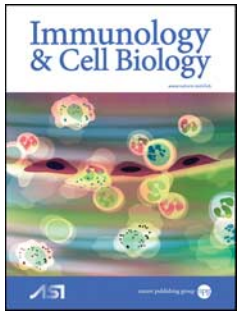
ACCEPTED ARTICLE PREVIEW

Accepted Article Preview: Published ahead of advance online publication



Processing, visualising and reconstructing network models from single cell data

Steven Woodhouse, Victoria Moignard, Berthold Göttgens, Jasmin Fisher

Cite this article as: Steven Woodhouse, Victoria Moignard, Berthold Göttgens, Jasmin Fisher, Processing, visualising and reconstructing network models from single cell data, *Immunology and Cell Biology* accepted article preview 18 November 2015; doi: 10.1038/icb.2015.102.

This is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication. NPG are providing this early version of the manuscript as a service to our customers. The manuscript will undergo copyediting, typesetting and a proof review before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Accepted manuscript

Received 21 October 2015; revised 3 November 2015; accepted 11 November 2015; Accepted article preview online 18 November 2015

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

1. Processing, visualising and reconstructing network models from single cell data

2. Steven Woodhouse[1,2], Victoria Moignard[1,2], Berthold Göttgens[1,2] and Jasmin Fisher[3,4*]

1.

[2.1]*Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK*

3.

[4.2]*Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK*

5.

[6.3]*Microsoft Research, Cambridge, UK*

3.

[4.] [4]*Department of Biochemistry, University of Cambridge, Cambridge, UK*

1.

*Correspondence should be addressed to Jasmin Fisher jf416@cam.ac.uk or jasmin.fisher@microsoft.com

Accepted manuscript

2. Abstract

3. New single-cell technologies readily permit gene expression profiling of thousands of cells at single-cell resolution. In this review we will discuss methods for visualisation and interpretation of single-cell gene expression data, and the computational analysis needed to go from raw data to predictive executable models of gene regulatory network function. We will focus primarily on single-cell real-time quantitative PCR and RNA sequencing data, but much of what we cover will also be relevant to other platforms, such as the mass cytometry technology for high dimensional single-cell proteomics.

4. Introduction

Recent advances in protocols, microfluidics technology, and a reduction in costs have opened up a new field of single-cell genomics. This new field promises to provide insights into cellular identity and decision making over more conventional bulk population data, which averages over the properties thousands of cells and therefore obscures the state of individual cells[1].

5. After experimental measurement, data must firstly be processed and normalised to ensure correct interpretation. We will begin with a discussion of the steps needed to process single-cell qPCR data, which can simultaneously measure the level of expression of tens to hundreds of genes, and the newer technique of single-cell RNA-sequencing, which can sample the whole transcriptome. Once these steps have been carried out subsequent analysis can be applied to answer specific biological questions.

Typically one of the first questions a researcher will want to ask about their single-cell expression data set is whether interesting sub-populations with characteristic gene expression profiles can be identified[2,3,4,5,6,7]. These sub-populations might represent previously unidentified cell types or cells with an abnormal phenotype. For example, in a study of the immune system, two separate populations might correspond to activated and naive cells, or in a patient sample, to cancerous and healthy cells[8,9,10,11,12]. Once identified, the sub-populations can be isolated and investigated further. Population-level gene expression data, on the other hand, would average out the differences

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

between these groups, giving a representative view of neither. We will discuss different algorithms for visualising and identifying structure in single-cell gene expression data sets (Figure 1).

Once structure has been identified, the researcher can investigate potential biological processes that have been captured in the data. Often, the data are representative of a developmental or differentiation time-course, with early cells such as stem cells or early progenitors progressing to more mature cells[13,14,15]. In this case, the single cell profiling data set can be used for gene regulatory network reconstruction. We will discuss several techniques for reconstructing the regulatory networks driving the journey from early to late cell types (Figure 1). Some of these methods have been adapted from analyses of population data, and some have been specifically developed to take advantage of single-cell resolution data.

1. Data Processing

1.1. qPCR on the Fluidigm BioMark

The Fluidigm BioMark platform uses microfluidics devices to scale back reagent and sample requirements, thereby facilitating thousands of parallel qPCR reactions and allowing up to 96 genes to be assayed in a single cell. Initial data processing takes place using the Fluidigm Real-Time PCR Analysis Software. Like conventional qPCR, the BioMark outputs Ct values, and the software allows sample and assay names to be assigned along with the quality thresholds, baseline correction methods and Ct thresholds used to calculate the final Ct values.

Next, expression values that fall outside of the linear range of the BioMark HD or the assays are excluded from further analysis. To do this, a limit of detection (LOD) is calculated from standard curves for each primer set as the last Ct value at which amplification can be reliably and repeatedly detected[13,17]. Ct values higher than the LOD, as well as samples where the amplification has failed entirely or where the amplification curves have failed quality control are usually given the same value as the limit of detection and treated as not detected.

Additional filtering can be used to exclude whole genes or samples. For example, genes may be excluded where there is amplification in typically >10% of no template controls, and where the amplification level in no template controls is too similar to that of single cells to be sure that the expression in the cells is real. In published studies, cells have been excluded from the analysis based on a number of criteria, including lack of expression of key or housekeeping genes, expression of no or low numbers of cells, or where the expression of particular genes differs significantly from the population[3,4,18,19], although these can also occur due to the choice of genes and transcriptional bursting rather than due to a poor quality or missing cell.

Single cell expression data are typically log-normally distributed[16] so it is useful to view data on a Log2 scale. The final step of processing therefore converts the data either to ΔCt values normalized against one or more housekeeping genes which exhibit stable expression across the populations[3,4,15,18,19,20,21], or as the Log2 expression above the LOD (PCR cycles above background; Log2Ex)[22,23]. Log2Ex values can be further normalized to remove variability due to factors such as cell size[17]. Fluidigm have now generated an R package, Singular, for processing and basic analysis of single cell qPCR and RNAseq data.

1.2 Single cell RNAseq

Single cell RNAseq (scRNAseq) has recently come to the fore for transcriptomics due to increases in multiplexing and concurrent decreases in price. Compared with qPCR, it offers the potential to

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

study the entire transcriptome rather than a specific set of pre-selected genes, so has a much wider potential for discovery. However, there are many current challenges both for processing samples and analyzing data[24,25].

There are many different scRNAseq protocols which can capture different aspects of the transcriptome depending on the priming and reverse transcription (RT) methods used. Typically, either the 5' or 3' end of the transcript is captured[26,27], although some methods can capture entire transcripts[28,29]. Samples are multiplexed using indexed primers during library preparation, with 96 to 384 individual cells sequenced per lane of a flow cell. After sequencing, samples are deconvoluted based on index sequences, and normalised read counts are generated for further analysis. Alternatively, short and unique DNA sequences (unique molecular identifiers, UMI) can be incorporated into every transcript during the RT step to act as barcodes to enable molecule counting. Regardless of how many times a transcript-UMI pair is sequenced, it can only have come from a single mRNA within the cell and so is only counted once, with the total number of UMIs per transcript summed to give an absolute expression count for each gene[30]. However, this currently only allows for the sequencing of the 3' end of the transcript, providing information about expression levels but not splicing.

Quality check of samples is an important step before downstream analysis. An important quality control method for scRNAseq is the inclusion of extrinsic standards to facilitate normalization and comparison between single cells. Typically, RNA standards of known concentration and sequence, such as the External RNA Control Consortium (ERCC) set of 92 artificial RNA molecules[31], are spiked into the reverse transcription step. These molecules should be amplified uniformly across samples, so can be used to estimate RT efficiency, technical variation in library preparation and to indicate which genes show real biological variation as well as technical noise. Spikes can additionally be used to identify cells with degraded RNA, for example where the percentage of mapped reads is particularly low compared with reads mapped to spike molecules. Other important metrics which are used for quality control and to discard poor-quality cells include the fraction of reads mapped to mitochondrial genes (a large fraction is believed to be indicative of the cell undergoing apoptosis)[25,32]. Principal component analysis (discussed later), can also be used to identify outlier cells, based upon the assumption that good-quality cells will cluster together while poor-quality cells will be isolated[25].

Samples undergo initial quality control prior to alignment, with tools originally developed for bulk RNA-seq such as fastqc which monitor sequencing quality, GC nucleotide content, sequence length and so on. Reads are assigned to individual cells based on their indexes, the sequencing adapters are trimmed off and the resultant sequences are mapped to a reference transcriptome using existing alignment tools such as TopHat[33], Star[34] or GSNAP[35]. Tools such as HTseq[36] are then used to generate read counts per gene. Further quality control, as discussed above, can then be carried out. Normalisation is required to account for differences in sequencing depth between samples, which is calculated from the total mappable reads and the ratio of mapped reads to those coming from spike molecules. However, adequate normalization of scRNAseq data is an ongoing challenge[25] as much is still unknown about technical variation in library preparation and sequencing bias towards particular transcripts.

Once we have processed and normalised our data set we can begin to ask interesting biological questions about the cells that we have measured. Usually the first thing we would like to do is to try to identify and visualise structure in the data and establish which biological processes have been captured. We will discuss methods for doing this in the next section.

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

6. 2. Visualisation

High-dimensional data sets can be hard to visualise. A two or three dimensional data set can be directly plotted to try to reveal structure in the data (Figure 2a). This is not possible with high dimensional data such as a single-cell gene expression data set, which has a dimension corresponding to each measured gene. In the field of machine learning, a number of clustering and dimensionality reduction techniques have been developed to help aid visualisation of high-dimensional data[37,38]. Clustering algorithms attempt to group data points into subsets called clusters, where data points within a cluster are more similar to each other than to points from different clusters. Dimensionality reduction algorithms attempt to transform the high-dimensional data set into a lower-dimensional (2 or 3) representation that can then be directly plotted and visualised.

1. 2. 1. Hierarchical clustering

Agglomerative hierarchical clustering has been used to identify sub-populations in single-cell data[4,22]. Rather than seeking to identify a predetermined number of clusters, the algorithm recursively builds a hierarchical representation of the data where each level organises the data into a different number of clusters. This makes the algorithm useful for exploratory analysis.

At the beginning of the algorithm each data point is placed into its own cluster. Then, at each subsequent step the two most similar clusters from the previous iteration are merged into one. The algorithm terminates when all of the data lies in a single cluster[37].

The results of hierarchical clustering can be plotted as a heat map (a coloured representation of the data matrix, reorganised according to the clustering) with a *dendrogram*, which is a binary tree showing the hierarchical neighbour relationships between clusters. As we go to higher levels in the dendrogram, the dissimilarity between merged clusters increases. By examining the reorganised expression matrix, and the cell types and gene expression patterns of closely placed points, natural clusters can often be discerned by eye (Figure 2b).

Before hierarchical clustering can be performed, two measures of similarity need to be specified: a notion of distance between pairs of data points, and a notion of distance between clusters (the linkage criterion), defined in terms of the distance between data points. For the distance between data points, the Euclidean, Manhattan or Spearman correlation distance can be used. For the linkage criterion between clusters A and B, one distance is the nearest neighbour distance (known as single linkage), which is the distance between the point in A and the point in B which are most similar. A second distance is the farthest neighbour (known as complete linkage), which is the distance between the point in A and the point in B which are least similar.

Care must be taken when interpreting the results of hierarchical clustering, keeping in mind that different choices of dissimilarity measure and linkage criterion will result in different hierarchies, and that the algorithm will always impose a hierarchy on the data whether or not one truly exists.

Many other clusterings algorithms exist, but hierarchical clustering and related methods stand out in their utility for exploratory visualisation. Spectral clustering is closely related to diffusion maps[39] (discussed later). DBSCAN is a very commonly used algorithm which groups together points with many nearby neighbours[40]. K-means clustering clustering places each point into the cluster with the closest mean, but requires the desired number of clusters to be specified *a-priori* (and is therefore best used for classification after using another method for exploratory visualisation)[41].

The SPADE algorithm was introduced specifically for the analysis of single-cell data, and is based upon firstly applying hierarchical clustering, and then linking clusters together using a minimum spanning tree to infer developmental progression, while taking into account the existence of rare cell populations via density-dependent downsampling[42]. The BackSPIN algorithm is conceptually

© 2015 Macmillan Publishers Limited. All rights reserved.

similar to hierarchical clustering, but seeks to avoid noise from cell dissimilarity caused by uninformative genes. It works by sorting the gene expression matrix through cell-cell and gene-gene similarity[43]. Grün et al. recently introduced an algorithm, RaceID, designed specifically for identification of rare cell types in single cell data[44].

2. 2. 2. Principal component analysis

The most ubiquitous tool used for dimensionality reduction is principal component analysis (PCA). PCA is used to find a projection of the data onto a smaller linear subspace, such that the variance of the projected data is maximised (the data points are spread out as much as possible)[37,38].

PCA finds a sequence of uncorrelated best linear approximations of the data, which are ordered in decreasing order of variance and are known as *principal components*. The first two or three of these components can be retained and plotted as a scatter plot to perform dimensionality reduction[4,22,45]. Equivalently, PCA can viewed as an instance of the Multidimensional Scaling (MDS) algorithm with Euclidean distances. MDS attempts to preserve all pairwise distances between data points in the high-dimensional space, as best as possible[46]. In general, the method will fail to preserve all pairwise distances perfectly. For example, in a ten dimensional data set, up to 11 data points may be mutually equidistant, while there is no way to accurately represent this in a three dimensional plot.

The advantages of principal component analysis are its simplicity, its computational efficiency and its direct interpretation in terms of linear combinations of genes. A disadvantage is that it fails to capture non-linear structure in the data. Single-cell gene expression data in particular can be expected to be highly non-linear (Figure 3d). Manifold learning and graph-based visualisations, discussed next, attempt to address this weakness. Non-linear generalisations of PCA also exist, most notably kernel PCA, which also falls into the class of manifold learning algorithms[47].

3. 2.3. Non-linear manifold learning

In general, there is no way to represent a high-dimensional data set in a lower dimensional space without discarding information. Different dimensionality reduction tools therefore aim to embed the data in a way that preserves some particular property of interest. We will focus the remainder of our discussion on two non-linear dimensionality reduction methods that have recently been used to visualise single cell gene expression data: t-Distributed Stochastic Neighbor Embedding (t-SNE) and diffusion maps.

t-SNE aims to preserve the pairwise distance between points, but (unlike MDS/PCA) only between those points which are very close neighbours in the high dimensional space, focusing only on preserving local structure rather than attempting to preserve pairwise distances between all points[48]. This allows the global structure of the embedding to become non-linear, as distances at different regions of the embedding are allowed to correspond differently to distances in the high dimensional space. Diffusion maps attempt to reconstruct the global non-linear connectivity of the data from a local random walk on the data, and place points close together in the low-dimensional map if they are connected by many short paths in the high-dimensional space[49,50].

Diffusion maps and t-SNE belong to a class of techniques known as manifold learning algorithms. Manifold learning is based on the hypothesis that the dimensionality of the data under consideration is only artificially high, and that rather than being uniformly distributed throughout the high dimensional space it actually lies on a lower dimensional non-linear manifold that curves through the high dimensional space (Figure 3). This manifold hypothesis seems particularly appropriate for single-cell gene expression data as the expression states that a cell can take are highly constrained

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

by an underlying gene regulatory network. A cellular state therefore has relatively few degrees of freedom in terms of the states it can immediately progress to, an idea that was formalised in Waddington's epigenetic landscape[51]. This landscape can also be expected to be non-linear because of complex gene interactions, waves of gene expression and positive and negative feedback loops in the gene regulatory network. PCA can be considered as a linear manifold learning algorithm, that assumes data lies on a linear hyperplane.

The aim of a non-linear manifold learning algorithm is to reconstruct the geometry of the low-dimensional manifold the data lies on from the only information we have: the similarities between data points. Key to these algorithms is the idea that it is local distances, similarities between nearby points that are important for reconstructing this geometry.

1. ***2.3.1. t-SNE***

t-SNE defines a Gaussian probability distribution over pairs of data points in the high-dimensional space, that captures the pairwise similarity of points. The probability of a pair being chosen is high if the points are very similar in terms of their high-dimensional gene expression profiles, and very close to zero if they are dissimilar. A second distribution over pairs of points in the low dimensional embedding is then defined, this time as a Student's t-distribution. Points are placed on the two- or three- dimensional plot, and the discrepancy between these two probability distributions (the Kullback–Leibler divergence) is iteratively minimised via a gradient descent optimisation method, shifting points around until this discrepancy reaches a minimum[48].

A disadvantage of t-SNE is that it can be slow to compute. For this reason, a Barnes–Hut approximation algorithm has been developed which can scale better to larger data sets[51]. t-SNE has been used very successfully to dissect heterogeneity in leukemia samples using single-cell mass cytometry data[5], and to identify an improved cell-sorting strategy for hematopoietic stem cells by separating true stem cells from non-stem cells in combined single-cell qPCR and single-cell indexed flow cytometry data[6].

2. ***2.3.2. Diffusion maps***

Unlike t-SNE, which tends to pull data apart into separate clusters, diffusion maps tend to organise the data into a single continuous manifold and are therefore particularly appropriate when the data is sampled from a developmental or differentiation process that we wish to reconstruct (Figure 3). The algorithm was first introduced in the context of biology by Haghverdi, Buettner, and Theis, adapting it to deal with uncertainties or missing measurement values in qPCR data, and adding density normalisation to cope with heterogeneities in data sampling[15,53].

Diffusion maps are based upon the idea of reconstructing the global geometry of the data set by constructing and iterating a random walk on the data points, and attempt to accurately approximate the so-called "diffusion distance" between data points when mapping to a lower-dimensional space[49,50]. This diffusion distance is small if there are many high-probability short paths connecting the two points, and large if the points are connected only by long paths or low-probability transitions. When reducing to a lower-dimensional space, the diffusion algorithm attempts to place points with a low diffusion distance nearby in the map.

The diffusion map algorithm works by constructing a transition matrix on the data, where the probability of jumping from one data point to another in one step is high if the two data points are similar in the high-dimensional space. If the points are dissimilar, this probability is very close to zero. It then employs results from a branch of mathematics known as spectral theory to approximate the diffusion distance in lower dimensional space without explicitly iterating the random walk, which would be computationally expensive.

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

The diffusion map algorithm is computationally efficient, and, because it integrates over all paths, robust to noise, unlike some manifold learning algorithms. For a review of other manifold learning approaches, see Lin et al.[54]

3. 2.3.3. Graph-based representations

Another way to represent the relationships between single-cell gene expression profiles is as an undirected graph, where an edge connecting a pair of profiles indicates that they are similar in the high-dimensional space. There are essentially two ways to do this.

Firstly, a unit distance graph can be constructed, where cells are connected by an edge if they are a distance of exactly 1 away from each other[15]. This requires a metric that specifies when a cell should be considered a neighbour of another. A useful notion of neighbour is that the two cells are different in the binary expression of exactly one gene. That is, one of the cells expresses the gene and the other does not. This representation is particularly useful for gene regulatory network reconstruction, which we will discuss in the next section. A disadvantage of this approach is the large number of cells which need to be measured in order to construct a connected graph. In order for two cells to be connected, we need to have measured any intermediate cellular states.

A second approach is to construct a k-nearest neighbours graph. Here, each cell is connected to the k cells which are most similar to it. This representation requires fewer cells than the state-transition graph and retains continuous gene expression levels. However, it fixes all cells to have the same number of neighbours. This representation has recently been use to discover subpopulations in leukemia samples measured by mass cytometry[55].

Usually it makes sense to apply a range of different visualisation approaches to a single-cell data set and to compare and contrast the results. Reassuringly, there will often be good high-level agreement between different methods, but depending on the data, the specific biological processes under study and the specific questions the researcher is interested in, one representation may be more appropriate. If we are interested in a developmental or differentiation process, a method which attempts to reconstruct a developmental time-course, such as a diffusion map or graph representation may be best suited. In this case, the next biological question we will usually ask is whether we can learn the gene regulatory networks underlying development. We will cover this in the next section.

7. 3. Network reconstruction

1. 3.1. Statistical relationships between genes

2. When trying to infer regulatory interactions between genes one of the most obvious things to look for is correlation in gene expression levels. If there is a strong correlation between two genes, this may indicate that one directly regulates the other. Performing this analysis on all possible pairs and selecting strong and statistically significant relationships results in a relevance network, which is an undirected graph where edges between genes indicate a potential interaction (Figure 4a). There are two types of edge: positive edges where strong positive correlation indicates a potential activation and negative edges where strong negative correlation indicates a potential repression[56].
3.
4. The standard Pearson correlation coefficient is a measure of the linear dependence between two variables. As genes may not exhibit a linear relationship the Spearman rank correlation is generally preferred. Spearman correlation measures how well the relationship between the two variables can be fit by a monotonic function. A measure from information theory called mutual information is more general still and can capture more complex relationships.
5.
6. These statistical relationships are very simple to compute, can scale to huge data sets, and have

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

been successfully applied to find previously unknown regulatory links in single cell data[4]. However, relevance networks are undirected and can be very dense, with almost all gene pairs showing significant correlation. Partial correlation attempts to address this second issue, by calculating correlation after first controlling for the effect of all other genes, and therefore retaining the links that are most likely to be direct interactions (Figure 4a).

7.

8. To understand another problem, consider two subpopulations, one of which expresses gene A but not gene B, the other expresses gene B but not A. Correlation would suggest a very strong negative link between the two genes, although there is no strong reason to believe they directly regulate each other. One way to address this is to compute the correlation only on cells which coexpress the genes of interest[57].

9.

10. Other methods for detecting statistical signals in gene expression data exist. One notable method is GENIE3, which constructs random forests of decision trees[58]. GENIE3 was best performer in the DREAM5 Network Inference challenge for population data[59], and has been applied to single-cell data[60]. A reweighted mutual information measure known as DREMI, specifically designed for single-cell data has recently been introduced[61].

11.

8. 3.2. Learning Bayesian networks

9. A Bayesian network is a probabilistic model defined by a directed acyclic graph coupled with a joint probability distribution[62]. Nodes in the graph correspond to variables in the model (which in our case represent genes) and edges indicate direct influence between variables (Figure 5). A variable is conditionally independent from variables it is not directly connected to, given the value of its parents. The global joint probability distribution for the model can therefore be given by a set of much smaller conditional probability tables that give the probability of a variable given the value of its parents. Given a Bayesian network, inference can be performed in the model to predict the effect of perturbations on the probability distributions of downstream genes. It is the directed acyclic graph and conditional independence structure of Bayesian networks which allows efficient inference to be performed, and permits efficient learning of models from data.

10. Bayesian networks were first applied in the context of genomics by Friedman et al.[63] to infer networks from population microarray data. They have since been applied by Sachs et al.[64] to reconstruct signalling networks from single cell flow cytometry data taken from primary human T cells. Sachs et al. measured 11 phosphorylated proteins and phospholipids in 5400 individual cells spanning nine different conditions. Seven of these conditions directly perturbed variables of the network by activating or inhibiting phosphorylation. The differences between these perturbed populations were then used to infer causality. Data were first discretised to 3 levels (low, medium and high expression), and then learning algorithms were applied to construct a Bayesian network which was subsequently successfully validated against existing literature.

11. One of the key insights of this paper is the need for perturbation data to reconstruct an accurate Bayesian network from single cell data. If we have two correlated variables, X and Y, and we find that direct inhibition of X affects the value of Y and that direct inhibition of Y does not affect X, we can conclude that Y is downstream of X. A learning algorithm can then often determine the direction of additional edges downstream of the perturbed variables, even when these edges were not directly perturbed.

12. Bayesian networks are an attractive model class for encoding directed, causal relationships between genes, which support efficient structure and parameter learning from data and can naturally cope with noise due to their probabilistic interpretation. Bayesian networks have been successfully applied to dissect connections between components of signalling pathways. However, they do suffer from two drawbacks that limit their application to reconstruction of wider gene regulatory networks.

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

13. Firstly, as concluded by the Sachs study, to infer accurate networks the single cell data needs to be coupled with intervention data. Generating such intervention data is very time consuming and often impractical, and cannot be done without disturbing the wild-type system that we are supposed to be studying. Secondly, Bayesian networks are acyclic, and have no feedback. Feedback is a crucial component of gene regulatory networks.

14. 3.3. Synthesising executable gene regulatory networks

15. We recently introduced a method for reconstructing mechanistic models of gene regulatory networks directly from single cell gene expression data[15,65]. This method, the Single Cell Network Synthesis toolkit, is based on constructing a state transition graph from the data, and then finding a model which matches this graph.
16. A state transition graph is a unit distance graph, where nodes correspond to measured cell states, and edges between states are changes in single genes. We view this graph as a branching time course reconstructed from the single-cell snapshot measurements, where each edge is a potential transition between individual cellular states. We then ask for a model with a set of rules that allow us to walk via a series of single-gene changes from the earliest cell states in the graph (e.g. stem cells or cells measured on day 1 of a differentiation time course) to the latest cell states (differentiated cells). The resulting network models this differentiation journey.
17. The synthesis method results in an asynchronous Boolean network, which is a mechanistic model that can be directly executed on a computer and used to make predictions. Each gene is given a Boolean rule that specifies how its on/off expression value changes over time due to regulation by other genes in the network. The reconstructed networks are directed and can have cycles with feedback and auto-regulation, and mechanistic logic (Figure 4b).
18. Once a matching model has been found, analysis can be performed to find the stable states of the model, which may correspond to mature, differentiated cell types. In-silico over-expression or knock outs can then be introduced to assess their effect on the model's behaviour.
    We applied the SCNS toolkit to study early blood development in the mouse embryo, using single-cell qRT-PCR analysis of 33 transcription factors and additional marker genes in 3934 cells with blood-forming potential captured at four sequential time points between embryonic day 7.0 and day 8.5. Several novel predictions from the model about the role of Hox and Sox factors in blood development were validated experimentally (data shown in Figure 2b, resulting network in Figure 4b).
19. Synthesis approaches generally lead to combinatorial rather than statistical problems, which are then exactly solved using algorithms which leverage highly optimised specialist solvers[66]. Synthesis yields a globally-optimal model which satisfies the specification given by the data completely, or otherwise informs the user that no such model exists. We can also use synthesis to find all models that satisfy a given set of model specifications. Experiments can then be designed which distinguish between these different possibilities. The major disadvantages of the approach are the need for a very large number of cells (thousands rather than hundreds) in order to construct a connected state transition graph, and the restriction to only binary, rather than continuous, gene expression levels.

20. Concluding remarks

21. The field of single-cell genomics is still relatively young, and there is a sparsity of high-quality data sets with a large number of cells. This is set to change as adoption of these protocols becomes more widespread. New single-cell studies will give us insights into organ development and human disease. A larger number of data sets will allow comprehensive comparison of the techniques covered in this review and allow assessment of the advantages and disadvantages of different approaches and improvement of the methods.

    The move towards whole-transcriptome RNA-sequencing data removes the selection bias of qPCR

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

data and allows analysis of the full genetic program of the cell but presents its own challenges. RNA-sequencing will also allow the impact of genetic variation in cancer upon gene expression to be assessed. Finally, new, emerging methods may allow spatial information to be incorporated into whole-transcriptome studies of gene expression[67,68].

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

22. Conflict of interest

The authors declare no competing financial interests.

23. Acknowledgements

S.W is supported by a Microsoft Research PhD Scholarship.

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

1. Moignard V, Göttgens B. Transcriptional mechanisms of cell fate decisions revealed by single cell expression profiling. *Bioessays* 2014; **36**:419-26.
2. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, Zabala M, Bueno J, Neff NF, Wang J, Shelton AA, Visser B, Hisamori S, Shimono Y, van de Wetering M, Clevers H, Clarke MF, Quake SR. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011; **29**:1120-1127.
3. Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, Jaenisch R. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* 2012; **150**:1209–1222.
4. Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol* 2013; **15**:363–372.
5. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 2013; **31**:545-552.
6. Wilson NK, Kent DG, Buettner F, Shehata M, Macaulay IC, Calero-Nieto F, Sánchez Castillo M, Oedekoven CA, Diamanti E, Schulte R, Ponting CP, Voet T, Caldas C, Stingl J, Green AR, Theis FJ, Göttgens B. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* 2015; **16**:712-724.
7. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014; **343**:776-779.
8. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 2015; **525**:251-255.
9. Mahata B, Zhang X, Kolodziejczyk AA, Proserpio V, Haim-Vilmovsky L, Taylor AE, Hebenstreit D, Dingler FA, Moignard V, Göttgens B, Arlt W, McKenzie AN, Teichmann SA. Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids De Novo to Contribute to Immune Homeostasis. *Cell Rep* 2014; **7**:1130-1142 .
10. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A, Bernstein BE. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 2014; **344**:1396-1401.
11. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, Schwartz S, Fowler B, Weaver S, Wang J, Wang X, Ding R, Raychowdhury R, Friedman N, Hacohen N, Park H, May AP, Regev A. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 2014; **510**:363-369.
12. Spitzer MH, Gherardini PF, Fragiadakis GK, Bhattacharya N, Yuan RT, Hotson AN, Finck R, Carmi Y, Zunder ER, Fantl WJ, Bendall SC, Engleman EG, Nolan GP. An interactive reference framework for modeling a dynamic immune system. *Science* 2015; **349:**1259425.
13. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014; **32**:381-386 .
14. Bendall SC, Davis KL, Amir ED, Tadmor MD, Simonds EF, Tiffany JC, Shenfeld DK,

© 2015 Macmillan Publishers Limited. All rights reserved.

Accepted manuscript

ACCEPTED ARTICLE PREVIEW

Nolan GP, Pe'er D. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B cell Development. *Cell* 2014; **157**:714-725 .
15. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol* 2015; **33**:269-276.
16. Fluidigm, SINGuLAR Analysis Toolset User Guide. 2014.
17. Livak KJ, Wills QF, Tipping AJ, Datta K, Mittal R, Goldson AJ, Sexton DW and Holmes CC. Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* 2013; **59**:71–79.
18. MacArthur BD, Sevilla A, Lenz M, Müller FJ, Schuldt BM., Schuppert AA et al. Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. *Nat. Cell Biol* 2012; **14**:1139–1147.
19. Pina C, Fugazza C, Tipping AJ, Brown J, Soneji S, Teles J, Peterson C and Enver T. Inferring rules of lineage commitment in haematopoiesis. *Nat. Cell Biol* 2012; **14**:287–294 2012.
20. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* 2010; **18**:675–685.
21. Swiers G, Baumann C, O'Rourke J, Giannoulatou E, Taylor S, Joshi A *et al*. Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nat. Commun* 2013; **4**:2924.
22. Guo G, Luc S, Marco E, Lin TW, Peng C, Kerenyi MA et al. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell Stem Cell* 2013;**13**:492–505.
23. Ståhlberg A, Andersson D, Aurelius J, Faiz M, Pekna M, Kubista M, Pekny M. Defining cell populations with single-cell gene expression profiling: correlations and identification of astrocyte subpopulations. *Nucleic Acids Res* 2011; **39**:e24.
24. Macaulay IC and Voet T. Single cell genomics: advances and future perspectives. *PLoS Genet* 2014; **10**:e1004126.
25. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet* 2015; **16**:133-145.
26. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2012; **2**:666–673.
27. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, Linnarsson S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 2011; **21**:1160–1167.
28. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 2013; **10**:1096–1098.
29. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 2009; **6**:377–382.
30. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 2012; **9**:72–74.
31. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 2011; **21**:1543–1551.
32. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014; 11:163-166.
33. Trapnell C, Pachter L, Salzberg SL. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**:1105–1111.

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29:**15–21.
35. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010 **26**:873–881.
36. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015; **31**:166-169.
37. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer: Berlin, Germany, 2011.
38. Bishop C. *Pattern Recognition and Machine Learning*. Springer: Berlin, Germany, 2007.
39. Nadler B, Galun M. Fundamental limitations of spectral clustering. In: *Advances in Neural Information Processing Systems*, 2006, pp. 1017-1102.
40. Ester, M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*, 1996, pp. 226-231.
41. MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1996, pp. 281–297.
42. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 2011; **29**:886-891.
43. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A1, *et al*. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015; **347**:1138-1142
44. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N1, *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 2015; **525**:251-255.
45. Kumar RM, Cahan P, Shalek AK, Satija R, DaleyKeyser AJ, Li H, Zhang J, Pardee K, Gennert D, Trombetta JJ, Ferrante TC, Regev A, Daley GQ, Collins JJ. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 2014; **516**:56-61.
46. Borg I, Groenen P. *Modern Multidimensional Scaling: theory and applications*, 2nd edn. Springer: Berlin, Germany, 2005.
47. Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 1998, **10**:1299-1319.
48. van der Maaten LPJ, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *JMLR* 2008; **9**:2579-2605.
49. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS* 2005 **102**: 7426–7431.
50. Nadler B, Lafon S, Coifman R, Kevrekidis IG. Diffusion Maps - a Probabilistic Interpretation for Spectral Embedding and Clustering Algorithms. In: *Gorban AN, Kégl B, Wunsch DC, Zinovyev A (eds). Principal Manifolds for Data Visualization and Dimension Reduction*, Springer: Berlin, Germany, 2008, pp 238-260.
51. Goldberg AD, Allis CD, Bernstein E. Epigenetics: A landscape takes shape. *Cell* 2007; **128**: 635-638.
52. van der Maaten LPJ. Accelerating t-SNE using Tree-Based Algorithms. *JMLR* 2014; **15**:3221−3245.
53. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 2015; **31**:2989-2998.
54. Lin B, He X, Ye J. A geometric viewpoint of manifold learning. *Applied Informatics* 2015; **2**:3.
55. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, Litvin O, Fienberg

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 2015; **162**:184-197.

56. Butte AJ, Kohane IS. Relevance Networks: A First Step Toward Finding Genetic Regulatory Networks Within Microarray Data. In: Parmigiani G, Garett ES, Irizarry RA, Zeger SL (eds). *The Analysis of Gene Expression Data,* Springer:Berlin, Germany, 2003, pp 1-45.
57. Pina C, Teles J, Fugazza C, May G, Wang D, Guo Y, Soneji S, Brown J, Edén P, Ohlsson M, Peterson C, Enver T. Single-Cell Network Analysis Identifies DDIT3 as a Nodal Lineage Regulator in Hematopoiesis. *Cell Rep* 2015;**11**:1503-1510 .
58. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* 2010, **5**:e12776.
59. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, The DREAM5 Consortium, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. *Nature Methods* 2012; **9**:796-804.
60. Ocone A, Haghverdi L, Mueller NS, Theis FJ. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics* 2015; **31**:i89-96.
61. Krishnaswamy S, Spitzer M, Mingueneau M, Bendall SC, Stone EL, Litvin O, Pe'er D, Nolan GP. Conditional Density-based Analysis of T cell Signaling in Single Cell Data. *Science* 2014; **346**:1250689.
62. Heckerman D. A Tutorial on Learning With Bayesian Networks. *Technical Report MSR-TR-95-06* 1996.
63. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol* 2000; **7**:601-620.
64. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 2005; **308**:523-529.
65. Fisher J, Koksal AS, Piterman N, Woodhouse S. Synthesising Executable Gene Regulatory Networks from Single-Cell Gene Expression Data, *27th International Conference on Computer Aided Verification, Springer* 2015.
66. Fisher J, Piterman N, Bodik R. Towards Synthesizing Executable Models in Biology. *Front Bioeng Biotechnol* 2014; **19**:2-75
67. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science* 2014; **343**:1360-1363.
68. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015; **33**:495-502.

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

FIGURE LEGENDS

Figure 1: Overview of the different analyses covered in this review.

Figure 2: High-dimensional data can be hard to visualise. **A.** The classic Iris flower dataset has four dimensions, measured in three different species. As this is a low-dimensional dataset, we can directly plot it to try to undercover structure. Here we are plotting sepal length against petal width and petal length. **B.** Hierarchical clustering of a high-dimensional single-cell qPCR dataset with 40 genes and 3934 cells[15]. Rows represent genes and columns represent cells. Left-hand side colour bar shows measured ΔCt level of expression of genes. Top colour bar shows cell types - blood cell progenitors (red) fall into one large cluster while other cell types separate into two more large clusters and do not separate by cell type.

Figure 3: Manifold learning. **A.** A two-dimensional curving manifold embedded in three dimensions. **B.** Diffusion map applied to "unfold" the manifold to a rectangle, giving one possible way of representing the three-dimensional data in two dimensions. **C.** t-SNE separates bone marrow cells measured by cytometry into different immune cell types[5]. Points are coloured by CD20 expression, a B-cell cell-surface lineage marker. **D.** PCA, a linear projection method, fails to separate between the different immune subtypes on the first two principal components.

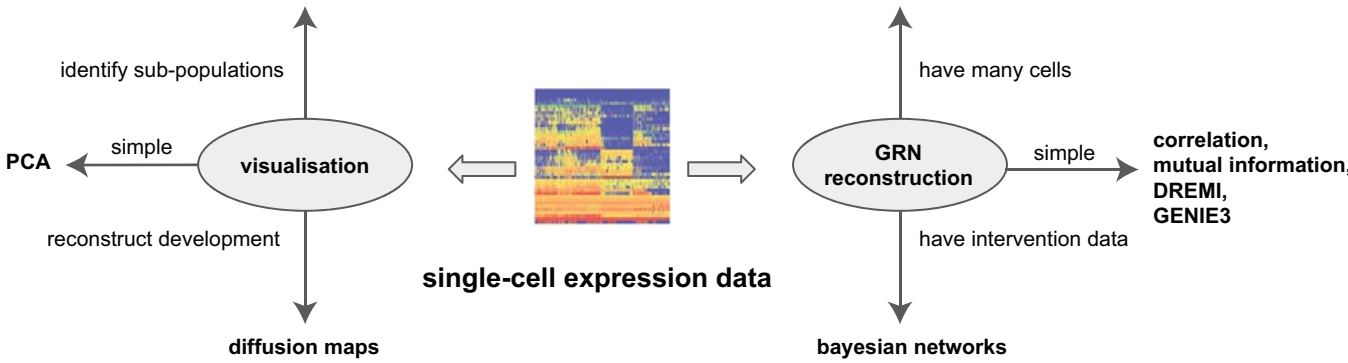
Figure 4: A. Relevance network obtained for early blood development from partial correlation analysis[15]. Green: activation; red: repression. Data used to reconstruct network shown in Figure 1a. **B.** Asynchronous Boolean network obtained from same data set using the SCNS toolkit.

Figure 5: Bayesian network for T cell signalling[64].

Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

ACCEPTED ARTICLE PREVIEW

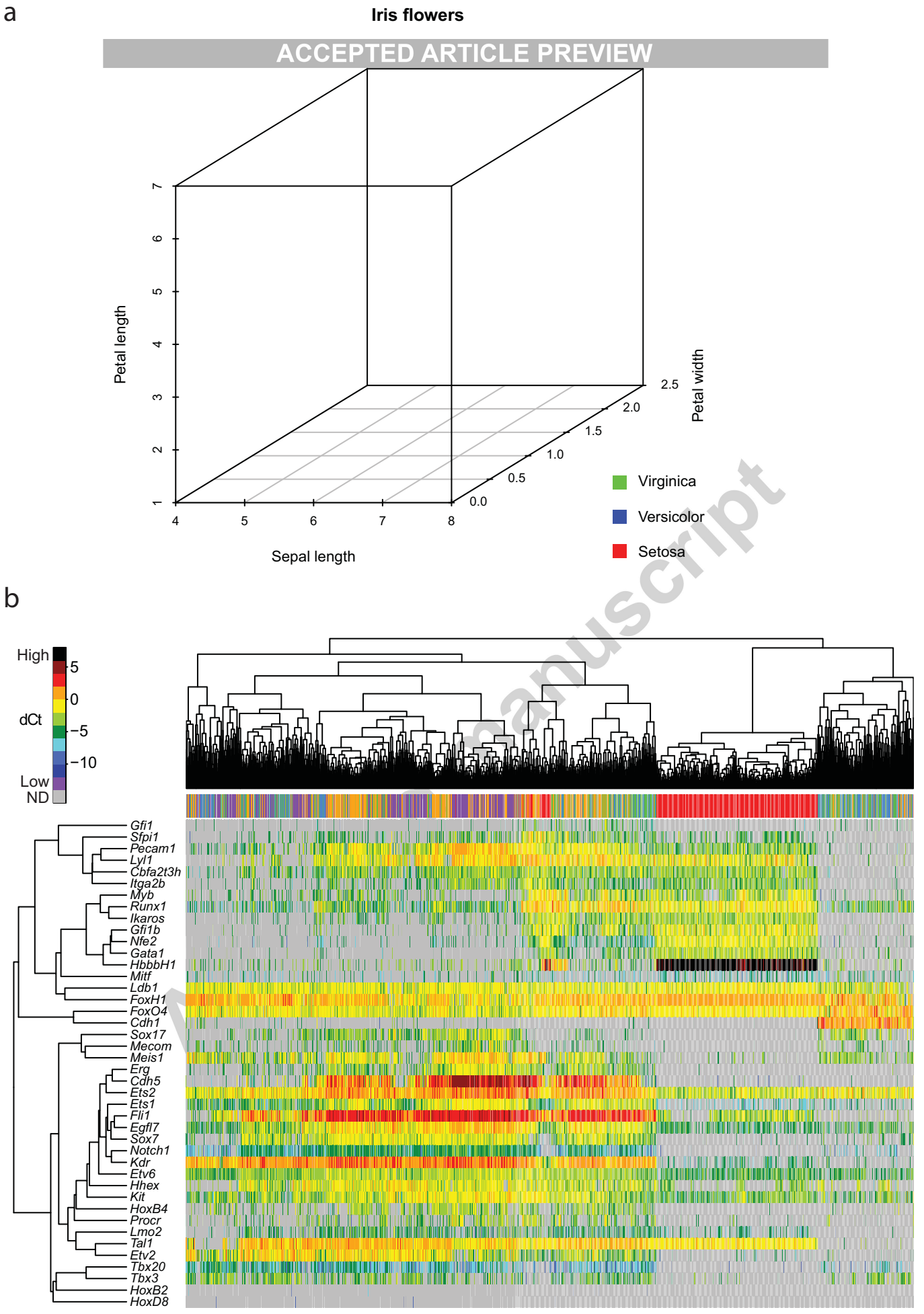


Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

FIGURE 1

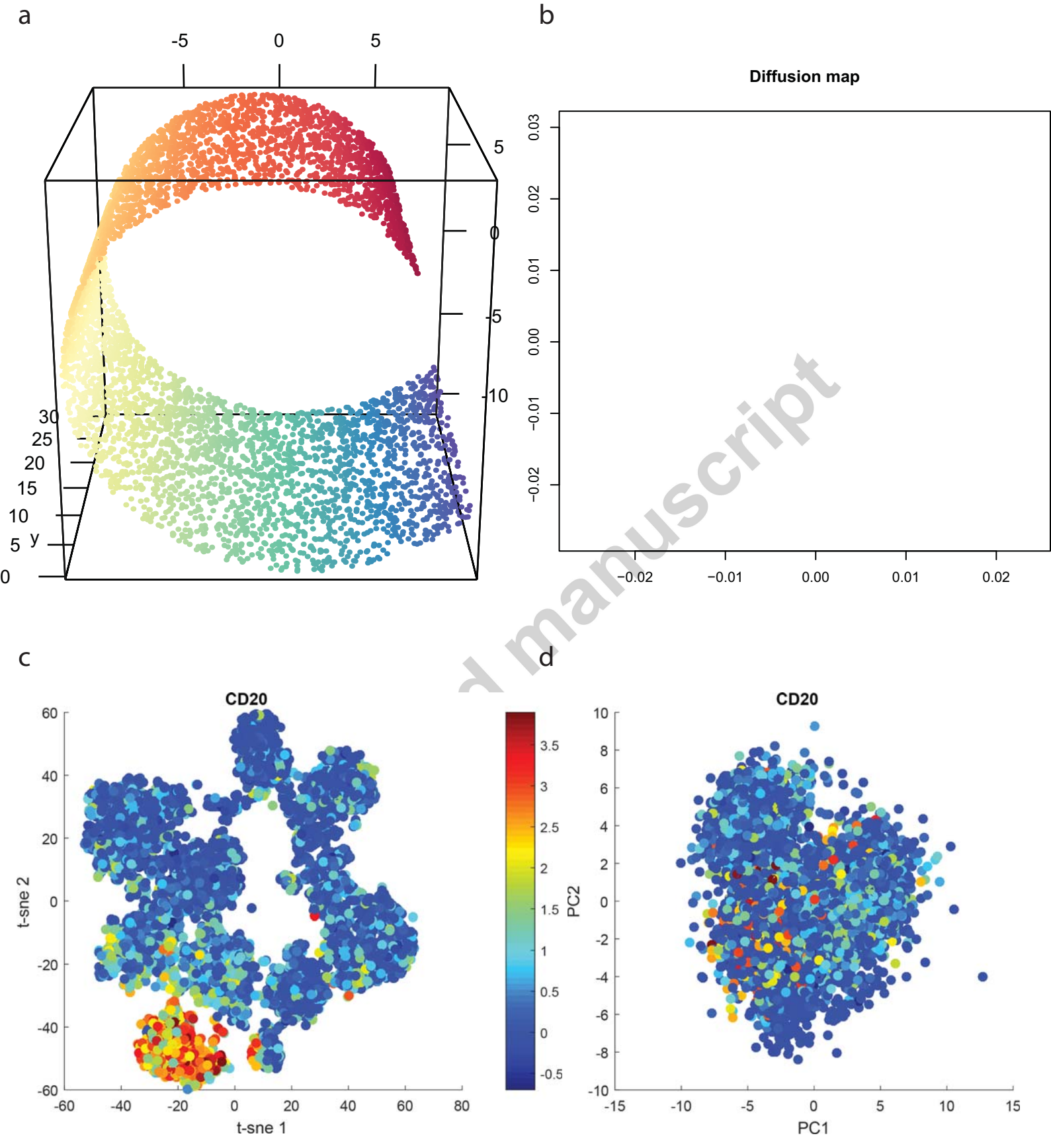
ACCEPTED ARTICLE PREVIEW



Author manuscript

FIGURE 2

© 2015 Macmillan Publishers Limited. All rights reserved.



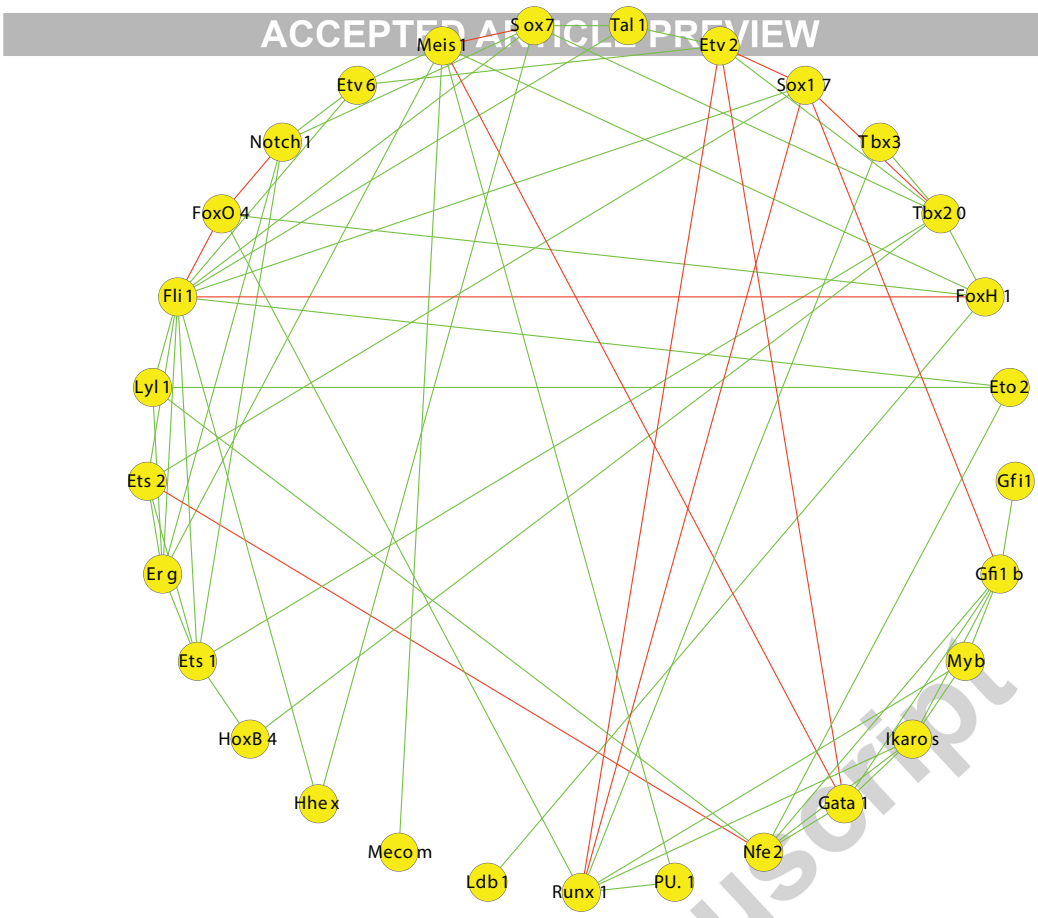
manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

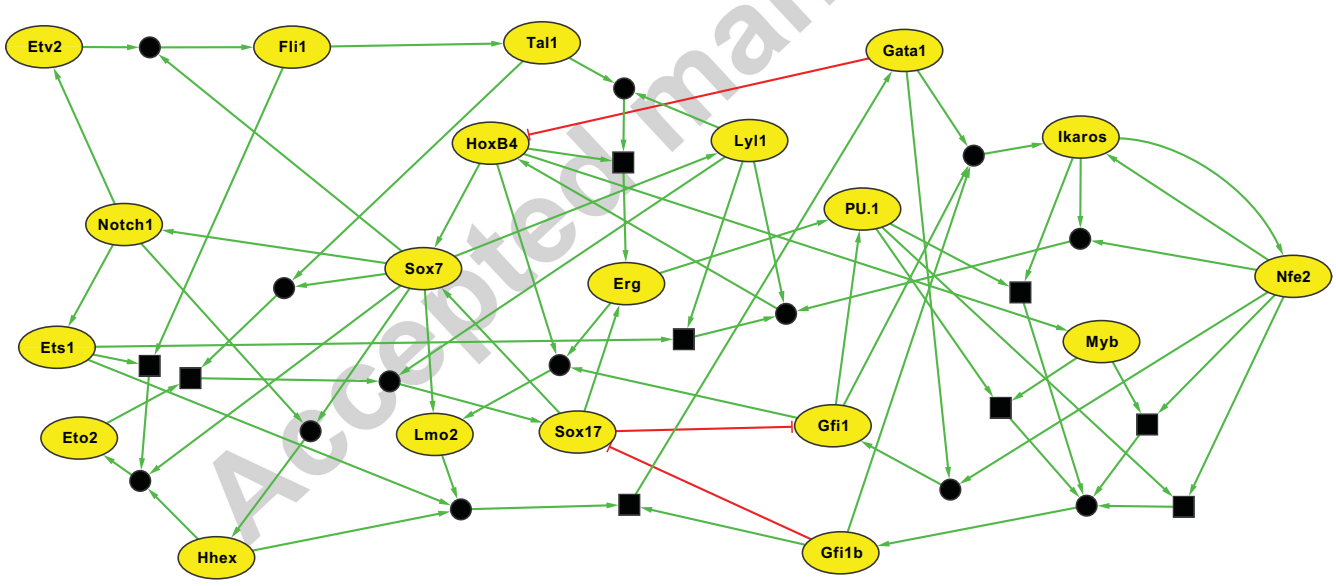
FIGURE 3

ACCEPTED ARTICLE PREVIEW

a



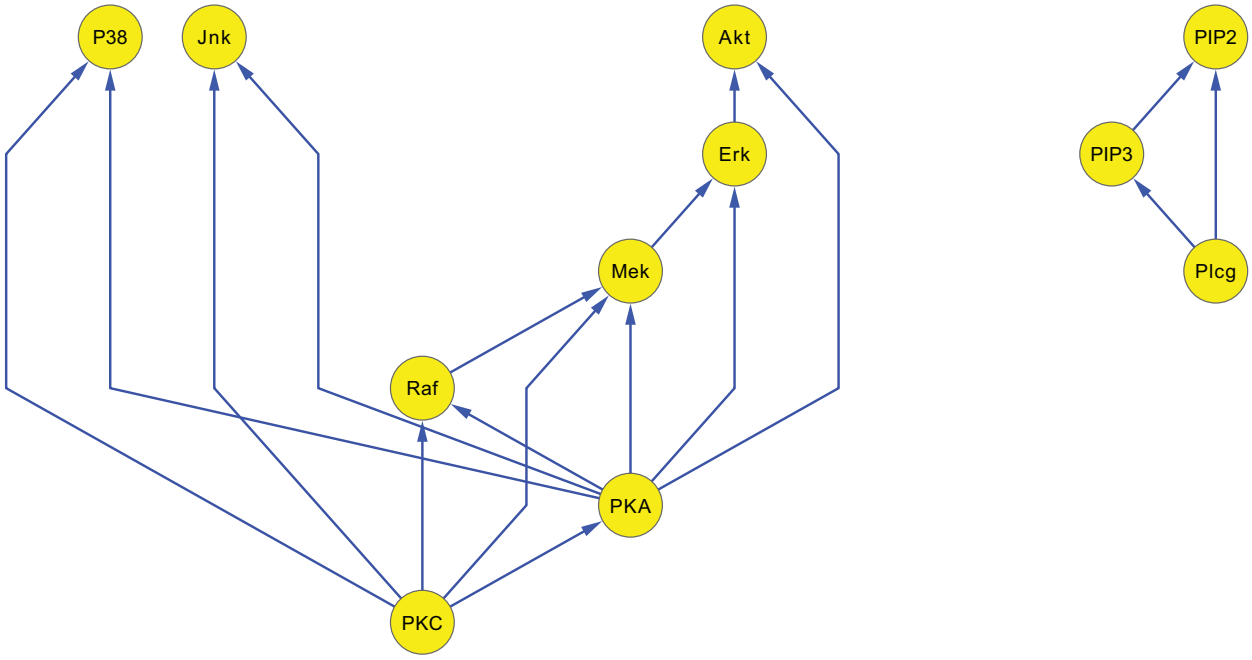
b



Accepted manuscript

© 2015 Macmillan Publishers Limited. All rights reserved.

FIGURE 4



Accepted manuscr

© 2015 Macmillan Publishers Limited. All rights reserved.

FIGURE 5