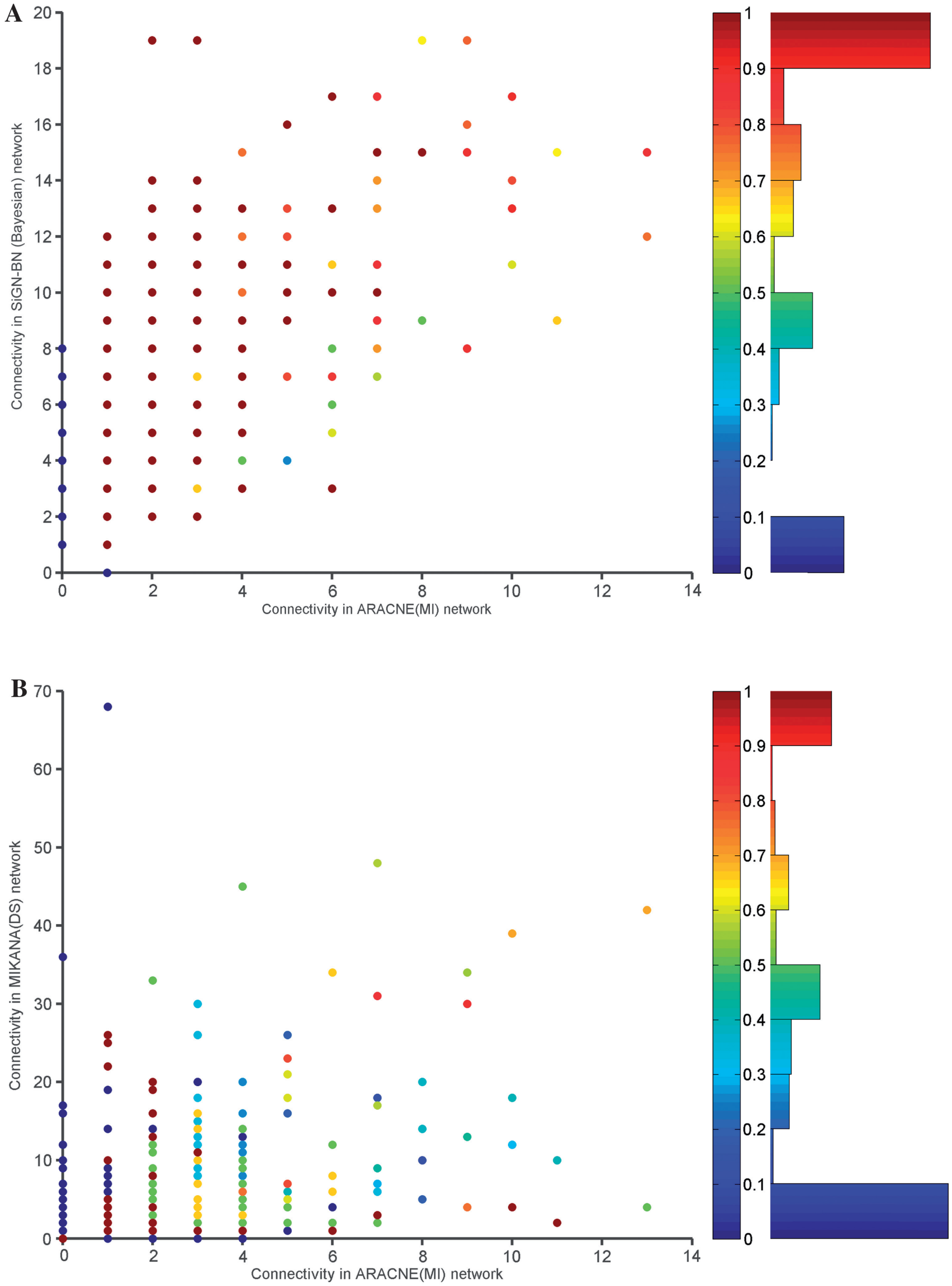

Figure 8. Assessment of higher structures in networks. (**A**) Comparison of connectivity in the ARACNE (MI) network and the SiGN-BN (Bayesian) network. (**B**) Comparison of connectivity in the ARACNE (MI) network and the MIKANA (DS) network. Each point represents a single node. Points are coloured according to the fraction of relationships each node has in common between the two networks, from blue (few common relationships) to red (many common relationships). The histogram on the right of the plots indicates the relative fraction of nodes in each degree of commonality, from blue to red.

relationships in a second network. The motivation for this work was that a pathway-level comparison may reveal similarity between two networks that was not evident from direct comparison between the edges of the two networks.

Figure 11A illustrates an application of this principle to compare one network (left) with another network (right) (in this case a reference network and an inferred network are compared) by determining the number of 'hops' between two genes. We can find the shortest path in the inferred network (smallest number of 'hops' between nodes) required to link each pair of nodes in the reference network connected by a single edge, and plot the distribution of the number of 'hops'. Figure 11B shows an example based on the gene networks inferred from our 379 NFκB-associated mRNA/400 siRNA dataset, illustrating the distribution of the shortest paths in the MIKANA network to trace each edge in the SiGN-BN

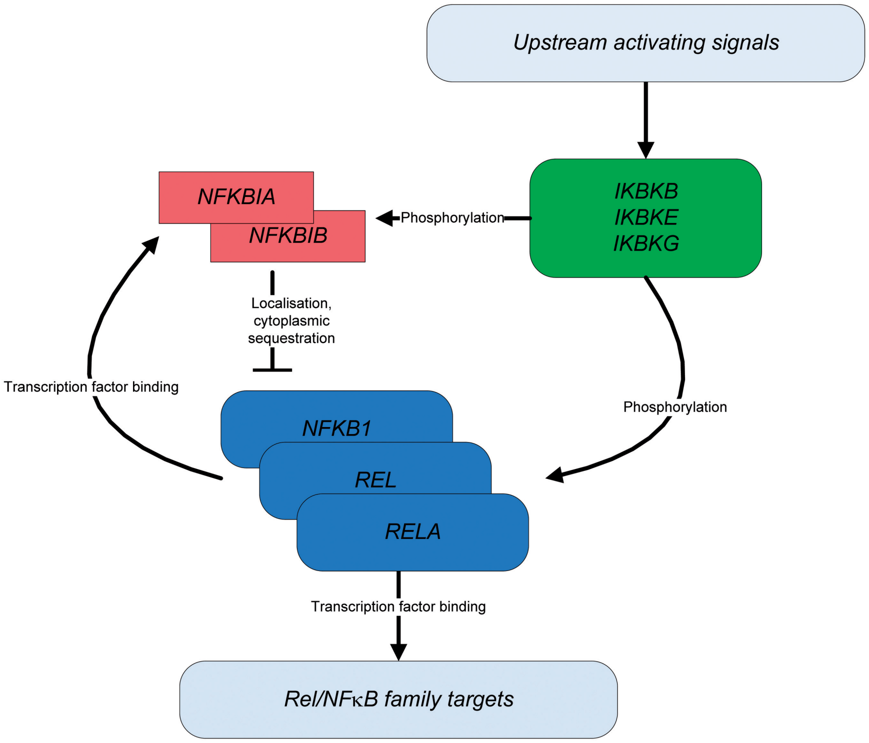
Upstream activating signals
IKBKB
IKBKE
IKBKG
Phosphorylation
NFKBIA
NFKBIB
Localisation, cytoplasmic sequestration
Transcription factor binding
Phosphorylation
NFKB1
REL
RELA
Transcription factor binding
Rel/NFκB family targets

Figure 9. Schematic of key relationships surrounding the Rel/NFkB family. Red upstream regulators are primarily associated with localization, and green upstream regulators are primarily associated with phosphorylation.

(Bayesian) network. At $x = 1$ 'hop' in this plot, there are 500 edges in the SiGN-BN network which exactly match edges in the MIKANA network (red curve), and this is many more than seen in 100 randomly relabeled MIKANA networks (i.e. more than expected due to chance alone grey curves). This is not surprising, given our previous observation that these two network inference methods recover edges with similar characteristics. However, Figure 11B also shows that 350 edges present in the SiGN-BN network are represented by two hops in the MIKANA network ($x = 2$), and this is also more than in the random networks. This suggests that, in addition to the intersection of 500 edges between these two networks, the two networks share more two-step pathways than would be expected by chance.

In the previous section, we showed that, on average, there was stronger correlation and MI between upstream regulators of *NFKB1* and *NFKB1* targets, than between *NFKB1* itself and its targets. Therefore, to illustrate a different way in which this shortest-path analysis method may be used, we identified the distribution of the number of hops in the reference network required to trace the specific *NFKB1* regulator to *NFKB1* target edges from each inferred network. Figure 11C shows the distribution of shortest paths in the reference network for the *NFKB1* regulator to *NFKB1* target inferred network edges. More of the *NFKB1* regulator to *NFKB1* target inferred network edges are represented by two hops in the reference network than would be expected by chance. This illustration is in line with expectations, since the relationships identified in the published literature between *NFKB1* regulators to *NFKB1* targets (and present in the reference network) in many cases involve more than one hop (e.g. *NFKB1* regulator-to-*NFKB1*-to-*NFKB1* target).

Edge directionality

We then explored the ability of the directional network inference methods (SiGN-BN Bayesian, Banjo Bayesian and MIKANA dynamical systems models) to correctly identify the directionality of edges—i.e. to identify cause-and-effect relationships from transcriptome data. We did this by comparing our reference networks to both forward inferred networks (the inferred networks discussed earlier) and reversed inferred networks (networks in which the parents and children had been swapped with one another). We were surprised to find that for all networks inferred from either the siRNA or timecourse data, the reversed networks contained many more reference network edges than expected by chance, just as the forward networks had done (data not shown). This was apparent even when we restricted the inferred networks to those edges we had the most directional information about—for example, using those edges in bootstrapped SiGN-BN Bayesian networks which had the same direction in networks inferred from ≥50% of 1000 bootstrapped data sets (Figure 12).

This observation may be in-part due to the complexity of the 'real biology' associated with *Rel*/*NFκB*, *TNF* and similar transcriptional pathways (47). For example, often

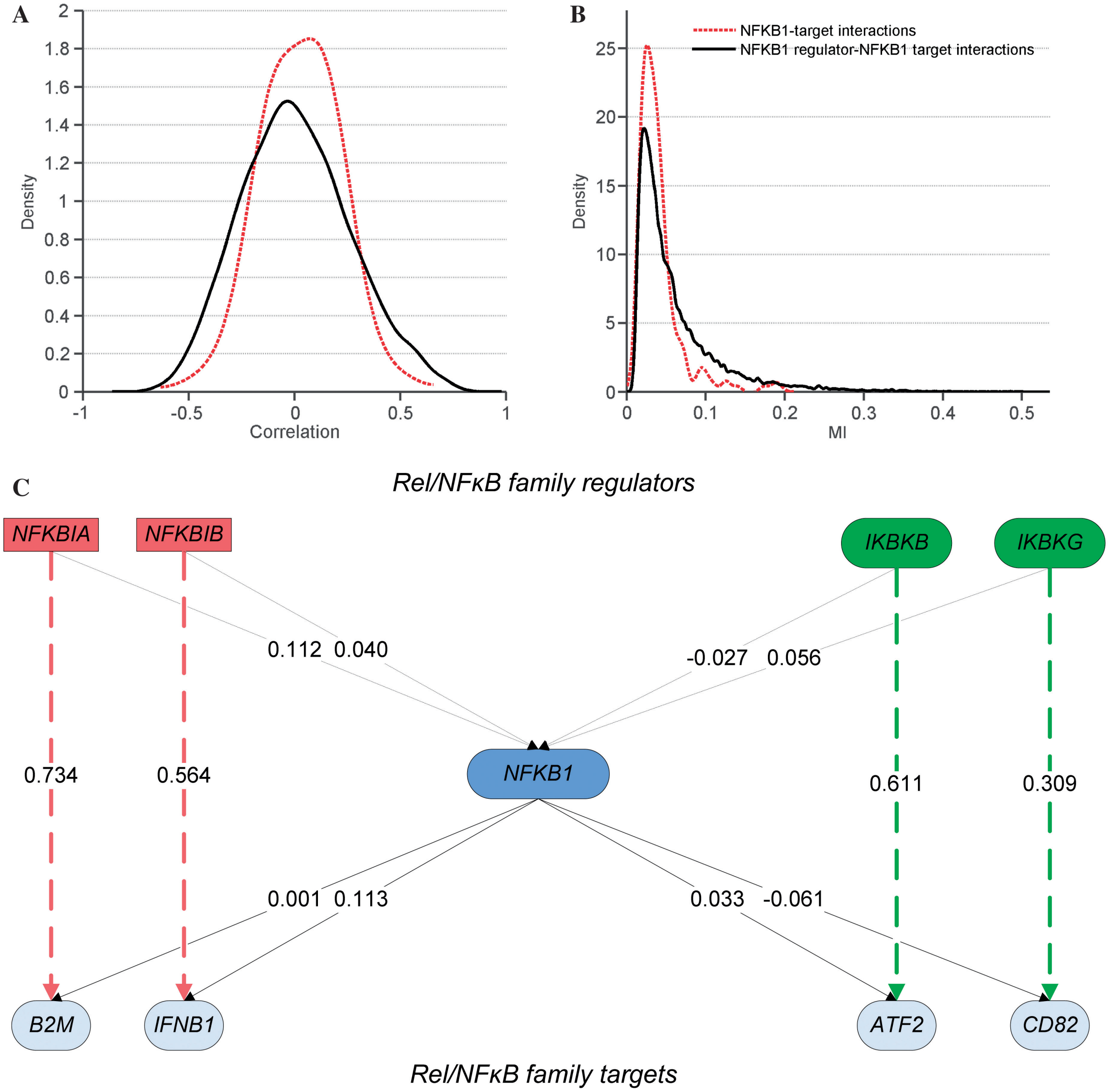

Figure 10. Comparison of correlation and MI between NFKB1 and its targets and also between regulators of NFKB1 and its targets. Relationships in the whole dataset (black line) with relationships between transcripts identified as being experimentally related in the reference network (red line). (**A**) Comparison of correlation and (**B**) MI: (i) between all possible relationships between upstream regulators of NFKB1 and NFKB1 targets (black line) and NFKB1 and its targets (red line). (**C**) Schematic showing examples of Spearman's correlation coefficients between reference network NFKB1 regulators, NFKB1 targets and NFKB1 itself, taken from the key regulators shown in Figure 9. As in Figure 9, red upstream regulators are primarily associated with localization, and green upstream regulators are primarily associated with phosphorylation.

the parent and child of a directional reference network edge are themselves both regulated by the same upstream transcription factors. To illustrate this point, in the *TNF* reference network, in 603 of the 699 directional edges, both the parent and child are also the targets of a common upstream regulator, so that these 603 parent-child RNA pairs are also connected as edges in the *TNF* coregulatory reference network. In addition, several of the 'key players' in *Rel/NFκB* and *TNF* pathways regulate one another. For example, *NFKB1* and its heterodimeric partner *Rel* are known to regulate the expression of one another, as well as regulating the expression of their direct upstream regulators *NFKBIA* and *IKBKE,* and their indirect upstream regulator *TNF.*

In addition to this reference network complexity and circularity, it is possible that there is relatively little information about cause and effect in the highly correlated relationships that the inference methods identified in the siRNA and timecourse datasets. As a further level of complexity, in some cases we observed positive correlation between two RNAs in one subset of the siRNA data but anti-correlation between the same two RNAs in another subset of the data (e.g. CXCL2 and IL8, e.g. ORC5L and CD45L). Presumably, in these situations the RNA-to-RNA

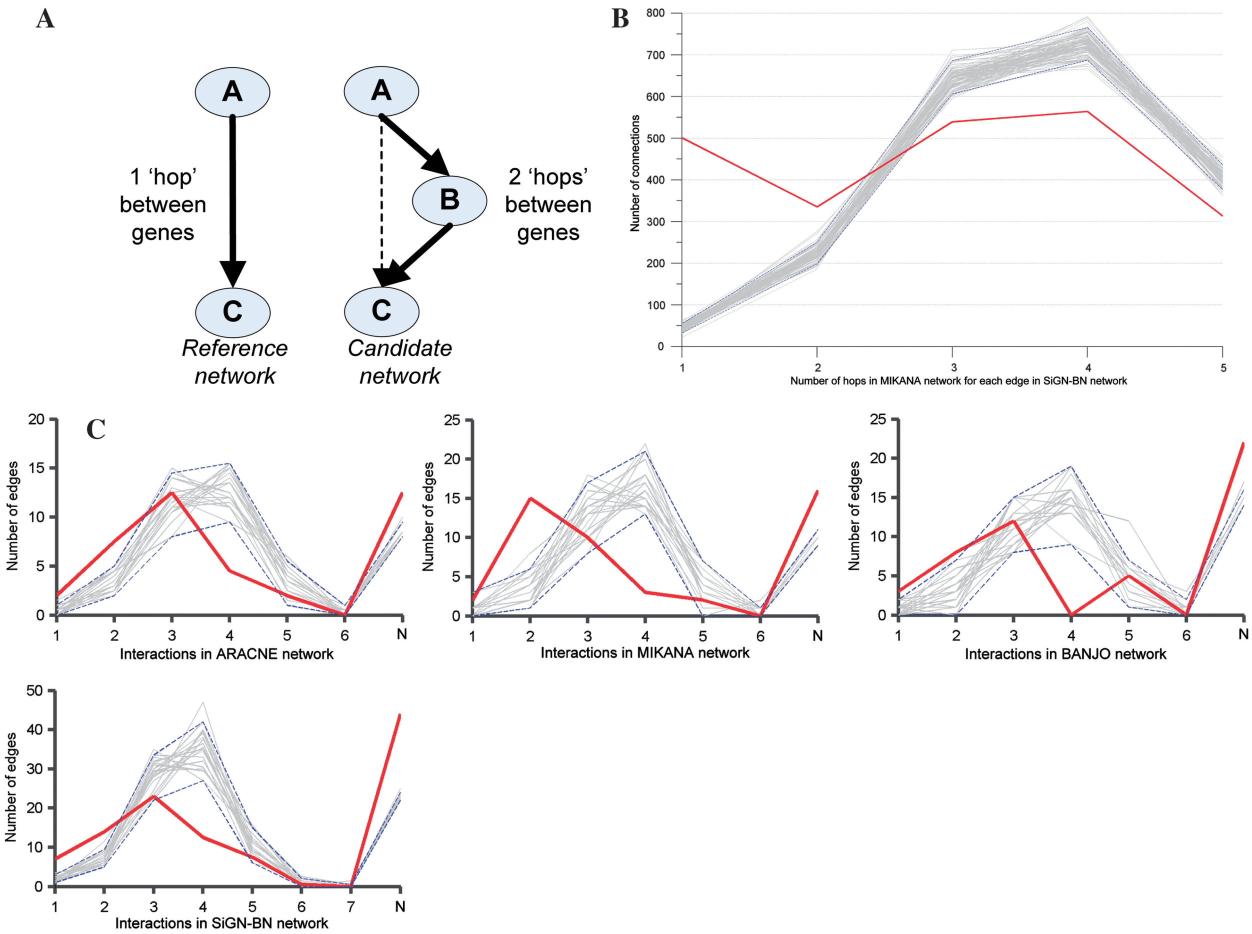

Figure 11. Assessment of pathways in selected networks. (**A**) Basic principle of comparing the number of 'hops' between transcripts. In the reference network, transcripts A and C are connected by a single interaction. In the inferred network, A is indirectly connected to C via B. (**B**) Pathway comparison of a MIKANA network against each edge in the SiGN-BN (Bayesian) network. The red line indicates the number of transcripts connected by a particular number of hops in the MIKANA network, required to trace every directly connected pair of nodes in the SiGN-BN (Bayesian) network. The grey lines indicate the number of transcripts connected by a particular number of hops for the randomly relabelled networks, and the dotted blue lines indicate the 95% confidence interval for the distribution of randomly relabelled networks. (**C**) Pathway comparison showing the number of hops required to trace NFKB1 regulator to NFKB1 target inferred network edges in the NFKB1-associated reference network. X = N represents the situation where no path exists.

relationship depends on the state of the cells and the expression of additional functionally related molecules.

Given this complexity in both the biology underlying reference networks and consequently in the data sets, it is not surprising that the inferred networks frequently identify reference network edges in the reversed direction. Additional work using gene sets with more simple regulation (if these can be found) may be required to robustly study the ability of gene network inference to correctly identify cause and effect relationships.

DISCUSSION

Addressing barriers to adoption of network inference techniques

In this study, we have developed methods and resources to address three of the barriers to the widespread adoption of gene network inference techniques. First, to address the shortage of appropriately dimensioned disruptant transcriptome data in mammalian cells, we generated a new 400-chip microarray dataset in HUVECs using siRNA-mediated knockdown of 400 specific signalling molecules and transcription factors. There are very few datasets generated from normal primary human cells of this scale publicly available. We believe the large number of observations in this dataset will facilitate further development and optimization of gene network inference methods. In addition, it is likely that analysis of this dataset will provide new insights into vascular biology and pathology. We also generated a new triplicate eight timepoint TNF response timecourse data set in HUVECs for use by researchers developing timecourse-based gene network inference methods. A detailed biologically-focused analysis of these two datasets is beyond the scope of this article and will be fully described in a subsequent publication.

Second, to address the problem that many network inference approaches require specialized programming knowledge, we have produced a comprehensive set of

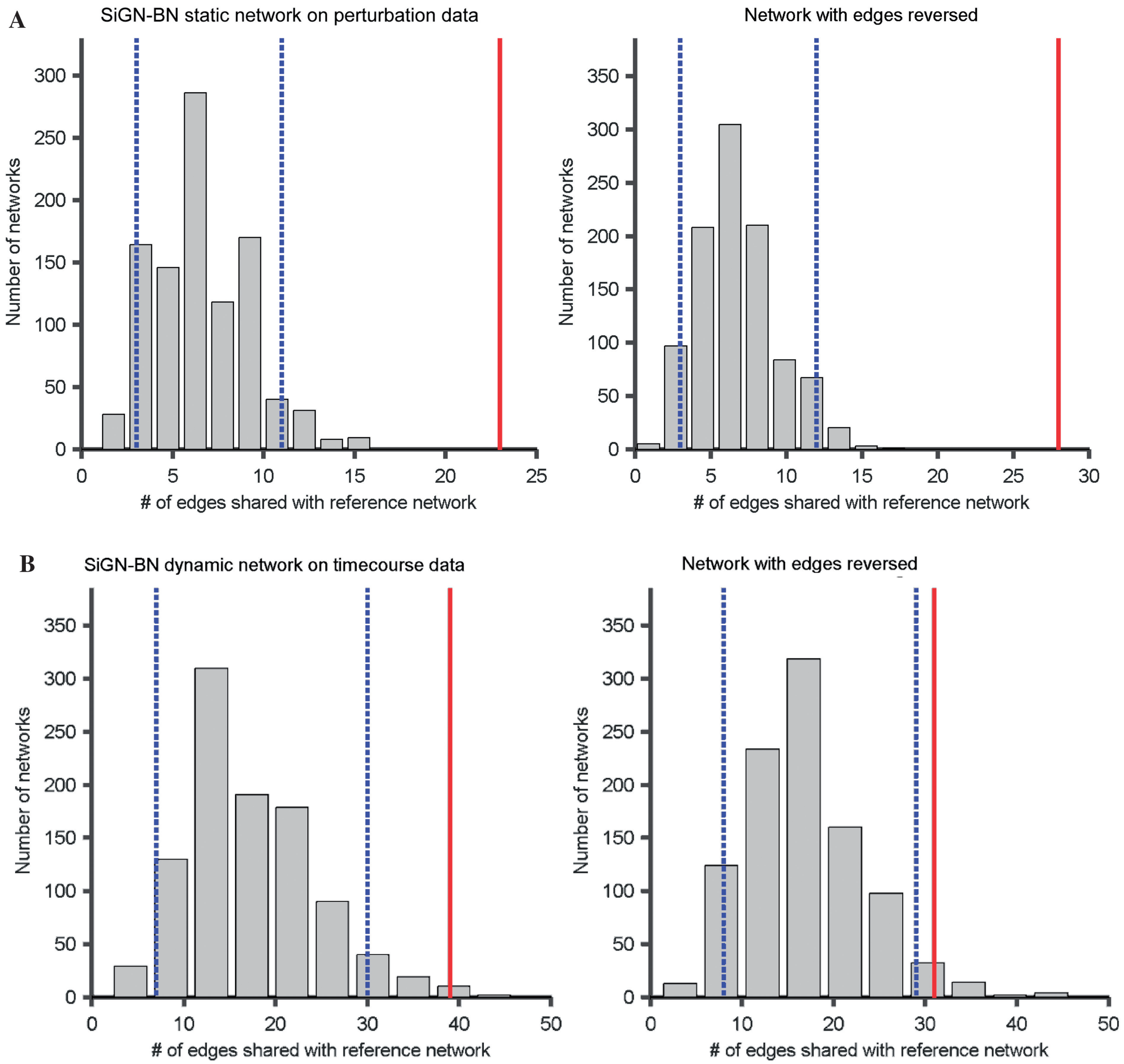

Figure 12. Assessing directionality of relationships in the perturbation and timecourse datasets. (**A**) Relationships in the Rel/NFκB-based reference network present in forward (left) and reversed (right) SiGN-BN Bayesian networks inferred from the siRNA dataset, compared to reference network relationships recovered at random. (**B**) Relationships in the TNF-based reference network present in forward (left) and reversed (right) SiGN-BN Bayesian networks inferred from the siRNA data set, compared to reference network relationships recovered at random. The red line indicates the number of reference network relationships present in the forward and revered inferred networks, while the blue lines indicate the 95% confidence interval for the distribution of edges recovered from 1000 randomly relabelled networks.

GenePattern modules for gene network inference methods. These contain one method from each of four major classes (dynamical systems models—MIKANA, mutual information—ARACNE, Bayesian networks—BANJO and correlation). These modules are available from the authors. They are readily usable by biologists with minimal programming experience, and can be readily set up on a GenePattern server.

Third, to make different inference methods easily comparable, we have developed a set of tools to compare networks generated by different methods and extract features of interest from them. These tools were used here to examine relationships present in subsets of our new microarray data, and to propose hypotheses about the way in which functional relationships between proteins are represented in transcriptome-level data.

Comparison of networks

We found that most of the different types of networks we inferred recovered more experimentally verified relationships than would be expected by chance, and we observed a high degree of coherence between the relationships recovered by five different gene network inference methods. Relationships between transcripts with high linear correlation were common in all of the inferred networks. This was not unexpected, since a simple measure of non-linearity (MI/correlation ratio) demonstrated that non-linear relationships do not make up a large part of either the relationships identified in the literature, or of the two microarray datasets, in agreement with previous findings (64). This suggests that whatever the amount of non-linearity that exists in biology, it is not well represented in these two new datasets, nor in the body of previously published experimental relationships, which in general appear to be enriched for linear relationships. It is possible that linear methods may be sufficient to capture most of the testable relationships in microarray datasets.

All five inference methods tested identify edges in different ways and assume a fundamentally different meaning for an edge. Nevertheless, all methods identified predominately linear relationships. For the ARACNE mutual information networks, since MI is a generalization of correlation, a highly correlated interaction will tend to have high MI and consequently be identified. For the MIKANA dynamical-systems networks inferred from the siRNA data set, a parent that is highly correlated with a particular child RNA is likely to be a good fit to explain much of the variation of that child RNA, so will likely be selected as one of the first parents in the model for that transcript. For the MIKANA dynamical-systems networks inferred from the time course dataset, edges have a different meaning—they imply that the abundance of a parent RNA influences the rate of change of abundance of a child RNA. For these networks, high-correlation edges would not be expected to dominate (and this is the case, see Figure 5B). For the SiGN-BN and BANJO Bayesian methods, a transcript A that is highly correlated with another transcript B is highly predictive of transcript B's behaviour. Some Bayesian network methods have a significant advantage over others in that they are capable of inferring non-linear relationships. Nevertheless, the probability function associated with the abundance of each Bayesian network child node is likely to be especially strongly influenced by input from parent nodes with strong linear relationships (highly correlated) with the child node.

Interestingly, the SiGN-BN Bayesian networks were the only gene networks inferred using bootstrapping, in which 1000 networks were generated from random samples (with replacement) of the 400 microarrays, and edges included if they were present in at least 5% of the inferred networks. We have found previously that the use of bootstrapping allows estimation of the degree of support across the data set for each edge, but does not fundamentally change the nature of the inferred networks (data not shown). The framework we developed here allows for the effects of bootstrapping to be explored using our GenePattern modules, however this does add considerable computing time for each network inference process.

Despite similarity in the types of relationships identified by the different inference methods, we also found that networks inferred by different methods do differ to some degree. Although there was considerable overlap between edge types and in some cases individual edges, hubs in the different networks were frequently different, as were the edges that constituted them, and pathways as measured by shortest path analysis were also different in some cases. This suggests that differences between networks may not always be clear at the level of primary structure (direct connections); methods for visualizing intermediate network structures, such as hubs and pathways, are also required.

The types of edges identified by network inference methods is also likely to be affected by the methods used to select the genes to be included in the network. Recently, this problem may have been overcome, by the development of network inference methods using high-performance computing techniques. For example, Tamada *et al.* (65) have successfully inferred Bayesian networks from >13 000 probes in microarray data generated from the data presented in this study, and have shown that the resulting network replicated many features of smaller networks generated from filtered data, as well as new relationships and hubs missing from the smaller networks. Although the computational power to infer networks from all genes represented on a microarray is not readily available to most researchers due to the computational hardware required, this is an encouraging development that suggests results from smaller studies on a limited subset of probes may still be valid when cells/tissues are considered as whole transcriptome-scale systems.

We were unable to identify any network inference methods that were more efficient than other methods at identifying directional edges from either the siRNA or timecourse data sets. As discussed earlier, we believe that this may be due to the complexity of the 'real biology' that underlies these datasets (47) as well as due to the very limited directional information available in the data for the highly correlated edges that the inference methods tended to identify.

Experimentally verified relationships present in the dataset

Only a small subset of the experimentally verified reference network relationships was present in the inferred networks, or present as high absolute value of correlation or high-MI RNA pairs in the data sets themselves. While this may initially appear to be disappointing it is not very surprising. Both the IPA and BIOBASE databases have been generated using numerous experimental cell, tissue, animal and pathological systems. In contrast, the dataset used here is from a single highly specialized primary cultured human cell type. Endothelial cells are highly differentiated and have a specialized transcriptome, with tightly regulated expression of the RNAs that encode proteins necessary for their specific function. Thus, they may not express the full gamut of RNAs found in less differentiated and other cell types (66) which may mean that potential gene network parents found in reference networks may be absent from the HUVEC data. In addition, the 400 siRNA knockdowns applied to the HUVECs were presumably able to place these cells into only a finite number of different states—additional disruptions and growth conditions may be required to reveal all the RNA-to-RNA relationships involved in the regulation of function in these cells. In the future, it will be interesting to explore whether this complexity can be explicitly modelled in gene networks using mixture modelling approaches such as that taken by Hansen *et al.* (49). The large number of strong RNA-to-RNA relationships found in the literature but not in our dataset, as well as the numerous strong RNA-to-RNA relationships found in our dataset but not in the literature, suggest that systems biology databases constructed from multiple sources, while very useful, may need to be applied to specialized cell types like endothelial cells with caution. The availability of appropriate data sets as described here are likely to facilitate refinement of these data bases.

Generation skip hypothesis

We hypothesized that previously reported relationships involving transcription factors like the Rel/NFκB family may not be recovered by gene network inference methods because the activity of these factors is strongly regulated at the post-translational level, with RNA abundance only making a minor contribution to activity. However, relationships involving transcription factors like the Rel/NFκB may still be identifiable by analysing relationships between upstream regulators and downstream targets. We investigated this ‘generation skip’ hypothesis in the case of *NFKB1*, and found preliminary evidence supporting it; *NFKB1* regulators appeared to share higher correlation overall with *NFKB1* targets than did *NFKB1* itself. Following this proof-of-principle experiment, this hypothesis now needs more rigorous statistical investigation including the investigation of other transcription factor families.

We hope that the principles this publication outlines, as well as the data sets and the software resources it provides, closes some of the current gaps that prevent gene network inference from becoming a commonly used tool in biological research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Files S1–S6.

ACKNOWLEDGEMENTS

We wish to thank the patients and staff at Rosie Hospital, Cambridge, UK, for their donation and collection of umbilical cords used to isolate HUVECs. Computation for the SiGN-BN method was provided by the Super Computer System, Human Genome Centre, Institute of Medical Science, University of Tokyo. We wish to thank Manjunath Kustagi and Andrea Califano of the Califano lab at Columbia University for their assistance with configuration and optimization for the ARACNE algorithm, and Alexander Hartemink and Eric Jarvis of Duke University for their assistance with configuration and optimization for the BANJO algorithm. Generation of the microarray data was supported by research collaboration agreements between the University of Cambridge and GNI Ltd. We also wish to thank Atsushi Doi and Masao Nagasaki for assistance with gene network digraph mapping using the Cell Illustrator application.

FUNDING

Research collaboration agreements between the University of Cambridge and GNI Ltd for generation of the endothelial cell data used in this work, and the implementation of Bayesian gene network inference methods; the University of Auckland Doctoral Scholarship (to D.H.); Maurice Wilkins Centre for Molecular Biodiscovery Studentship (to D.H.); Auckland Bioengineering Institute (to D.H.); Marsden Fund of the Royal Society of New Zealand (06-UOA-182 to E.J.C.); University of Auckland Early Career Research Excellence Award (#3606318 to E.J.C.); Health Research Council of New Zealand International Investment Opportunities Fund (06/581 to C.P.); University of Auckland Faculty of Medical and Health Sciences Research Development Fund (# 8303460 to C.P.); University of Auckland Vice-Chancellor’s Development Fund (# 23285 to C.P.); Maurice Wilkins Centre of Research Excellence Fund. Funding for open access charge: Departmental Grant funding.

Conflict of interest statement. H.A., Y.T., K.Y. and C.S. are past employees of GNI Ltd., a Japanese company which supported the generation of microarray data used in this study. Y.T., B.D., D.S., S.H., M.A., S.I., K.Y., C.S., S.S., S.K., S.M., D.S.C.-J. and C.P. have held/hold stock or stock options in GNI Ltd. None of the authors are current employees of GNI Ltd.

REFERENCES

1. Andreopoulos,B., An,A.J., Wang,X.G. and Schroeder,M. (2009) A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform.*, **10**, 297–314.
2. Clarke,R., Ressom,H.W., Wang,A.T., Xuan,J.H., Liu,M.C., Gehan,E.A. and Wang,Y. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer*, **8**, 37–49.
3. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**.
4. Marbach,D., Prill,R.J., Schaffter,T., Mattiussi,C., Floreano,D. and Stolovitzky,G. (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl Acad. Sci. USA*, **107**, 6286–6291.
5. Penfold,C.A. and Wild,D.L. (2011) How to infer gene networks from expression profiles, revisited. *Interface Focus*, **1**, 857–870.
6. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64–68.
7. Ma,H.W., Kumar,B., Ditges,U., Gunzer,F., Buer,J. and Zeng,A.P. (2004) An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res.*, **32**, 6643–6649.
8. Kohanski,M.A., Dwyer,D.J., Wierzbowski,J., Cottarel,G. and Collins,J.J. (2008) Mistranslation of membrane proteins and two-component system activation trigger antibiotic-mediated cell death. *Cell*, **135**, 679–690.
9. Yoon,H.J., McDermott,J.E., Porwollik,S., McClelland,M. and Heffron,F. (2009) Coordinated regulation of virulence during systemic infection of *Salmonella enterica* Serovar Typhimurium. *PLoS Pathog.*, **5**, e1000306.

10. Bonneau,R., Facciotti,M.T., Reiss,D.J., Schmid,A.K., Pan,M., Kaur,A., Thorsson,V., Shannon,P., Johnson,M.H., Bare,J.C. *et al.* (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell*, **131**, 1354–1365.
11. Basso,K., Margolin,A.A., Stolovitzky,G., Klein,U., Dalla-Favera,R. and Califano,A. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.
12. Belcastro,V., Siciliano,V., Gregoretti,F., Mithbaokar,P., Dharmalingam,G., Berlingieri,S., Iorio,F., Oliva,G., Polishchuck,R., Brunetti-Pierri,N. *et al.* (2012) Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. *Nucleic Acids Res.*, **40**, D715–D719.
13. Ljung,L. (1999) *System Identification: Theory for the User*, 2nd edn. Prentice Hall, Upper Saddle River, NJ.
14. Stolovitzky,G., Prill,R.J. and Califano,A. (2009) Lessons from the DREAM2 Challenges. *Ann. NY Acad. Sci.*, **1158**, 159–195.
15. Marbach,D., Schaffter,T., Mattiussi,C. and Floreano,D. (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.*, **16**, 229–239.
16. Gardner,T.S., di Bernardo,D., Lorenz,D. and Collins,J.J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.
17. Pisarev,A., Poustelnikova,E., Samsonova,M. and Reinitz,J. (2009) FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Res.*, **37**, D560–D566.
18. Shi,L.M., Reid,L.H., Jones,W.D., Shippy,R., Warrington,J.A., Baker,S.C., Collins,P.J., de Longueville,F., Kawasaki,E.S., Lee,K.Y. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.
19. Scheinine,A., Mentzen,W.I., Fotia,G., Pieroni,E., Maggio,F., Mancosu,G. and Fuente,A.D.L. (2009) Inferring gene networks: dream or nightmare? *Ann. NY Acad. Sci.*, **1158**, 287–301.
20. Faith,J.J., Hayete,B., Thaden,J.T., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J.J. and Gardner,T.S. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, 54–66.
21. Yu,J., Smith,V.A., Wang,P.P., Hartemink,A.J. and Jarvis,E.D. (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.
22. Imoto,S., Goto,T. and Miyano,S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pac. Symp. Biocomput.*, **7**, 175–186.
23. Kim,S., Imoto,S. and Miyano,S. (2004) Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, **75**, 57–65.
24. Imoto,S., Tamada,Y., Araki,H., Yasuda,K., Print,C.G., Charnock-Jones,S.D., Sanders,D., Savoie,C.J., Tashiro,K., Kuhara,S. *et al.* (2006) Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. *Pac. Symp. Biocomput.*, 559–571.
25. Wildenhain,J. and Crampin,E.J. (2006) Reconstructing gene regulatory networks: from random to scale-free connectivity. *Syst. Biol.*, **153**, 247–256.
26. Srividhya,J., Crampin,E.J., McSharry,P.E. and Schnell,S. (2007) Reconstructing biochemical pathways from time course data. *Proteomics*, **7**, 828–838.
27. Della Gatta,G., Bansal,M., Ambesi-Impiombato,A., Antonini,D., Missero,C. and di Bernardo,D. (2008) Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome Res.*, **18**, 939–948.
28. Werhli,A.V., Grzegorczyk,M. and Husmeier,D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**, 2523–2531.
29. Soranzo,N., Bianconi,G. and Altafini,C. (2007) Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, **23**, 1640–1647.
30. Bansal,M. and di Bernardo,D. (2007) Inference of gene networks from temporal gene expression profiles. *IET Syst. Biol.*, **1**, 306–312.
31. Lauria,M., Iorio,F. and Bernardo,D.D. (2009) NIRest: a tool for gene network and mode of action inference. *Ann. NY Acad. Sci.*, **1158**, 257–264.
32. Cosgrove,E.J., Zhou,Y., Gardner,T.S. and Kolaczyk,E.D. (2008) Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics*, **24**, 2482–2490.
33. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Dalla Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**.
34. Tamada,Y., Kim,S., Bannai,H., Imoto,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19(Suppl. 2)**, 227–236.
35. Ergun,A., Lawrence,C.A., Kohanski,M.A., Brennan,T.A. and Collins,J.J. (2007) A network biology approach to prostate cancer. *Mol. Syst. Biol.*, **3**, 6.
36. Tamada,Y., Araki,H., Imoto,S., Nagasaki,M., Doi,A., Nakanishi,Y., Tomiyasu,Y., Yasuda,K., Dunmore,B., Sanders,D. *et al.* (2009) Unraveling dynamic activities of autocrine pathways that control drug-response transcriptome networks. *Pac. Symp. Biocomput.*, **14**, 251–263.
37. Franke,L., van Bakel,H., Fokkens,L., de Jong,E.D., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
38. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
39. Oinn,T., Greenwood,M., Addis,M., Alpdemir,M.N., Ferris,J., Glover,K., Goble,C., Goderis,A., Hull,D., Marvin,D. *et al.* (2006) Taverna: lessons in creating a workflow environment for the life sciences. *Concurr. Comput. Pract. Exp.*, **18**, 1067–1100.
40. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
41. Floratos,A., Smith,K., Ji,Z., Watkinson,J. and Califano,A. (2010) geWorkbench: an open source platform for integrative genomics. *Bioinformatics*, **26**, 1779–1780.
42. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
43. Bo,T.H., Dysvik,J. and Jonassen,I. (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.*, **32**, e34.
44. Wingender,E. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.*, **9**, 326–332.
45. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
46. Denk,A., Goebeler,M., Schmid,S., Berberich,I., Ritz,O., Lindemann,D., Ludwig,S. and Wirth,T. (2001) Activation of NF-kappa B via the I kappa B kinase complex is both essential and sufficient for proinflammatory gene expression in primary endothelial cells. *J. Biol. Chem.*, **276**, 28451–28458.
47. Clark,K., Peggie,M., Plater,L., Sorcek,R.J., Young,E.R.R., Madwed,J.B., Hough,J., McIver,E.G. and Cohen,P. (2011) Novel cross-talk within the IKK family controls innate immunity. *Biochem. J.*, **434**, 93–104.
48. Daub,C.O., Steuer,R., Selbig,J. and Kloska,S. (2004) Estimating mutual information using B-spline functions - an improved

similarity measure for analysing gene expression data. *BMC Bioinform.*, **5**.

49. Hansen,M., Everett,L., Singh,L. and Hannenhalli,S. (2010) Mimosa: mixture model of co-expression to detect modulators of regulatory interaction. *Algorithms Mol. Biol.*, **5**.
50. Lee,H.K., Hsu,A.K., Sajdak,J., Qin,J. and Pavlidis,P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.
51. Han,J.D.J., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J.M., Cusick,M.E., Roth,F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
52. Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.
53. Green,T. (1996) Haematopoiesis - master regulator unmasked. *Nature*, **383**, 575–577.
54. Jiang,W., Li,X., Rao,S.Q., Wang,L.H., Du,L., Li,C.X., Wu,C., Wang,H.Z., Wang,Y.D. and Yang,B.F. (2008) Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. *BMC Syst. Biol.*, **2**, 72.
55. Goh,K.I., Cusick,M.E., Valle,D., Childs,B., Vidal,M. and Barabasi,A.L. (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.
56. Yan,K.-K., Fang,G., Bhardwaj,N., Alexander,R.P. and Gerstein,M. (2010) Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc, Natl Acad. Sci USA*, **107**, 9186–9191.
57. Perkins,N.D. (2006) Post-translational modifications regulating the activity and function of the nuclear factor kappa B pathway. *Oncogene*, **25**, 6717–6730.
58. Viatour,P., Merville,M.P., Bours,V. and Chariot,A. (2005) Phosphorylation of NF-kappa B and I kappa B proteins: implications in cancer and inflammation. *Trends Biochem. Sci.*, **30**, 43–52.
59. De Bosscher,K., Vanden Berghe,W. and Haegeman,G. (2006) Cross-talk between nuclear receptors and nuclear factor κB. *Oncogene*, **25**, 6868–6886.
60. Stark,L.A. and Dunlop,M.G. (2005) Nucleolar sequestration of RelA (p65) regulates NF-kappa B-driven transcription and apoptosis. *Mol. Cell. Biol.*, **25**, 5985–6004.
61. Saccani,S., Pantano,S. and Natoli,G. (2003) Modulation of NF-kappa B activity by exchange of dimers. *Mol. Cell.*, **11**, 1563–1574.
62. Ladunga,I., Everett,L., Hansen,M. and Hannenhalli,S. (2010) Regulating the regulators: modulators of transcription factor activity. In: Ladunga,I. (ed.), *Computational Biology of Transcription Factor Binding*, Vol. 674. Humana Press, NY, pp. 297–312.
63. Everett,L., Vo,A. and Hannenhalli,S. (2009) PTM-Switchboard - a database of posttranslational modifications of transcription factors, the mediating enzymes and target genes. *Nucleic Acids Res.*, **37**, D66–D71.
64. Steuer,R., Kurths,J., Fiehn,O. and Weckwerth,W. (2003) Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019–1026.
65. Tamada,Y., Imoto,S., Araki,H., Nagasaki,M., Print,C., Charnock-Jones,S. and Miyano,S. (2010) Estimating genome-wide gene networks using nonparametric Bayesian network models on massively parallel computers. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **8**, 683–697.
66. Thorrez,L., Laudadio,I., Van Deun,K., Quintens,R., Hendrickx,N., Granvik,M., Lemaire,K., Schraenen,A., Van Lommel,L., Lehnert,S. *et al.* (2010) Tissue-specific disallowance of housekeeping genes: the other face of cell differentiation. *Genome Res.*, **21**, 95–105.