Bioinformatics, 36(8), 2020, 2515–2521
doi: 10.1093/bioinformatics/btz949
Advance Access Publication Date: 24 December 2019
Original Paper

Systems biology

Causal network perturbations for instance-specific analysis of single cell and disease samples

Kristina L. Buschur[1,2], Maria Chikina[1] and Panayiotis V. Benos[1,*]

[1]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA and [2]Joint CMU-Pitt PhD Program in Computational Biology, Pittsburgh, PA 15260, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on June 20, 2019; revised on November 22, 2019; editorial decision on December 18, 2019; accepted on December 19, 2019

Abstract

Motivation: Complex diseases involve perturbation in multiple pathways and a major challenge in clinical genomics is characterizing pathway perturbations in individual samples. This can lead to patient-specific identification of the underlying mechanism of disease thereby improving diagnosis and personalizing treatment. Existing methods rely on external databases to quantify pathway activity scores. This ignores the data dependencies and that pathways are incomplete or condition-specific.

Results: ssNPA is a new approach for subtyping samples based on *deregulation* of their gene networks. ssNPA learns a causal graph directly from control data. Sample-specific network neighborhood deregulation is quantified via the error incurred in predicting the expression of each gene from its Markov blanket. We evaluate the performance of ssNPA on liver development single-cell RNA-seq data, where the correct cell timing is recovered; and two TCGA datasets, where ssNPA patient clusters have significant survival differences. In all analyses ssNPA consistently outperforms alternative methods, highlighting the advantage of network-based approaches.

Availability and implementation: http://www.benoslab.pitt.edu/Software/ssnpa/.

Contact: benos@pitt.edu

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

1 Introduction

Gene expression profiling by RNA-sequencing has become routine tool in biomedical research. Similarly, on the clinical side, RNA-seq has now been introduced as a cost-effective diagnostic tool (Cummings *et al.*, 2017; Kremer *et al.*, 2017). Moreover, recent technological advances have made the assessment of gene expression at single-cell level (scRNA-seq) feasible, opening new avenues to developmental biology and the study of dynamic networks (Chen *et al.*, 2018; Villani *et al.*, 2017; Zhao *et al.*, 2018). Consequently, the number of large RNA-seq datasets keeps growing with hundreds or thousands of samples representing a single clinical or cellular condition. As a result, the scientific questions have shifted away from simple differential expression to characterizing the molecular heterogeneity of disease phenotypes. One simple way to characterize sample heterogeneity is via clustering and/or dimensionality reduction. This approach will often reveal distinct sample groups within the population but ignores the fact that genes are organized in regulatory networks. On the other end of the spectrum, there has been considerable development in methods that quantify pathway activation on a single sample level [ssGSEA (Barbie *et al.*, 2009), PLAGE (Tomfohr *et al.*, 2005), GSVA (Hanzelmann *et al.*, 2013), Pathifier (Drier *et al.*, 2013)]. However, these methods rely heavily on existing pathway information (e.g. from KEGG, BioCarta, The Nature Pathway Interaction Database), which may be incomplete, not well annotated or irrelevant to the studied phenotype or condition. A related method, N-of-1-pathways (Gardeux *et al.*, 2014; Li *et al.*, 2017a), predicts deregulated pathways from a single patient but requires multiple measurements for each patient (e.g. matched cancer-control), which are not usually available; and not applicable to scRNA-seq data. Other methods (e.g. Mohammadi *et al.*, 2018) quantify a sample-to-sample similarity aiming to identify similarities and differences between cell functions.

In this article, we present a different approach for assessing, qualitatively and quantitatively, how the gene network of a set of control samples is perturbed in a newly presented (query) sample. Our approach, *Single Sample Network Perturbation Assessment* (ssNPA), uses causal modeling to first learn the gene expression interaction network from a set of reference samples; and for each query sample the method assesses the part(s) of the 'reference sample network' which are deregulated. The rationale behind this is that in many diseases an observed phenotype may be due to changes in different parts of the 'healthy' gene network.

Causal graphs have been used in the past to learn gene networks from expression data (Friedman, 2004; Sachs, 2005; Sedgewick *et al.*, 2016) or gene features that are highly predictive of certain phenotypes (Huang *et al.*, 2015; Raghu *et al.*, 2018a,b,c; Sedgewick *et al.*, 2019). ssNPA learns a causal graph from expression data and for every gene it builds a predictive model based on its Markov blanket. Applying the model to a new sample produces a vector of residuals which quantifies the network level gene dysregulation. These vectors can also be used to cluster samples into groups and assess their group characteristics (e.g. developmental time, survival, molecular mechanisms of phenotype, etc.) or to assign an individual patient to a disease subcluster. We use this property to evaluate ssNPA on existing datasets. Specifically, we show that ssNPA separates well the mouse liver developmental trajectory and cell types in a scRNA-seq dataset (Yang *et al.*, 2017). We also use RNA-seq data from The Cancer Genome Atlas (TCGA) (lung and breast cancer datasets) (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2012) to demonstrate that ssNPA-identified subtypes have better accuracy than alternative approaches with respect to patient survival and molecular subtype.

2 Materials and methods

Datasets. A murine liver cell development scRNA-seq dataset (Yang *et al.*, 2017) was obtained from GEO (GSE90047). The dataset consists of gene expression measurements in 447 cells over the course of embryonic days E10.5–E17.5. Cells were first sorted with fluorescence-activated cell sorting (FACS) according to the cell surface markers Delta-like (DLK) to identify hepatocytes and epithelial cell adhesion molecule (EpCAM) to identify cholangiocytes. scRNA-seq counts were used as input to ssNPA.

The Cancer Genome Atlas (TCGA) RNA-seq data from breast invasive carcinoma (BRCA) and lung adenocarcinoma (LUAD) were downloaded from the Broad Firehose (Broad Institute Cancer Genome Analysis group, https://gdac.broadinstitute.org). The BRCA dataset consists of RNA-seq normalized gene counts for 1100 cancer samples and 112 normal samples (Cancer Genome Atlas Network, 2012). The LUAD dataset consists of RNA-seq normalized gene counts of 517 cancer samples and 59 normal samples (Cancer Genome Atlas Research Network, 2012).

Data preprocessing. Lowly expressed genes were filtered out with the filterByExpr function (edgeR v. 3.26.8)(Robinson *et al.*, 2010), and the RNA-seq counts were transformed to log2 counts per million through mean-variance modeling by the voom function (Limma v. 3.40.06) (Ritchie *et al.*, 2015). For speed and accuracy, we selected the top 3000 most variant genes for each dataset for input to ssNPA.

Sample clustering. In order to better evaluate the efficiency of the various methods for single sample subtyping, we performed sample clustering using Seurat (Butler *et al.*, 2018), and we examined various external characteristics of the clusters. Samples were clustered in their feature space. First, the samples are projected into principal component space. The number of principal components to retain in the projection is determined heuristically by identifying the elbow of the scree plot. Then clustering is performed with a graph-based clustering that constructs the shared nearest neighbor graph and then optimizes the modularity function (Waltman and van Eck, 2013). Finally, the clusters are visualized with a nonlinear dimensionality reduction (t-SNE) (van der Maaten and Hinton, 2008).

Method comparison. ssNPA methods were compared to Pathifier (Drier *et al.*, 2013) and single sample gene set enrichment analysis (ssGSEA) (Barbie *et al.*, 2009). All methods were tested on the same input data and reference sample selections. For Pathifier, we provided gene lists for all KEGG pathways provided by KEGGRest in R (Tenenbaum, 2016) for the appropriate organism, and used the R implementation of Pathifier with the quantify_pathways_deregulation() function and default parameters. For ssGSEA, we used the gene sets from the C2 collection of the Molecular Signatures Database (Subramanian *et al.*, 2005) version 7.0 for the TCGA datasets and version 5.2 with mouse identifiers downloaded from http://bioinf.wehi.edu.au/software/MSigDB/index.html for the murine liver cell development scRNA-seq dataset. We applied the implementation of ssGSEA provided within the GSVA() function of the GSVA R package with default parameters. For fairness, we use 10 principal components for clustering with each method.

3 Results

3.1 Description of ssNPA and ssNPA-LR algorithms

ssNPA learns the global gene expression network as a directed (causal) graph from a set of reference samples using the Fast Greedy Equivalent Search (FGES) algorithm (Ramsey *et al.*, 2017). FGES calculates a directed acyclic graph (DAG) over all data by maximizing the Bayesian Information Criterium (BIC) score of the data given the model (network). The BIC score is given by the formula:

$$\mathrm{BIC} = -2 \cdot \mathcal{L}(\mathcal{D}) + \mathrm{PD} \cdot \mathrm{df} \cdot \ln n \qquad (1)$$

where $\mathcal{L}(\mathcal{D}) = lnP(\mathcal{D}|\theta, \mathcal{M})$ is the maximum log-likelihood of the data (D) given the structural model (M) and its parameters (θ); PD is a penalty value (‘penalty discount’) that controls sparsity (PD = 1 in the standard BIC definition); df is the degrees of freedom; and n is the sample size. This score is decomposable and the total BIC of the graph is the sum of the BIC of its nodes and their parents. FGES starts with an empty graph then adds single edges while the BIC score increases. Next, the algorithm removes single edges while the BIC score increases.

For DAGs, the Markov blanket of a gene G_i (MB(G_i)) consists of the parents, children and spouses of G_i in the graph. If the graph contains also undirected edges then the Markov blanket includes also nodes on those edges, if they could potentially be part of the Markov blanket in the directed graph. Once the graph has been learned from the reference (control) samples, then ssNPA uses the Markov blanket around each gene, G_i, to build a predictor of its expression. This is because in this type of directed graph every variable G_i is independent of any variable that does not belong to the MB(G_i), conditioned on the MB(G_i). Therefore, a highly predictive regression model can be learned for the expression of each gene:

$$G_i = \beta_{0,i} + \sum\nolimits_{G_k \in \mathrm{MB}(G_i)} \beta_{k,i} \cdot G_k + \varepsilon \qquad (2)$$

Then for each new sample this model can be used to calculate the deviation of the expression of G_i in this sample compared to the reference samples. We are only interested in the magnitude of the deviation, so we calculate the squared regression residual:

$$\left(G_i - \beta_{0,i} + \sum\nolimits_{G_k \in \mathrm{MB}(G_i)} \hat{\ } \beta_{k,i} \cdot G_k \right)^2 \qquad (3)$$

where G_i is the observed gene expression value of gene i in the query sample. Thus, the new sample can be represented as a vector of deviations of expression of every gene from the reference samples. Given that genes are connected through the network of interactions, in this way, we assess both the topology and the magnitude of network perturbations. The idea behind this approach is that diseases are usually defined by symptoms, but the underlying molecular mechanisms may differ from patient to patient. In other words, a group of patients may have perturbations in subnetwork A, another in subnetwork B, etc. (Fig. 1). Identifying which part of the reference network is perturbed in each patient can provide insights on the mechanisms of disease and can be used to identify new subphenotypes.

For comparison purposes, we also implemented ssNPA-LR, in which causal learning is substituted by lasso regression, resulting in an undirected graph. ssNPA and ssNPA-LR analysis procedures have the following steps:

1. *Reference samples.* For disease data, we used the controls as reference sample set. For the liver scRNA-seq, we tested each developmental stage (as determined by external cell markers) individually and all together as potential reference groups.
2. *Gene network learning (ssNPA).* A directed graph is learned from expression data from the reference group (FGES algorithm)

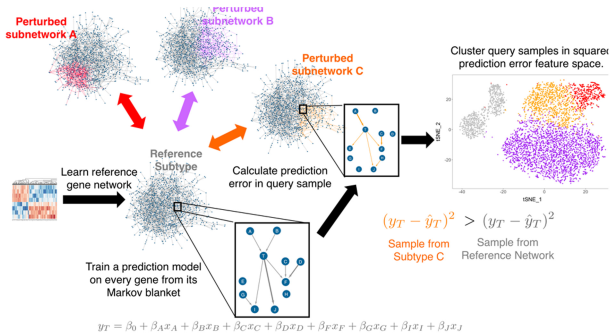
Perturbed subnetwork A
Perturbed subnetwork B
Perturbed subnetwork C
Cluster query samples in squared prediction error feature space.
Reference Subtype
Learn reference gene network
Calculate prediction error in query sample
Train a prediction model on every gene from its Markov blanket
A
B
T
C
D
E
F
G
H
I
J
tSNE_1
tSNE_2
-40
-20
0
20
$(y_T - \hat{y}_T)^2 > (y_T - \hat{y}_T)^2$
Sample from Subtype C
Sample from Reference Network
$y_T = \beta_0 + \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_D x_D + \beta_F x_F + \beta_G x_G + \beta_I x_I + \beta_J x_J$

Fig. 1. Overview of ssNPA. Gene interaction network is learned as directed graph from reference samples and a regression model is trained for each gene using its Markov blanket as predictors. For each new query sample, the expression of each gene is predicted from its model and squared prediction errors (deviations) calculated between this and the observed values. Query samples may deviate from the reference network at different subnetworks (**A**, **B**, **C**, etc.). Clustering can then be performed on the deviation space to produce sample subtypes with deregulated subnetworks

(Ramsey *et al.*, 2017). For this work, we scan over a number of PD values in the range [4, 12], and we choose a PD for each dataset that balances grouping the reference samples together while not overfitting. The Markov blanket around every gene in this network is used for predicting its expression on any given sample with a linear model, and the deviation from the observed value is a measure of network perturbation.

3. *Feature selection (ssNPA-LR)*. In this case, we used the glmnet package in R (v. 2.0.18) to learn a lasso regression prediction model for every gene across the reference samples (Friedman *et al.*, 2010). We chose each sparsity parameter (λ) with 10-fold cross-validation, selecting the value of λ corresponding to minimum mean cross-validated error.

3.2 ssNPA correctly identifies embryonic stage and cell type in murine liver cells from single-cell RNA-seq data

We used a recently published liver development scRNA-seq dataset to test ssNPA and compare it to other methods. This dataset is composed of multiple types of liver cells samples at a series of developmental timepoints. The early hepatoblast cell differentiates into two lineages (hepatoblasts and cholangiocytes). In this dataset, the time point and cell-identity are experimentally controlled and thus can serve as the ground truth. We hypothesize that information regarding the cell-type and developmental stage is reflected in the gene expression data and leveraging information about gene regulatory network deregulation can improve separation according to these classes. To quantitatively compare the clustering performances of all methods, we used the normalized mutual information (NMI) and adjusted Rand index (ARI) to assess how well the cluster assignments match true cell-type and developmental stage classes.

When gene expression data used directly for clustering (Waltman, 2013), we identified five clusters (NMI = 0.50, Supplementary Fig. S1A). These clusters separated well the extreme developmental time points: cells measured at day E10.5 and hepatocytes from day E17.5 (Fig. 2A). However, all of the differentiated cholangiocytes were grouped together in a single cluster and although they were somewhat stratified within the cluster, their embryonic stage was not distinguishable. The remaining clusters contained a mix of intermediate timepoints (days) and hence they did not accurately represent the developmental trajectory.

By contrast, the eight identified clusters based on the network perturbation features of ssNPA (Supplementary Fig. S1B) separated well all stages (NMI = 0.62, Fig. 2B). In particular, hepatoblasts from days E10.5 and E11.5 as well as mature cholangiocytes and E17.5 hepatocytes were separated into four distinct clusters. Because the E15.5 cells were used as the reference set, these cells were well-separated into two clusters, one comprised of the hepatoblasts and another comprised of the cholangiocytes. Since there was not an obvious reference set of samples in this dataset, we examined the utility of each group as a potential reference. We additionally evaluated a range of PD (penalty discount) parameter values for FGES [4, 12]. We found the late intermediate stages (E14.5 or E15.5) to show better performance than the extremes when they were used as reference set (Supplementary Fig. S2). We also observed that performance increase with smaller PD values, although we expect this to be a dataset-dependent effect.

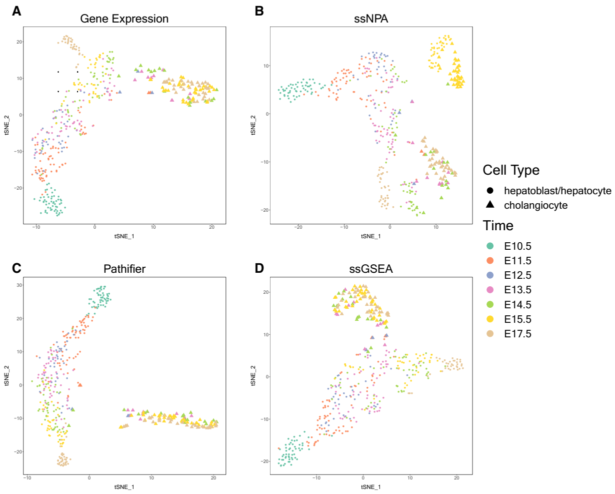

Fig. 2. Comparison of how well (**A**) gene expression, (**B**) ssNPA, (**C**) Pathifier and (**D**) ssGSEA separate murine liver cell scRNA-seq samples by developmental stage and cell type. ssNPA was used with the E15.5 cells as the reference set and PD = 4. Pathifier was applied with the E13.5 cells as the reference set. Clustering for every method was performed with the first 10 principal components

We further evaluated the importance of the reference group selection by choosing random groups of cells to use as the reference. We generated five separate lists of the 70 cells (the median number of cells across time points) randomly chosen from the 447 cells in the dataset and used these as the reference group with ssNPA. These reference groups did not lead to good separation of the cells according to timepoint and cell type, and notably the performance decreased sharply with lower PD values (Supplementary Fig. S2). This suggested that fitting the network to a group of cells that do not share similar gene regulatory structure was very detrimental to ssNPA performance, and thus choosing a group of reference samples that we expect to have similar regulatory patterns is the more important consideration compared to which particular group we choose.

scRNA-seq is still a developing technology and many of scRNA-seq experiments do not provide as high-quality data as the dataset we have used for benchmarking here. To simulate this phenomenon, we downsampled the number of genes and then the number of cells we use as input to ssNPA (Supplementary Fig. S3). Using E15.5 as the reference group and PD = 4, we observed that ssNPA performance was relatively robust with at least 85% of the 3000 genes selected (mean NMI = 0.58) but began to decrease by 80% (mean NMI = 0.56) and more steeply by 50% (mean NMI = 0.52). With only 25% of the genes, performance was very poor (mean NMI = 0.39). Similarly, we simulated smaller datasets with fewer cells by downsampling but maintaining the same proportions of cells in each timepoint and cell-type category. Performance bottomed out with 80% of cells (mean NMI = 0.50). It was not feasible to test smaller subsets with 75% of cells or less. As the number of cells decreased, any given timepoint did not include enough cells to use as an appropriate reference group for the network-learning step of ssNPA.

Next, we compared ssNPA to Pathifier and ssGSEA. Both methods quantify gene interactions but require pathway information from an external database. For Pathifier, we used the KEGG pathway database (Kanehisa and Goto, 2000). Using the cells from

Table 1. Comparison of different feature calculation methods

Method	NMI	ARI	Avg no. feat.
ssNPA	0.62	0.49	3.6
Gene expression	0.50	0.37	NA
Pathifier	0.53	0.40	NA
ssGSEA	0.48	0.35	NA
ssNPA-LR	0.60	0.48	13.5

Note: Clustering for every method was performed with the first 10 principal components. E15.5 were used as reference cells for ssNPA and E13.5 were used as the reference for Pathifier and ssNPA-LR. PD = 4 for ssNPA.

NMI, normalized mutual information; ARI, adjusted Rand Index.

E13.5 as the reference group with Pathifier led to the best performance (NMI = 0.53), but the performance was quite consistent across all the timepoints we tested as reference groups and even with randomly chosen reference cells (Supplementary Fig. S4). With E13.5 as the reference group, Pathifier produced six distinct clusters (Supplementary Fig. S1C), but with the exception of E10.5 and E17.5, it did not separate the developmental stages very well (Fig. 2C). The cholangiocytes from all stages were grouped together, but all of the intermediate stage hepatocytes were mixed together and distributed across three clusters. Furthermore, the runtime of Pathifier was very long compared to ssNPA (on the order of hours compared to minutes). Finally, we tested ssGSEA which calculates a gene set enrichment score for every sample. We used ssGSEA with default parameters and the gene sets from the C2 collection of the Molecular Signatures Database version 5.2 with mouse identifiers (Subramanian *et al.*, 2005). ssGSEA does not require the user to provide a reference set. Clustering with the ssGSEA produced five clusters (Supplementary Fig. S1D), but in general, these were not well-separated according to developmental time point (NMI = 0.48, Fig. 2D). The cholangiocytes were grouped into one cluster and the hepatocytes from E10.5 and E17.5 were each in their own cluster. However, the remaining hepatoblasts/hepatocytes spanning E11.5–E15.5 were mixed together and divided between two clusters.

We additionally developed and tested a variation of ssNPA, the ssNPA-LR algorithm, which uses lasso regression instead of causal network learning to choose the features predicting the expression of a gene (Supplementary Fig. S5A). We found seven clusters (Supplementary Fig. S5B), which separated well both the early and late developmental stages hepatocytes and the cholangiocytes, as well as the E13.5 cells which were used as the reference group (NMI = 0.60). The other three clusters contained more of a mix of the intermediate timepoint hepatocytes. Although using E13.5 as the reference group led to highest NMI, E14.5 and E15.5 had almost identical performance and all timepoint groups were mostly consistent (Supplementary Fig. S6). As with ssNPA, however, choosing random cells as the reference group did lead to poor clustering results (mean NMI = 0.33).

Table 1 presents all the clustering performance results for the various methods we tested. We found that ssNPA and ssNPA-LR clearly outperform Pathifier, ssGSEA and gene expression by maximizing NMI (0.62 and 0.60, respectively). ssNPA also returned the highest ARI of these methods (0.49). However, we note a strong advantage to ssNPA over ssNPA-LR when we consider how many genes they utilized. On average, ssNPA used only 3.6 predictors for every gene, while ssNPA-LR needed 13.5 genes. This could suggest that the lasso regression step is overfitting compared to feature selection by causal network which jointly models the expression of all 3000 genes. Additionally, the runtime for ssNPA-LR is much slower than for ssNPA because of the cross-validation step needed to select a lasso regression sparsity parameter for every gene.

Finally, we note that the ssNPA gene network deregulation features offer the additional benefit of being directly interpretable and can identify which points in the network are being deregulated in ways that lead to different subtypes. In order to investigate which genes contributed to cluster identification, we used the magnitude of the PCA loadings for the principal components used in clustering (Supplementary Table S1). These genes with the highest loadings are the ones whose network is most deregulated in all the cells compared to the reference group (E15.5) and are differentially regulated across development and differentiation. Mbd3 encodes a transcription factor and was a top gene for PC 1. It is known to be involved in the nucleosome remodeling deacetylase (Mi-2/NuRD) corepressor complex in mice (Hendrich *et al.*, 2001) and in separate, recent scRNA-seq study of murine liver cell development was linked to hepatoblast-specific network regulation (Su *et al.*, 2017). Additionally, the human ortholog MBD3 plays an important role in TGFβ/Smad signaling during the epithelial-mesenchymal transition in pancreatic cancer (Xu *et al.*, 2017) and inhibits the formation of liver cancer stem cells (Li *et al.*, 2017b). Fubp3 is another gene encoding a transcription factor, FBP, that had a top loading for PC 1. FBP is an important regulator of Myc (He *et al.*, 2000), which in turn is one of the most important regulators of cell differentiation. FBF knockout in mice is embryonic lethal from E10.5 to birth, and FBP loss is associated with pale liver, trilineage anemia and number of other severe phenotypes in mice (Zhou *et al.*, 2016). These results lend support to our finding that differential regulation of genes like Mbd3 and Fubp3 occurs throughout liver development and demonstrate how ssNPA can directly highlight some of the most important effectors in a gene regulatory network.

3.3 ssNPA clusters in breast cancer samples matching the molecular subtypes and significantly differ in survival

We also applied the various methods on breast cancer RNA-seq data from tissues of known subtype. Our dataset was comprised of 127 basal, 41 HER2+, 488 luminal A and 144 luminal B samples. ssNPA identified five clusters that were significantly associated with molecular subtype ($\chi^2 = 609.52$· $P < 2.2\text{e-}16$). The majority of the basal tumor samples (85%) were grouped together in cluster 3 (Fig. 3A and Supplementary Fig. S7A). The two luminal subtypes were not completely separated; 88% of the luminal A samples were distributed evenly among the final three clusters (0, 1 and 2), and two of these clusters (0 and 1) also contained a large number of the luminal B samples (51%). However, an additional 35% of the luminal B samples were in cluster 4, which also contained 73% of the HER2+ samples but only 8% of the luminal A samples. Additionally, the five ssNPA clusters were significantly associated with survival ($P = 0.0015$), which further suggests that ssNPA is able to use information about gene network deregulation to separate the BRCA tumor samples in a clinically meaningful way (Fig. 3D and Supplementary Fig. S8A).

The other methods produced a similar result with respect to molecular subtype clustering (Fig. 3 and Supplementary Figs S7 and S8). Pathifier identified nine clusters that were also significantly associated with molecular subtype ($\chi^2 = 596.23$, $P < 2.2\text{e–}16$). These clusters were also significantly associated with survival, but with a slightly larger *P*-value than ssNPA ($P = 0.0021$). Finally, ssGSEA identified four clusters which were also significantly associated with molecular subtype ($\chi^2 = 535.62$, $P < 2.2\text{e-}16$). However, these clusters did not significantly associate with survival ($P = 0.18$). Altogether, we see that ssNPA can perform similarly well to comparable methods in separating breast cancer samples according to their molecular subtypes with the additional advantage of separating samples according to high-level clinical endpoints such as patient survival.

Finally, we again note that the ssNPA gene network deregulation features can directly offer insight into where the gene regulatory networks differ across clusters. The genes with the top PC loadings are those whose deregulation is most responsible for separating the BRCA sample clusters (Supplementary Table S2). The top gene for PC 1 was ITM2A, and a recent study demonstrated that the downregulated expression of this gene is associated with breast cancer and poor survival outcomes (Zhou *et al.*, 2019). Additionally, the authors found that overexpression of ITM2A significantly inhibited breast cancer cell proliferation and was involved in positive regulation of autophagy. We observed a similar trend in our results;

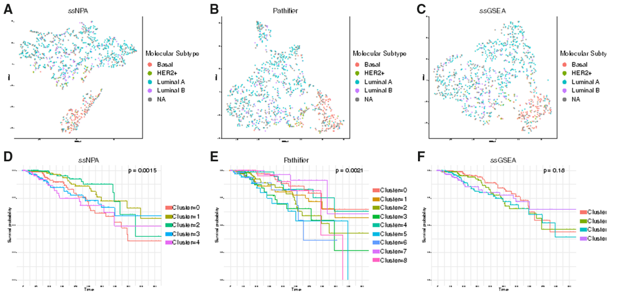

Fig. 3. Separation of breast cancer RNA-seq samples according to tumor molecular subtype by (**A**) ssNPA, (**B**) Pathifier and (**C**) ssGSEA. ssNPA was used with PD = 10. Clustering for all methods was performed with the first 10 principal components. Molecular subtype was assigned according to estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) status. We define ER-negative, PR-negative and HER2-negative as basal (triple-negative); ER-negative, PR-negative and HER2-positive as HER2+; ER-positive, PR-positive and HER2-negative as luminal A; and ER-positive, PR-positive and HER2-positive as luminal B. Breast cancer subject survival by clusters discovered with (**D**) ssNPA, (**E**) Pathifier and (**F**) ssGSEA. Survival curves are truncated to only display points for which at least 10 subjects survive; however, analysis was performed with the full dataset as shown in Supplementary Figure S8

expression of ITM2A was the highest for the samples in cluster 2 which was also the group that had the best survival outcomes (Supplementary Fig. S9). Clusters 0, 3 and 4 had lowest expression of ITM2A and were grouped together with worst survival outcomes. TNN had the second largest loading for PC 1, and its expression is known to be induced in breast cancer metastasis in the bone via regulation by SMAD4 and TGFβ1-signaling (Chiovaro *et al.*, 2015). As with ITM2A, expression of TNN varied in a cluster-dependent manner in our dataset (Supplementary Fig. S9B). Reduced expression of another top loading gene, SCN4B, has been associated with high-grade primary and metastatic tumors, increased RhoA activity and increased cell migration and invasiveness (Bon *et al.*, 2016). On the other hand, overexpression of SCN4B led to reduced tumor progression. Again, we saw in our data that expression of SCN4B varied according to cluster and tracked with the survival outcomes of these clusters (Supplementary Fig. S9C). As a final example, HOXA7 was another top gene that plays an important role in breast cancer. Knockdown of HOXA7 leads to decreased cell proliferation, ERα expression and PR expression (Zhang *et al.*, 2013). HOXA7 is regulated through the HMGA2/TET1/HOX signaling pathway in breast cancer, which is also predictive of patient survival (Sun *et al.*, 2013). This agrees with our observation that HOXA7 was most highly expressed in cluster 2 which is mostly comprised of luminal A samples that are ER+ and PR+ (Supplementary Fig. S9D). Mean expression of HOXA7 was lowest in cluster 3 which is mostly made up of basal samples that are ER- and PR-. These are just a few such examples from the list of top PC loading genes identified by ssNPA, but they highlight the utility of the approach for identifying the key players in the underlying molecular mechanisms of disease subtypes. Other top genes we identified do not yet have well-documented roles in breast cancer, but our work suggests they would be strong targets for research into the mechanism of the disease. For example, CLDN11 encodes a tight junction-associated protein and is known to play a role in head and neck (Li *et al.*, 2018), gastric (Agarwal *et al.*, 2009) and brain cancers (Katsushima *et al.*, 2012). Computational analyses have identified the gene as a biomarker for breast cancer survival (Meng *et al.*, 2016), but its mechanistic role in the disease has not yet been directly investigated.

3.4 ssNPA identifies patient subclusters with different survival rates in lung adenocarcinoma

We also tested ssNPA in subtyping patients in the context of lung adenocarcinoma, a disease for which there is no subtyping ground truth. The normal samples were used as reference dataset when needed (ssNPA, Pathifier), but they were omitted during clustering in order to better facilitate the discovery of new disease subtypes. ssNPA features resulted in four clusters with significantly different survival rates ($P < 0.0001$, Fig. 4A and Supplementary Fig. 10A). Subjects who maintain the greatest survival probability through the first 1500 days were grouped into clusters 1 and 2 (Fig. 4D and Supplementary Fig. S10A). Similarly, subjects with the worst survival were all clustered together in cluster 3. The final cluster was comprised of subjects with an intermediate survival phenotype. While we observe large differences in survival curve at later time points (after 1000 days), these have very few subjects and thus contribute little to the reported P-value. Survival differences across Pathifier and ssGSEA clusters (Supplementary Fig. S10B and C) were also significant ($P = 0.00059$ and $P < 0.0001$, respectively). This suggested that these pathway-based approaches are better able to reflect variation in this particular lung cancer dataset compared to the other examples we tested; however, it is was not obvious *a priori* that this would be the case, and ssNPA was able to perform equally well without requiring any prior information.

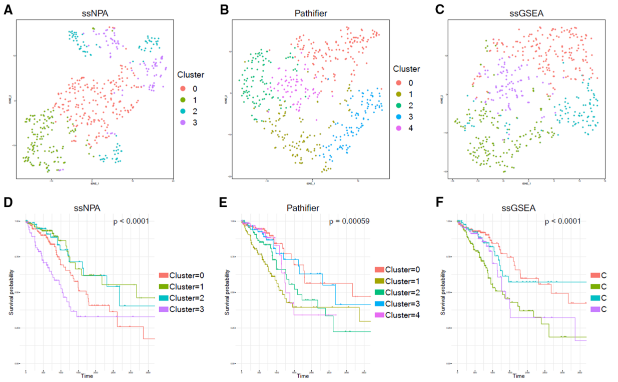

Fig. 4. Lung adenocarcinoma RNA-seq sample clusters (top row) and subject survival by cluster (bottom row) as discovered with (**A**) ssNPA, (**B**) Pathifier, (**C**) ssGSEA. Survival curves are truncated to only display time points for which at least 10 subjects survive; however, analysis was performed with the full dataset (see Supplementary Fig. S10). ssNPA was used with PD = 8. Clustering for all methods was performed with the first 10 principal components

Similar to our analysis of breast cancer, ssNPA features can be used not only to separate patients into groups with coherent clinical phenotypes but also investigate the specific network perturbations that underlying differences among clusters. Supplementary Table S3 lists the top 10 genes based on their factor loadings for the first 10 principal components of the ssNPA features. Notably, many of these genes have well-documented connections to lung physiology and cancer biology. For example, PRR11 was a top gene for PC 1 and its expression is known to play an important role in promoting cell proliferation in non-small-cell lung cancer (Ji *et al.*, 2013). Silencing this gene inactivates the Akt/mTOR signaling pathway, suppresses cell proliferation and triggers autophagy in human lung cancer cells (Zhang *et al.*, 2018). Indeed, we saw that PRR11 expression tracked with survival outcomes, with samples in cluster 3 having both the highest expression of PRR11 and the worst survival outcomes cluster expression decreasing in the order of improving survival (Supplementary Fig. S11A). Another gene we identified, FAM83, is a known oncogene involved in many types of cancer (Snijders *et al.*, 2017). FAM83 plays a role in activating the CRAF/MAPK/mTOR signaling, and its expression is associated with higher tumor grade and worse survival (Cipriano *et al.*, 2012, 2014). Again, in our data FAM83D was most highly expressed in cluster 3 which also had the worst survival outcome (Supplementary Fig. S11B). As a final example, we found the deregulation of AQP1 played an important role in separating the LUAD clusters. This gene has been found to be overexpressed in lung cancer cell lines and linked to increased cell proliferation (Hoque *et al.*, 2006; Xie *et al.*, 2012). Knockdown of the gene leads to decreased migration and invasiveness and decreased expression of MMP-2 and MMP-9 (Wei and Dong, 2015). It is also thought to interact with many other important signaling pathways in the cell including the Wnt/β-catenin/Lin-7 and

FAK/PI3K/AKT pathways (Tomita *et al.*, 2017). Interestingly, its expression varied in a cluster-dependent manner in our samples but was highest in cluster 1 which had relatively good survival outcomes and lowest in cluster 3 which had poor outcomes (Supplementary Fig. S11C). Further study will useful in understanding its exact role and regulation in lung cancer.

These are just a few of the genes we identified with ssNPA, but there were many others that have important connections to lung cancer. Additionally, several of the genes we found do not yet have well-documented links to lung cancer but are highly suggestive. For example, the top gene we identified for PC 1 was CAPN3 whose mutation is associated with dominantly inherited limb girdle muscular dystrophy. The gene encodes an unusual protein that is activated by Na+ and undergoes fast and exhaustive autodegradation which makes it difficult to study, although it can still carry out some of its protease function afterward (Ono *et al.*, 2016). Other members of the calpain family have been implicated in cancer development and progression and have been proposed as possible targets for treatment (Leloup and Wells, 2011), and CAPN3 specifically has been linked melanoma progression through p53 accumulation and regulation of oxidative stress-related pathways (Moretti *et al.*, 2015). We saw that CAPN3 expression varied by cluster with highest expression in cluster 1 and lowest expression in cluster 3 (Supplementary Fig. S11D). Because deregulation of CAPN3 played such a large part in separating the LUAD clusters with ssNPA, further study of its role in lung cancer specifically is merited.

4 Discussion

We presented ssNPA, a new method to assess gene network perturbations in single samples. The method first infers the global network from a set of reference samples using causal graph learning. In the following step, given a new sample, the method calculates its deviation from the reference network at every gene, thus providing information about both the topology and the magnitude of network perturbations. The perturbation feature vector can be used to cluster samples into cell or disease subtypes. We demonstrated the performance of ssNPA by using it to evaluate cluster memberships of datasets with known ground truth; specifically, liver development cells (time course scRNA-seq data) and TCGA breast and lung cancer data. In the first case, we showed that ssNPA performs better than currently used methods and from simple gene-based clustering on finding the true developmental stage and type of the cell. This showed that network perturbation features can recapitulate the time course data. In this dataset, we found that using one of the middle developmental stages (which are equadistant from both progenitor and fully differentiated extremes) as reference point allows for better results. This is likely to be the case in all datasets where changes of the regulatory network depend on time. This is because using a middle point as reference is more likely to be able to capture most regulatory changes in whole time spectrum.

In the cancer data we identified clusters of patients either with good agreement with known histologically determined cancer subtypes (breast cancer) or with significant differences in survival (lung adenocarcinoma). Both these cases demonstrate the ability of ssNPA to identify disease subtypes, which is the most significant problem in developing personalized medicine strategies, especially in complex diseases.

We also compared ssNPA to ssGSEA and Pathfinder, two known methods for single sample analysis. In all cases ssNPA performed better than these methods as evidenced by the greater agreement of the ssNPA-identified clusters to the ground truth and the more significant differences in survival rates in the cancer cases. Network deregulation features also capture differences in the topology of the network of each sample from the reference samples as well differences in the resulting gene expression outputs from those networks. The better performance of ssNPA versus ssGSEA and Pathfinder might reflect the fact that the latter depend on prior knowledge that might not be very accurate or might not reflect the particular conditions in the studied dataset.

In summary, ssNPA is a new method for characterizing single samples of gene expression and offers significant advantages over existing methods. Unlike ssGSEA and Pathifier, it does not require prior pathway knowledge; it is substantially faster than Pathifier; and can be used to produce high-quality sample clusters that reflect the underlying mechanisms of the disease condition or phenotype. Although in this work we used only 3000 genes so we can compare ssNPA to Pathifier, ssNPA can run on any arbitrary number of genes, if it is used with the parallelized FGES version. In the future, ssNPA can be used for analyzing disease data to identify disease subphenotypes and help develop personalized intervention strategies.

Data access

All data used in this project come from published papers and publicly accessible data sources (TCGA) as it is described in Materials and Methods.

Software availability

ssNPA is currently freely available from http://www.benoslab.pitt.edu/Software/ssnpa/. We also include code and all software version information for the analyses for this article in the form of R markdown files, as well as the input dataset files.

Acknowledgements

We would also like to thank the three reviewers, whose constructive comments and criticism helped us improve significantly this manuscript.

Funding

This work was supported by the National Institutes of Health (NIH) [U01HL137159, R01LM012087 to P.V.B.].

Author contributions

P.V.B. conceived the idea. K.L.B. developed the code and executed the experiments under the guidance of M.C. and P.V.B. The article was written by all authors collectively.

Conflict of Interest: none declared.

References

Agarwal,R. *et al.* (2009) Silencing of claudin-11 is associated with increased invasiveness of gastric cancer cells. *PLoS One*, **4**, e8002.

Barbie,D.A. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.

Bon,E. *et al.* (2016) SCN4B acts as a metastasis-suppressor gene preventing hyperactivation of cell migration in breast cancer. *Nat. Commun.*, **7**, 13648.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

Chen,Y. *et al.* (2018) Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res.*, **28**, 879–896.

Chiovaro,F. *et al.* (2015) Transcriptional regulation of tenascin-W by TGF-beta signaling in the bone metastatic niche of breast cancer cells. *Int. J. Cancer*, **137**, 1842–1854.

Cipriano,R. *et al.* (2012) FAM83B mediates EGFR-and RAS-driven oncogenic transformation. *J. Clin. Invest.*, **122**, 3197–3210.

Cipriano,R. *et al.* (2014) Conserved oncogenic behavior of the FAM83 family regulates MAPK signaling in human cancer. *Mol. Cancer Res.*, **12**, 1156–1165.

Cummings,B.B. *et al.* (2017) Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.*, **9**, eaal5209. doi: 10.1126/scitranslmed.aal5209.

Drier,Y. *et al.* (2013) Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. USA*, **110**, 6388–6393.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Gardeux,V. *et al.* (2014) N-of-1-pathways' unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine. *J. Am. Med. Inf. Assoc.*, **21**, 1015–1025.

Hanzelmann,S. *et al.* (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, **14**, 7.

He,L. *et al.* (2000) Loss of FBP function arrests cellular proliferation and extinguishes c-myc expression. *EMBO J.*, **19**, 1034–1044.

Hendrich,B. *et al.* (2001) Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev.*, **15**, 710–723.

Hoque,M.O. *et al.* (2006) Aquaporin 1 is overexpressed in lung cancer and stimulates NIH-3T3 cell proliferation and anchorage-independent growth. *Am. J. Pathol.*, **168**, 1345–1353.

Huang,G.T. *et al.* (2015) T-ReCS: stable selection of dynamically formed groups of features with application to prediction of clinical outcomes. *Pac. Symp. Biocomput.*, 431–442.

Ji,Y. *et al.* (2013) PRR11 is a novel gene implicated in cell cycle progression and lung cancer. *Int. J. Biochem. Cell Biol.*, **45**, 645–656.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Katsushima,K. *et al.* (2012) Contribution of microRNA-1275 to Claudin11 protein suppression via a polycomb-mediated silencing mechanism in human glioma stem-like cells. *J. Biol. Chem.*, **287**, 27396–27406.

Kremer,L.S. *et al.* (2017) Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.*, **8**, 15824.

Leloup,L. and Wells,A. (2011) Calpains as potential anti-cancer targets. *Expert Opin. Therap. Targets*, **15**, 309–323.

Li,H.-P. *et al.* (2018) Inactivation of the tight junction gene CLDN11 by aberrant hypermethylation modulates tubulins polymerization and promotes cell migration in nasopharyngeal carcinoma. *J. Exp. Clin. Cancer Res.*, **37**, 102.

Li,Q. *et al.* (2017a) N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes. *BMC Med. Genomics*, **10**, 27.

Li,R. *et al.* (2017b) MBD3 inhibits formation of liver cancer stem cells. *Oncotarget*, **8**, 6067.

Meng,L. *et al.* (2016) Biomarker discovery to improve prediction of breast cancer survival: using gene expression profiling, meta-analysis, and tissue validation. *Onco Targets Ther.*, **9**, 6177–6185.

Mohammadi,S. *et al.* (2018) A geometric approach to characterize the functional identity of single cells. *Nat. Commun.*, **9**, 1516.

Moretti,D. *et al.* (2015) Calpain-3 impairs cell proliferation and stimulates oxidative stress-mediated cell death in melanoma cells. *PLoS One*, **10**, e0117258.

Ono,Y. *et al.* (2016) An eccentric calpain, CAPN3/p94/calpain-3. *Biochimie*, **122**, 169–187.

Raghu,V.K. *et al.* (2018a) Biomarker identification for statin sensitivity of cancer cell lines. *Biochem. Biophys. Res. Commun.*, **495**, 659–665.

Raghu,V.K. *et al.* (2018b) Evaluation of causal structure learning methods on mixed data types. In: Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery. *Proc. Mach. Learn. Res.*, 48–65.

Raghu,V.K. *et al.* (2018c) Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *Int. J. Data Sci. Anal.*, **6**, 33–45.

Ramsey,J. *et al.* (2017) A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int. J. Data Sci. Anal.*, **3**, 121–129.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sachs,K. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Sedgewick,A.J. *et al.* (2016) Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*, **17**, 175.

Sedgewick,A.J. *et al.* (2019) Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics*, **35**, **1204–1212**.

Snijders,A.M. *et al.* (2017) FAM 83 family oncogenes are broadly involved in human cancers: an integrative multi-omics approach. *Mol. Oncol.*, **11**, 167–179.

Su,X. *et al.* (2017) Single-cell RNA-Seq analysis reveals dynamic trajectories during mouse liver development. *BMC Genomics*, **18**, 946.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Sun,M. *et al.* (2013) HMGA2/TET1/HOXA9 signaling pathway regulates breast cancer growth and metastasis. *Proc. Natl. Acad. Sci. USA*, **110**, 9920–9925.

Tenenbaum,D. (2016) KEGGREST: client-side REST access to KEGG. R Package Version, 1.

Tomfohr,J. *et al.* (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Tomita,Y. *et al.* (2017) Role of aquaporin 1 signalling in cancer development and progression. *Int. J. Mol. Sci.*, **18**, 299.

van der Maaten,L.J.P. and Hinton,G.E. (2008) Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

Villani,A.C. *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.

Waltman,L. and van Eck,N.J. (2013) A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B*, **86**, 471.

Wei,X. and Dong,J. (2015) Aquaporin 1 promotes the proliferation and migration of lung cancer cell in vitro. *Oncol. Rep.*, **34**, 1440–1448.

Xie,Y. *et al.* (2012) Aquaporin 1 and aquaporin 4 are involved in invasion of lung cancer cells. *Clin. Lab.*, **58**, 75–80.

Xu,M. *et al.* (2017) Methyl-CpG-binding domain 3 inhibits epithelial-mesenchymal transition in pancreatic cancer cells via TGF-beta/Smad signalling. *Br. J. Cancer*, **116**, 91–99.

Yang,L. *et al.* (2017) A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology*, **66**, 1387–1401.

Zhang,L. *et al.* (2018) Silencing of PRR11 suppresses cell proliferation and induces autophagy in NSCLC cells. *Genes Dis.*, **5**, 158–166.

Zhang,Y. *et al.* (2013) Homeobox A7 stimulates breast cancer cell proliferation by up-regulating estrogen receptor-alpha. *Biochem. Biophys. Res. Commun.*, **440**, 652–657.

Zhao,T. *et al.* (2018) Single-cell RNA-seq reveals dynamic early embryonic-like programs during chemical reprogramming. *Cell Stem Cell*, **23**, 31–45.e7.

Zhou,C. *et al.* (2019) Integral membrane protein 2A inhibits cell growth in human breast cancer via enhancing autophagy induction. *Cell Commun. Signal.*, **17**, 105.

Zhou,W. *et al.* (2016) Far upstream element binding protein plays a crucial role in embryonic development, hematopoiesis, and stabilizing Myc expression levels. *Am. J. Pathol.*, **186**, 701–715.