Molecular Human Reproduction, Vol.21, No.2 pp. 115–125, 2015
Advanced Access publication on September 19, 2014 doi:10.1093/molehr/gau083

molecular human reproduction

NEW RESEARCH HORIZON *Review*

Bioinformatics approaches to single-blastomere transcriptomics

Leila Taher[1], Martin J. Pfeiffer[1,2], and Georg Fuellen[1,*]

[1]Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany
[2]Max Planck Institute for Molecular Biomedicine, Münster, Germany

*Correspondence address. Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany. E-mail: fuellen@uni-rostock.de

Submitted on March 18, 2014; resubmitted on August 27, 2014; accepted on September 8, 2014

ABSTRACT: The totipotent zygote gives rise to cells with differing identities during mouse preimplantation development. Many studies have focused on analyzing the spatio-temporal dependencies during these lineage decision processes and much has been learnt by tracing transgenic marker gene expression up to the blastocyst stage and by analyzing the effects of genetic manipulations (knockout/ overexpression) on embryo development. However, until recently, it has not been possible to get broader overviews on the gene expression networks that distinguish one cell from the other within the same embryo. With the advent of whole genome amplification methodology and microfluidics-based quantitative RT–PCR it became possible to generate transcriptomes of single cells. Here we review the current state of the art of single-cell transcriptomics applied to mouse preimplantation embryo blastomeres and summarize findings made by pioneering studies in recent years. Furthermore we use the PluriNetWork and ExprEssence to investigate cell transitions based on published data.

Key words: single-cell transcriptomics / embryo development / gene regulation / lineage decisions / network biology

Introduction

Importance of single-cell transcriptomics for embryology

Embryo development starts out with a single fertilized oocyte that has to give rise to all cell types and tissues needed for the establishment of a whole organism. After fertilization, the embryonic gene expression program is activated as the mouse zygote starts dividing and progresses through the preimplantation stages of development until, at Day 3.5, a blastocyst consisting of three distinct lineages, namely trophectoderm (TE), primitive endoderm (pEnd) and primitive ectoderm (pEct, or epiblast), is formed. While the TE and pEnd will give rise to extraembryonic tissues, the cells of the primitive ectoderm are considered pluripotent, and are the precursors of the actual fetus. How cellular heterogeneity arises from a single cell in a concerted manner is not fully understood. However, many hypotheses about the underlying regulatory mechanisms do exist, and most of them are not mutually exclusive. A common assumption is that lineage-determining molecules are distributed asymmetrically within a cell, so that after cell division the two daughter cells would not inherit the same set of instructions, leading to different cell types. Also much evidence hints at the importance of cell location within the developing embryo (Rossant and Tam, 2009). Once the embryo reaches the 16-cell stage, some cells will be located in the inside of the embryo while others will still be in contact with the surroundings. Cell location is assumed to impact fate decisions of these cells through cell–cell contact-dependent signaling pathways (Nishioka *et al.*, 2009). Further, it has been proposed that lineage decisions are also governed by stochastic processes at the level of gene regulation (Wennekamp and Hiiragi, 2012).

The above studies rely on labor intensive and time-consuming approaches, such as the use of transgenic mouse models (reporter genes), immunofluorescent analyses (labeling of marker genes) and extensive embryo manipulation (dissociation and aggregation). These approaches have substantially advanced our knowledge of how differences arise during preimplantation development. However, they are limited in that they are only able to assay relatively few genes at a time. Therefore, they cannot aspire to providing a comprehensive understanding of the transcriptional changes that ultimately determine cell fate. However, the recent availability of high-throughput single-cell gene expression data within the developing embryo has the potential to significantly advance the field of developmental biology.

Single-cell and blastomere transcriptomics: state of the art

Recent advances in single-cell transcriptomics

It is not until the last few years that methods have been developed that allow for the analyses of minute specimens such as single, isolated cells. The early embryo contains rare cell types, which exist only transiently. While microarray technology has become a standard procedure in

biology, microarrays can only detect sequences homologous to the probes on the array, making the technology prone to false negatives and largely blind to alternatively spliced transcripts. More recently, next generation sequencing (NGS) has become a driving force in molecular biology. In contrast to microarrays, almost all nucleic acids present in a cell can be sequenced and quantified by this method. Because it would enable the reconstruction and analysis of entire organismal cell lineage trees, single-cell NGS has the potential to significantly advance our understanding of the mechanisms that lie at the basis of developmental biology. However, despite advances in RNA sequencing, it is not currently practicable to sequence RNA directly from single cells. Instead, RNA first needs to be converted to cDNA and amplified. The losses and noise introduced by this procedure lead to appreciable errors. Although not routinely available yet, nanopore-based devices are now opening new horizons, and should eventually permit sequencing of RNA without amplification (Branton et al., 2008; Ayub et al., 2013). Nevertheless, microarrays remain popular. Microarrays are relatively inexpensive and widely available, and produce easy-to-analyze data. NGS data remain costly and decidedly difficult to process computationally, but provide substantially more detail. Finally, another highly quantitative and sensitive technique that is commonly used to analyze gene expression at the transcriptional level is real-time quantitative PCR (qPCR). However, its application is limited to relatively small numbers of different transcripts.

Recent advances in blastomere transcriptomics

Pioneering work has been done by Tang et al. (2009), who improved the cDNA amplification protocol to successfully quantify the transcriptome of single mouse oocytes and blastomeres from a 4-cell stage embryo using RNA sequencing (RNA-Seq). The same study showed for the first time that multiple transcript isoforms are simultaneously expressed in the same cell for hundreds of genes. Notably, most transcript isoforms remain undistinguished when using microarrays. A year later, Guo et al. (2010) employed high-throughput single-cell qPCR to describe the expression of 48 pre-selected genes in ~500 individual cells corresponding to mouse preimplantation embryos, from the zygote up to the blastocyst stage. We inspect their data in more detail below. Microarray technology was subsequently applied to single blastomeres of 2 and 3 cell embryos and to subcellular structures of oocytes and zygotes by VerMilyea et al. (2011), who described asymmetries in the transcriptomes within substructures of mouse oocytes, but were not able to detect differences between individual blastomeres after cleavage. More recently, Tan et al. (2013) followed up on the fate of mouse blastomeres using a combination of microarrays and qRT-PCR. They profiled pooled embryos, single embryos, and individual blastomeres, and conjectured that *Oct4*, *Sall4* and *Nanog* act as the main regulators of preimplantation development. Focusing on human, Yan et al. (2013) applied single-cell RNA-Seq to 124 individual cells from oocytes and preimplantation embryos, from zygote to blastocyst, providing the most comprehensive description of the transcriptome landscape of human early embryos to date. Finally, Ohnishi et al. (2014) also used microarrays and qPCR on mouse single inner cell mass (ICM) cells. They found the cells of the ICM to be essentially indistinguishable at the 32-cell stage (embryonic day 3.25), identified novel pEnd and pEct markers, and proposed a model in which pEnd and pEct lineages segregate within a population of initially seemingly equivalent ICM cells.

Given a suitable genetic background, RNA-seq enables the simultaneous quantification of allele-specific differences in gene expression of thousands of genes. Particular genes, such as *Nanog* or *Pou5f1*, have been shown to be expressed monoallelically during preimplantation development using PCR-based methods (Miyanari and Torres-Padilla, 2012; Pfeiffer et al., 2013). However, the extent of this phenomenon has only become apparent with the advent of single-cell RNA-Seq. Thus, Tang and colleagues concluded in 2011 that a rather large number of genes are only monoallelically expressed in cleavage stage embryos (Tang et al., 2011). Two years later, Xue and colleagues generated single-cell RNA-Seq data for a more comprehensive set of preimplantation stages, ranging from oocytes to morula, in both mouse and human (Xue et al., 2013). Based on these data, they were able to identify and follow paternally and maternally expressed genes in individual mouse cells, finding that 53% of the 8-cell embryo transcripts and 23% of morula transcripts exhibit monoallelic maternal expression patterns. Finally, widespread allele-specific gene expression has been confirmed in a recent in-depth study based on single-cell RNA-Seq data of mouse blastomeres up to the late blastocyst stage (Deng et al., 2014). This study also suggested that monoallelic gene expression is random, and, possibly, a dynamic feature of mammalian cells. Random changes in gene expression are believed to contribute to cell plasticity.

Single-cell transcriptomics enables observations with an unprecedented high-resolution. For example, in mouse, based on their selection of 48 genes, Guo et al. (2010) showed that expression patterns specific to the TE, pEnd and pEct can be clearly distinguished by the 64-cell stage. In previous stages, and in particular in the morula, cells appear to co-express the transcription factors that have been associated with each of the three lineages. These results are consistent with the observations of Xue et al. (2013), who noted that individual cells at the same stage exhibit very similar global expression profiles from oocyte through morula. Additionally, splicing patterns can be measured for each individual cell. Thus, Yan et al. (2013) collected evidence indicating that the level of alternative splicing within individual human cells depends on the developmental stage, with ~1000 genes expressing multiple transcript isoforms within the same blastomere.

In summary, these results illustrate the power of single-cell gene expression profiling at revealing novel insights into developmental biology that could not have been gained through traditional bulk profiling approaches.

Analytical approaches and challenges in the bioinformatics of single-cell and blastomere transcriptomics

Common bioinformatics approaches to transcriptome analysis

As with other high-throughput datasets, careful interpretation of single-cell transcriptomic data is essential for generating hypotheses that guide further research. An individual cell contains a limited total number of mRNA transcripts. Hence, in contrast to bulk cell sequencing, single-cell sequencing requires an amplification step, which may introduce artificial biases. First, the amplification results in a relatively low coverage. Second,

some transcripts are preferentially amplified. Depending on their magnitude, these biases make the bioinformatics analysis challenging, and can significantly affect the interpretability of the data. Moreover, the total number of mRNA transcripts in a single cell varies with factors such as cell size and cell cycle phase (Marinov *et al.*, 2014). This has two consequences. First, the magnitude of the aforementioned biases is likely to depend on the individual cell. Second, the underlying assumptions of standard normalization procedures such as RPKM (reads per kilobase of sequence range per million mapped reads, (Mortazavi *et al.*, 2008)) are likely to be violated. Additional variability between cells arises from the fact, as indicated by recent studies, that some genes are expressed in only a fraction of cells, in a rather stochastic manner (Raj *et al.*, 2006). These are, presumably, genes expressed at low levels in bulk samples. Marinov *et al.* (2014) examined these issues in detail and concluded that, at present, it is not possible to confidently distinguish between biological and technical variability, and proposed the introduction of spike-in mRNA sequences of known abundance to estimate the number of RNA transcripts in each cell. Wu *et al.* (2014) also advocate for the use of spike-ins to estimate technical and biological variability between cells. With the same aim, they contrasted the number of genes detected in combinations of replicate pairs of samples to the mean total number of genes detected. Their results confirm that variability is substantially higher for single-cell than for bulk samples. Finally, as an alternative methodology that permits a ready comparison of results, they proposed to compute expression values relative to the median expression across all transcripts in the cell.

Because single-cell datasets became available only recently, the number of bioinformatics methods especially designed for the analysis of these data remains limited (Robert, 2010; Ning *et al.*, 2014). As a consequence, many bioinformatics tools developed for bulk cell sequencing are routinely applied to single-cell transcriptomic data despite their known limitations. Therefore, bioinformatics tools are in great demand to catch up with the increase in single-cell expression data. Analytical methods applied to the analysis of large-scale transcriptomic datasets from late-stage embryos produced artificially by *in vivo* insemination, IVF and somatic cell nuclear transfer have been reviewed in Rodriguez-Zas *et al.* (2008). Such methods comprise differential gene expression analysis and network inference (e.g. Martens and Apweiler, 2009). Functional analysis based on gene ontology, module, pathway and network knowledge often furthers our understanding of the underlying regulatory mechanisms. For example, a recent analysis compared gene expression data from morula to blastocyst in rat and mouse, looking for species-specific differences between several pathways (Casanova *et al.*, 2012). Among others, they found differences in the regulation of the Notch pathway, a highly conserved signaling pathway that influences differentiation, proliferation and apoptosis during development and whose manipulation could improve the efficiency of rat embryonic stem cell (ESC) derivation. All these analytical methods are, in principle, transferable to embryos at any stage.

One of the main challenges in the analysis of high-throughput data is the recognition of biologically meaningful patterns. One particular widely-used technique is clustering. Clustering algorithms detect patterns within datasets, grouping together the data to highlight those patterns. One of the simplest clustering approaches applied to the analysis of gene expression profiles is hierarchical clustering (Eisen *et al.*, 1998). In hierarchical clustering, relationships among genes and/or samples are represented by a dendrogram where branch lengths reflect distances between objects. The definition of the distance measure is crucial to identify meaningful relationships between those objects. Commonly used distances include Euclidean distance as well as Pearson and Spearman correlation. For instance, in order to identify relationships between gene expression profiles, each gene is initially assigned to a cluster containing only one element. Next, a pairwise distance matrix is calculated for all clusters. In this case, the Euclidean distance quantifies the absolute difference in expression level between two genes, across all samples; the Pearson (or Spearman) correlation compares the shape of the curves representing the expression patterns of two genes. Subsequently, the closest pair of clusters is merged together into a single cluster, reducing the number of total clusters by one, and the pairwise distances between clusters are re-computed to account for this change. Distances between pairs of clusters can be defined in several ways. For complete linkage, the distance between two clusters is the maximum distance from any object in one cluster to any object in the other cluster. This procedure is repeated until all objects in the dataset are clustered into a single cluster. Hierarchically clustered genes and samples can be analyzed and visualized using heatmaps. For example, Guo *et al.* (2010) applied hierarchical clustering to group the expression profiles of 48 genes across 159 cells obtained from ~64-cell blastocysts into TE, pEnd and pEct. The major shortcoming of this method is that gene expression patterns are not necessarily expected to be related in a hierarchical form. Other widely used clustering approaches are the k-means and k-medoids algorithms (Tavazoie *et al.*, 1999; Gasch and Eisen, 2002; Huang and Pan, 2006). K-means partitions genes into a set of *k* clusters, with the aim of minimizing the sum of squared distances between the mean of each cluster, or *centroid*, and its members. Several strategies have been proposed to initialize this algorithm. Most commonly initial centroids are simply chosen randomly from the input data. The initial centroids are then refined iteratively, re-calculating gene memberships and updating the centroids until convergence. K-medoids is a robust version of k-means (Van der Laan *et al.*, 2003). The main difference with k-means is that k-medoids uses an actual member of each cluster as its centroid (medoid). VerMilyea *et al.* (2011) compared transcriptome profiles between blastomeres of the same 2-cell embryo. For this purpose, they clustered the expression profiles into two groups using the k-medoids algorithm. These clusters were then compared with random clusters. No difference was detected between the transcriptomes of sister cells in 2-cell embryos. The unsupervised, unstructured nature of the aforementioned algorithms might pose some difficulties in the interpretation of the results. Self-organizing maps (SOMs) (Tamayo *et al.*, 1999) constitute an alternative to classical clustering algorithms, providing similar, and often superior performance on data with high levels of noise. In contrast to the strict hierarchical clusters and the completely unstructured clusters produced by k-means, SOMs allow the user to impose partial structure on the clusters. A SOM consists of components called 'nodes', which represent subsets of the original data. The nodes are arranged onto a low dimensional lattice. The size (number of nodes per dimension) and type (rectangular or hexagonal) of the lattice is chosen by the user. In addition, each node is associated with a vector of weights, or 'codebook vector', of the same dimension as the input data. The initial codebook vectors of the nodes are defined by randomly assigning input data to the nodes. Assuming that we are interested in classifying samples based on gene expression, the codebook vectors represent gene expression profiles. After initializing the lattice, an instance from the training data is selected randomly,

compared with the lattice, and assigned to the most similar node. Similarity is computed based on the codebook vectors, using Euclidean distance. The codebook vector of the node is then updated to reflect the new data assignment. The updated codebook vector is a weighted average not only of the data assigned to that particular node, but also of the data assigned to its immediate neighbors. As a consequence, nodes that are relatively close in the lattice will tend to represent similar original data. The procedure described above is repeated in a fixed number of iterations. SOMs have, for example, been successfully applied to identify and characterize subpopulations of glial cells throughout development based on gene expression changes in single cells (Rusnakova *et al.*, 2013). Many variations of these basic clustering approaches have been designed for the analysis of gene expression data. For example, a novel clustering approach (Buettner and Theis, 2012) that considers the temporal structure of expression datasets claims to be able to identify different subpopulations of blastomeres in the 16-cell stage in the dataset generated by Guo *et al.* (2010). Finally, dimensionality reduction techniques, such as principal component analysis (PCA), have also been successfully applied to the analysis of gene expression data (e.g. Yeung and Ruzzo, 2001). PCA applies an orthogonal transformation to the original data to obtain a set of uncorrelated variables, the principal components (PCs). The PCs are ordered such that the *k*th PC has the *k*th largest variance among all PCs. Thus, the first PC accounts for most of the variability in the data. The *k*th PC can be interpreted as the direction that maximizes the variation of the projections of the original data, such that it is orthogonal to the first $k-1$ PCs. Traditionally only the first two or three PCs are used for data analysis, since they are supposed to capture most of the variation in the original dataset. For instance, both Xue *et al.* (2013) and Yan *et al.* (2013) used the two and three first PCs, respectively, to show that cells from human oocytes and preimplantation embryos can be easily distinguished from each other according to their developmental stage.

Sample bioinformatics analysis of single-blastomere transcriptomic data

In the analysis of expression data generated from early-stage embryos the challenge is to identify cell subpopulations as early as possible, as well as the genes determining their differentiation. This is exemplified in Fig. 1, where we have reanalyzed Guo *et al.*'s single-cell gene expression data from 64-cell stage embryos (Guo *et al.*, 2010) to identify the most prominent markers of early embryo differentiation. Independent of the method chosen, we observed three main clusters of cells. The smallest cluster, marked in blue, expresses high levels of the classical pluripotency markers, including *Nanog*, *Klf2*, *Esrrb*, *Fgf4*, *Sox2*, *Sall4*, *Klf4*, *Pou5f1* and *Utf1*, and is representative of the pEct within the ICM. The largest cluster, marked in red, expresses high amounts of *Cdx2* and *Tctap2c*, and represents the TE. Finally, cells marked in green express high levels of *Gata6*, *Sall4*, *Klf4*, *Pou5f1* and *Utf1*, and represent the pEnd. While all these genes are considered markers for the aforementioned cell subpopulations, most of them are dispensable for blastocyst formation itself. Indeed, mutant embryos with knocked-out copies of these genes often form an apparently morphological normal blastocyst. Yet, these genes seem to become crucial later on. For example, *Cdx2* negative embryos form morphologically normal blastocysts even if the maternal *Cdx2* has been depleted, but fail to hatch and implant (Wu *et al.*, 2010). Also, total ablation (maternal and zygotic) of *Pou5f1* (more precisely, of its isoform *Oct4A*) leads to grossly normal blastocysts (Wu *et al.*, 2013). However, the ICM of such embryos is not pluripotent, and, therefore, neither ESCs can be derived from these embryos nor is the embryo able to develop into an embryo proper (Nichols *et al.*, 1998). Similar results are observed upon ablation of *Klf5*, a transcription factor that positively regulates *Pou5f1* (Parisi and Russo, 2011). *Esrrb* knockout mouse embryos fail to properly develop extraembryonic ectoderm. Therefore, although *Esrrb* is regarded as a marker for pluripotency, in our context it is actually crucial for the establishment of an extraembryonic structure that originates from the pEct (Luo *et al.*, 1997). Similarly, newborn *Utf1* double mutant mice are significantly smaller than their wildtype counterparts, a phenotype that has been ascribed to the loss of *Utf1* expression in the extraembryonic ectoderm (Nishimoto *et al.*, 2013). A deficiency in the expression of *Fgf4* has more severe consequences. Although *Fgf4* knockout embryos undergo the segregation of the TE and ICM lineages, their ICMs consist only of pEct. Furthermore, the level of *Fgf4* present in the embryo controls the relative proportion of pEct and pEnd in the resulting ICM (Kang *et al.*, 2013; Krawchuk *et al.*, 2013). Finally, although not completely understood, the ablation of *Sox2* has similarly drastic effects, causing developmental arrest at the morula stage (Keramari *et al.*, 2010). In summary, although all the aforementioned genes are recognized as markers for different lineages, among the genes analyzed only a few appear to be crucial to the lineage decisions preceding the formation of the blastocyst.

Interestingly, in the hierarchical clustering, cells of the pEnd appear closer to the TE than to the pEct. This observation may seem counterintuitive as the pEnd and pEct both segregate from the ICM. However, only the pEct gives rise to the embryo proper, while the pEnd and the TE are functionally linked in that they form extraembryonic lineages. The k-medoids approach results in similar clusters (Fig. 1B). Indeed, the smallest cluster, corresponding to the pEct, comprises exactly the same cells. The largest cluster, representing the TE, contains 5 cells more than its counterpart in the hierarchical clustering analysis, while the opposite is true for the cluster corresponding to the pEnd. Compared with the average pEnd cells, these cells express particularly low levels of *Esrrb*, *Pou5f1* and *Sox2*, and are highly variable for other markers, making their classification uncertain. Nevertheless, these discrepancies highlight that different clustering methods may lead to different outcomes. The clusters identified by the SOM (Fig. 1C) are in agreement in terms of their number, size and general composition with both hierarchical clustering and k-means. The number of SOM nodes in each cluster roughly reflects the number of cells segregating into each of the lineages. Finally, the results of the PCA were visualized using a biplot (Fig. 1D). The axes in the biplot are the two main PCs, namely PC1 and PC2, which explain, in total, ~60% of the variance in the data. Each point within the plot is representative of a single cell. Their relative location relates to the similarity of their expression patterns. Vectors represent the observed variables, i.e. genes, projected into the 2D plane of the biplot. The angle between two vectors corresponding to two genes shows their correlation (an angle $<90^{\circ}$ implies a positive correlation), while their length depicts the strength of the association between the genes and the PC in the axes. Thus, *Cdx2* and *Tcfap2c* are both highly expressed in a particular group of cells (in red), indicating TE. Also, the expression profiles of the two remaining groups are concordant with the other methods presented. In all cases, from the single-cell analysis it is evident that the cells do not express the marker genes in a mutually exclusive manner, suggesting a

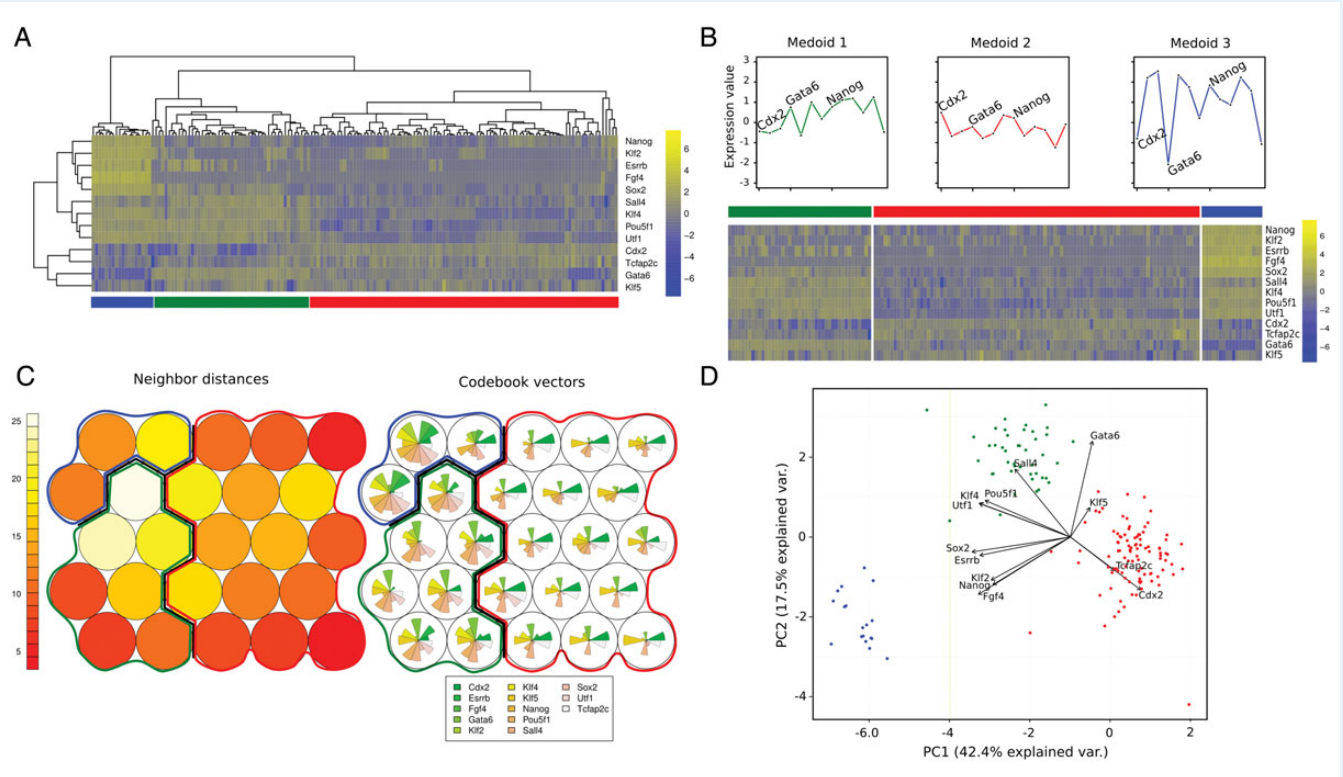

Figure 1 Clustering and visualization approaches applied to the expression data obtained for the mouse 64-cell stage cells by Guo *et al.* (2010). Only the 13 genes that are part of the PluriNetWork (Som *et al.*, 2010) were included in the analysis. Gene expression values were background-corrected, quantile-normalized and scaled across cells, on a per-gene basis (*z*-score) for clustering and visualization. (Groups of) Cells differentiating into trophectoderm (TE), primitive ectoderm (pEct) and primitive endoderm (pEnd) are indicated in red, blue and green, respectively. (**A**) Hierarchical clustering. Genes and cells were clustered with a Pearson correlation distance metric and a Spearman rank correlation distance metric, respectively, using complete linkage. (**B**) K-medoids for $k = 3$. The expression values for all PluriNetWork genes of the three cells that are the medoids of the corresponding clusters are shown on top. (**C**) Self-Organizing Map (SOM) with 5×5 hexagonal topology. The SOM nodes (shown as circles) represent groups of cells. Average Euclidean distances in gene expression between SOM nodes are indicated with different colors in the panel 'neighbor distances'. Average gene expression values for each SOM unit are shown in the panel 'codebook vectors'. SOM nodes with similar codebook vectors lie closer to each other than those with disparate codebook vectors. Hierarchical clustering of SOM codebook vectors clearly identifies the three clusters corresponding to TE, pEct and PrEnd. The boundaries between the three clusters are indicated on the SOM with a black line. In addition, the SOM nodes in each cluster are surrounded with a red, blue or green line, depending on the lineage. (**D**) Principal components analysis (PCA) of the expression data. Biplot of cells and genes according to the first two components of the PCA. Points represent cells. Vectors represent genes. The biplot visualizes the association of cells with gene expression levels.

high degree of cellular plasticity. These differences observed among cells committed towards the same lineage argue for a democratic approach to lineage patterning, as has been proposed previously (Tabansky *et al.*, 2013, Zernicka-Goetz, 2013).

Regulatory networks reveal underlying mechanisms

ExprEssence and the PluriNetWork

Clustering and data reduction procedures do not usually take into consideration the relationships among genes. Therefore, the biological insight obtained from such analyses is often limited. Alternatively, many studies have attempted to infer gene regulatory networks from gene expression data (Marbach *et al.*, 2012). For example, Xue *et al.* (2013) inferred a co-expression network directly from their single-cell RNA-seq data. The construction of such networks is usually conceptually simple, and different frameworks have been proposed with that purpose (e.g. Zhang and Horvath, 2005). However, such approaches have only been shown to be successful for small, well defined systems (Marbach *et al.*, 2012). Alternatively, network approaches (reviewed in Ideker and Krogan (2012)) that integrate expression data and expert knowledge have been proposed as a compromise solution. Thus, Xie *et al.* (2010) explored the dynamics of the networks regulating embryonic development in three mammalian species using the program MATISSE (Ulitsky and Shamir, 2007). Given a high-throughput dataset and a network of genes and proteins, MATISSE identifies functional modules by comparing inferred and known interactions. Likewise, the purpose of ExprEssence (Warsow *et al.*, 2010), which relies on similar sources of biological information, is 2-fold: first, it reduces the complexity of large high-throughput datasets by focusing on known relevant genes or proteins; second, it identifies the interactions associated with the strongest concerted changes in expression between two biological states or conditions.

ExprEssence is a Cytoscape (Saito et al., 2012) plugin for analyzing biological networks together with high-throughput data across different conditions. Given, for example, two transcriptomic datasets and an interaction network, ExprEssence highlights concerted expression changes in pairs of genes/proteins that are known to interact with one another. For this purpose, ExprEssence computes a *LinkScore* for every edge in the network and pair of conditions. The *LinkScore* is a measure of how concerted gene expression changes are along network edges, that is, along gene/protein interactions. Thus, the tails of the distribution of *LinkScores* constitute good hypotheses for the strongest regulatory mechanisms controlling the change of interacting genes. The quality of the results obtained with ExprEssence depends on both the expression data and the functional network taken as input. Additionally, the fact that two genes that are known to interact exhibit strong, concerted changes in expression does not necessarily imply a causal mechanism, but such genes and their interaction warrant further theoretical and experimental work. ExprEssence has been shown to perform well in comparative analyses (Hatem et al., 2012).

Pairs of concertedly up-regulated or down-regulated genes could also be identified using standard univariate analysis. However, because ExprEssence relies on an independent interaction network, it is highly specific, and biological interpretation of results is guided by the network. Generally, the use of a network has two consequences: first, expert-curated knowledge reduces the false positive rate; second, insights are limited to the genes and interactions represented in the network. STRING (Jensen et al., 2009) is probably the largest and most widely-used protein–protein interaction database. STRING contains information from numerous sources, including experimental data, computational predictions and database mining. The current version of STRING contains information on over 5 million proteins from over a thousand species. BioGRID (Stark et al., 2006) is another database of curated interactions obtained from both high-throughput data and knowledge mining. Its current version contains ~700 000 interactions between proteins from almost 50 species. IntAct (Kerrien et al., 2012) is also a database of molecular interactions, populated by data either curated from the literature or from direct data depositions. It contains over 300 000 interactions. Finally, the Human Protein Reference Database (HPRD, Keshava Prasad et al., 2009) is a database of curated information on human proteins. This resource relies solely in experimentally derived information, and includes most human proteins. Its current version contains information on over 40 000 interactions. A drawback of the aforementioned resources is their limited contextual information. As cellular behavior is dynamic, interactions are expected to change depending on context. Therefore, databases describing interactions in a particular context are, in principle, more informative. For example, the PluriNetwork (Som et al., 2010) is a manually curated protein/gene interaction network that currently contains 348 genes and has been developed with the specific purpose of describing pluripotency in mouse. Thus, it considers the context of pluripotency, but it does not distinguish among the differences between types of pluripotency (e.g. naïve and primed states in mouse ESCs). Most importantly for the analysis of blastomere developmental data, it only describes one 'target' of cellular differentiation, i.e. pluripotency (in the ICM/pEct), but not the 'source' cell attribute of totipotency (in the early blastomeres), and the majority of the interactions it describes are based on knowledge inferred from cultured cells. To our knowledge, no expert-curated network of gene/protein relationships underlying totipotency has been published to date.

Analyzing the mechanisms behind cell fate decisions with ExprEssence

To illustrate the application and potential of the aforementioned analytical methods we analyzed the mouse single-blastomere data of Guo et al. (2010) and Deng et al. (2014) using ExprEssence in the context of the PluriNetwork. Guo et al. generated data on ~500 individual cells, an unprecedented and still unduplicated number, from the 1-cell zygote to the 64-cell blastocyst, using qPCR. qPCR is both accurate and sensitive, and indeed, it is routinely applied to validate data obtained by microarrays and RNA-seq (e.g. Griffith et al., 2010; Knight et al., 2014). Deng et al. (2014) investigated gene expression in ~250 single cells from embryos across similar developmental stages using RNA-seq. The dataset of Guo et al. (2010) comprises mostly genes that are known to be relevant to developmental biology. This makes it particularly suitable for demonstrating the aforementioned analytical methods. These methods, however, have been developed for use on a much broader scale, and, as we show, can be readily applied to datasets as large as that of Deng et al. (2014), although with some considerations.

In particular, we investigated the transitions from the 8-cell to the morula stage (16-cell stage), and from the morula to the blastocyst stage (32-cell stages). At the morula stage, the embryo consists of both inner and outer cells. The inner cells are thought to predominantly be the precursors of the ICM that is established at the blastocyst stage. The outer cells are more likely to contribute to the TE (Nishioka et al., 2009). The blastocyst itself consists of cells of the pEnd and pEct, which constitute the ICM, and of the TE, which covers the ICM (polar TE) and the cavity (mural TE).

First, from the 48 genes for which Guo et al. (2010) measured expression values, we selected those that were also part of the PluriNetWork. This step restricted the analysis to 13 genes, and to 51 interactions in the PluriNetWork involving pairs of these genes. Next, we applied ExprEssence to quantify changes in those 51 interactions between two developmental stages. For the transition from the 8-cell stage to morula (Fig. 2A) we identified a total of 26 concerted changes in pairs of interacting genes (see Supplementary data). Only seven of these changes correspond to significant changes in the expression of the corresponding genes (see Supplementary data). The strongest changes among these seven involve genes with stimulating roles. The two strongest changes correspond to the increase in the expression of *Utf1* and its stimulator *Pou5f1*, and *Cdx2*, which stimulates itself.

The large number of genes considered in high-throughput datasets, such as that of Deng et al. (2014), poses an interpretive challenge. Specifically, a fundamental problem in the analysis of high-throughput data is the identification of differentially expressed genes. A number of statistical frameworks and software packages have been developed for this task, including edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), baySeq (Hardcastle and Kelly, 2010), NOIseq (Tarazona et al., 2011), SAMseq (Li and Tibshirani, 2013) and Cuffdiff (Trapnell et al., 2013). Further, biological interpretation of differentially expressed genes is necessary to ascertain their relevance. Deng et al. (2014) detected the expression of ~20 000 protein-coding genes. We normalized the raw counts and analyzed the data for differentially expressed genes using DESeq (Anders and Huber, 2010) in R/Bioconductor (Gentleman et al., 2004). Seventy percent of the measured genes can be considered as expressed above background levels. Among those, we identified differentially expressed genes with a Benjamini-Hochberg

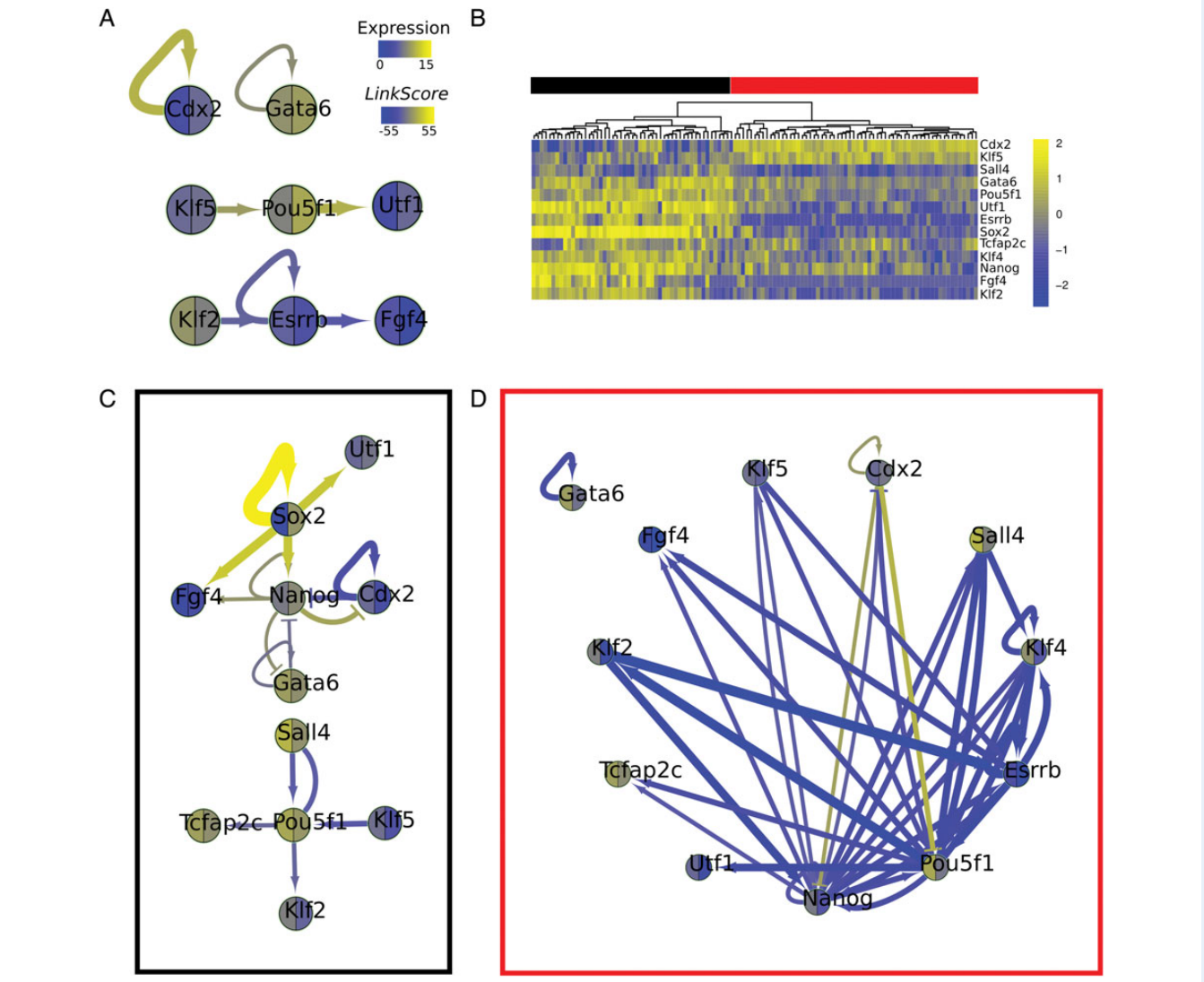

Figure 2 Concerted changes in pairs of genes for transitions between 8-, 16-, and 32-cell stages of mouse embryos. (**A**) Concerted changes in transitions from 8-cell stage to morula. The expression values for each gene at the two states considered are indicated by the color of the corresponding node (left half: 8-cell stage, right half: 16-cell stage). The direction and amount of change of a given interaction or *LinkScore* (see Supplementary data) is indicated by the color and width of the edges; in general, blue edges correspond to negative *LinkScore*s, and yellow edges to positive *LinkScore*s. (**B**) Hierarchical clustering of expression patterns in 32-cell stage cells. Genes and cells were clustered with a Pearson correlation distance metric and a Spearman rank correlation distance matrix metric, respectively, using complete linkage. Expression values for the 48 genes measured by Guo *et al.* (2010) were background-corrected, quantile-normalized and scaled across cells, on a per-gene basis (*z*-score) before clustering. Only genes that are part of the PluriNetWork (Som *et al.*, 2010) are visualized. (**C**) Concerted changes in transitions from morula to inner cell mass. (**D**) Concerted changes in transitions from morula to TE.

corrected *P*-value of <0.01. Approximately 3000 genes were considered differentially expressed between any pair of developmental stages. Strikingly, pluripotency markers such as *Pou5f1* and *Nanog* are not differentially expressed. Moreover, only 48 out of the ~3000 differentially expressed genes are included in the PluriNetWork. This lack of statistical enrichment may reflect the focus of the PluriNetWork on data derived from stem cells, and discrepancies with the biological functions and processes underlying embryonic development. Furthermore, it is noteworthy that mouse ESCs exist in at least two distinct states of pluripotency, naïve and primed, which resemble epiblast cells in embryonic days 4.5 and 5.5 mouse embryos, respectively. While all cultured pluripotent stem cells are thought to share the core pluripotency network of *Oct4*, *Nanog* and *Sox2*, different subtypes respond differently to certain signaling triggers. For example, FGF/Erk signaling promotes the transition from a naïve to a prime state, but prevents primed ESCs from reverting back to the naïve state (Nichols and Smith, 2009). The PluriNetWork describes direct interactions that have an influence on pluripotency in the mouse model system. The information on the interactions stems from all possible sources (different types of pluripotent embryonic cells and cultured pluripotent stem cells). Hence, the PluriNetWork can be regarded as broadly applicable to general pluripotency-related questions. However, it is less specific for the

analysis of data from early developing embryos, since the emergence of the pluripotent pEct is only one of the processes that unfold during the differentiation of the totipotent zygote to form a blastocyst containing all three germ layers. In addition, in this specific case, only a handful of genes are actually known whose mis-expression would halt the development of a mouse embryo prior to the blastocyst stage. Indeed, a PCA based on the ~3000 differentially expressed genes shows clusters of cells from embryos at the same developmental stage (Fig. 3A), suggesting that among these genes there are some that are either fundamental to development or tightly regulated by those genes that are. In any case, the set of identified differentially expressed genes most likely comprises both genes involved in specifying early lineage identities as well as false positives, and further experimentation (knockdown/overexpression of candidate genes) is necessary to confirm their individual roles. This is a typical example of how the bioinformatics analysis of high-throughput single-cell transcriptomics data can lead to the generation of novel hypotheses. Additionally, similarly to the analysis we performed on Guo *et al.*'s data, we explored the profiles of the ~3000 differentially expressed genes identified in Deng *et al.*'s data using ExprEssence and the PluriNetWork to identify the interactions associated with the strongest expression changes between the 8-cell stage and morula. Interestingly, the only significant concerted change corresponds to the self-activation of *Cdx2* as the expression of *Cdx2* increases. This result is in agreement with the critical role of *Cdx2* in TE specification discovered in the data of Guo *et al.* (2010) and reported in the literature (e.g. Jedrusik *et al.*, 2010).

The 32-cell stage cells investigated by Guo *et al.* (2010) can be clearly divided into two groups of similar size according to their expression

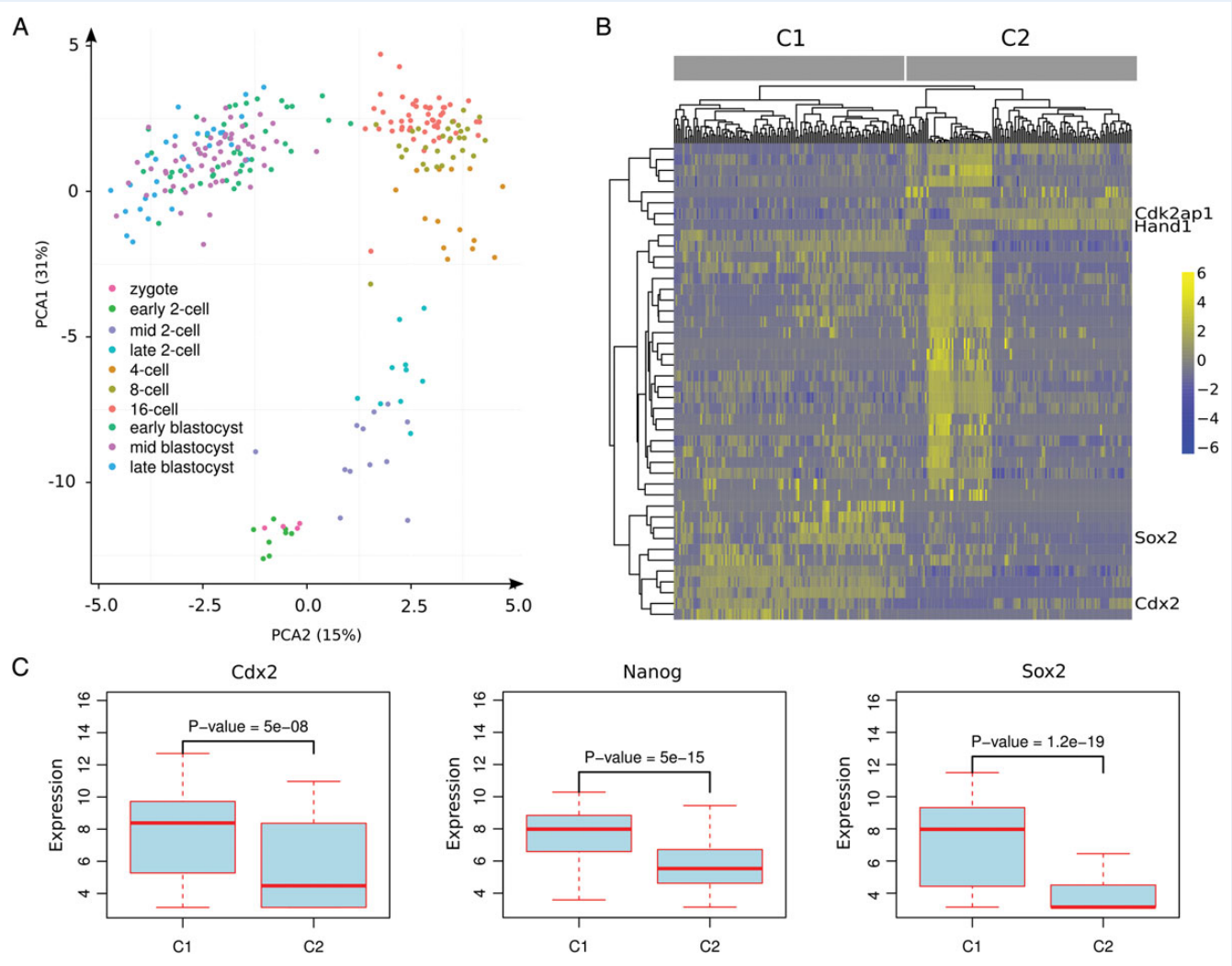

Figure 3 Clustering and visualization approaches applied to the expression data obtained for the mouse embryonic cells from zygote to late blastocyst by Deng *et al.* (2014). Analyses are based on the expression profiles of 48 genes that are both differentially expressed and part of the PluriNetWork (Som et al., 2010). Raw read counts for each protein-coding transcript were downloaded from the Gene Expression Omnibus (GEO, accession number GSM1112490). Normalization and differential expression analysis was performed using DESeq (Anders and Huber, 2010) in R/Bioconductor (Gentleman et al., 2004). Differentially expressed genes were identified as those genes with a Benjamini-Hochberg corrected *P*-value of <0.01. (**A**) PCA. (**B**) Hierarchical clustering. Genes and cells were clustered with a Pearson correlation distance metric and a Spearman rank correlation distance metric, respectively, using complete linkage. (**C**) Box plots indicating the expression values (log2) of three lineage-specific markers in the two main clusters defined by the hierarchical clustering in (B).

profiles (see Fig. 2B). Cells in one of the groups display significantly higher expression levels of *Pou5f1*, *Nanog*, *Gata6* and *Sox2*, and significantly lower expression levels of *Cdx2* when compared with the cells in the other group. This is consistent with differentiation into ICM and TE, respectively. From the 16-cell stage onwards, two clusters are also evident in the data generated by Deng *et al.* (2014) (Fig. 3B). However, the identity of the cells in the two clusters is ambiguous, at least if only traditional, well known cell lineage markers are considered. For example, the cells in one of the sub-clusters express relatively high levels of *Klf2* and *Klf4*, while the cells in the remaining sub-clusters express relatively high levels of *Cdx2*, consistent with differentiation into ICM and TE, respectively (Fig. 3C). However, those cells expressing relatively high levels of *Cdx2* also express *Nanog* and *Sox2*, which are markers characteristic of the ICM. Indeed, other genes, such as *Cdk2ap1* and *Hand1*, which have not been previously linked to mouse preimplantation development, appear to exhibit more regular expression patterns at these stages than those that are commonly assumed to be lineage-defining, and that are thus determining the clustering. This observation is rather unexpected. On the one hand, it could be an indicator of the high cellular plasticity within the developing embryo. For instance, not all lineage-defining genes will be expressed in all cells at exactly the same time, hindering the detection of a pattern. On the other hand, the genes dominating the clustering could constitute important genes that have not yet been linked to lineage decision processes. Additionally, the generation of the libraries is expected to have resulted in a certain amount of technical noise, affecting the effective coverage and read-depth, and ultimately, the differential expression analysis. In any case, novel candidates can only be revealed by unbiased approaches, such as RNA-Seq, and analytical techniques such as those presented here. Further, observed differences between the data presented by Guo *et al.* (2010) and Deng *et al.* (2014) might also be attributed to variations in the experimental protocols, such as mouse strains and pre-amplification protocols. Based on the above observations and for the sake of providing an easily comprehensible overview of bioinformatics approaches, we restricted the analysis of the transition from the morula to the 32-cell stage to the data generated by Guo *et al.* (2010).

For the transition from morula to ICM (Fig. 2C) we identified a total of 26 concerted changes in the Guo *et al.* (2010) data. Seventeen of these changes are considered significant. Concerted up-regulation of genes (in particular startups of stimulations, marked in yellow) are found in the network around *Sox2*, which is up-regulated in the ICM when compared with the morula. Other canonical pluripotency genes such as *Pou5f1* and *Sall4* and their interactions with stimulators such as *Klf2* and *Klf5* are down-regulated but still maintain relatively high levels. *Sox2* is a special case, because its expression is relatively low in the morula (the lowest across the developmental stages measured by Guo *et al.* (2010)). The up-regulation of *Sox2* in the ICM parallels the down-regulation of *Klf2*, reflecting the results of Guo *et al.* (2010). Furthermore, the up-regulation of *Sox2* is strikingly consistent across the single ICM cells. Indeed, despite the fact that the ICM lineage appears as a seemingly homogeneous group when compared with the TE, there is substantial variation among individual ICM cells (Fig. 2B), with *Sox2* being the notable exception. Based on expert knowledge in the PluriNetWork (in particular, from Chen *et al.* (2008) and Zhou *et al.* (2007)), we know that *Sox2* stimulates itself and *Nanog* (Fig. 2B) in ESCs, and we hypothesize that these stimulations are also relevant for the transition from morula to ICM. For the transition from morula to TE (Fig. 2D) we identified a total of 47 concerted changes in pairs of interacting genes. Most of them (38) also correspond to significant changes in the expression of the corresponding genes. Almost the entire pluripotency network is shut down (in Fig. 2D stimulation shutdowns are marked in blue; the few yellow edges indicate the startup of inhibitions involving *Cdx2*). The exception is *Sox2*, whose expression remains stable (and is therefore not shown in Fig. 2D). In fact, some expression of *Sox2* is necessary for the embryo to progress through the morula stage (Keramari *et al.*, 2010). Similarly to what we observed for the transition to ICM and consistent with the results of Guo *et al.* (2010), Fig. 2D shows the shutdown of interactions between the down-regulated pluripotency genes *Esrrb*, *Klf2*, *Nanog* and *Pou5f1*. Notably, in the single-cell data of Fig. 2B, *Cdx2* stands out as the most consistently up-regulated gene in the TE. Figure 2D suggests the hypothesis that its main function is the repression of *Nanog* and *Pou5f1* (interactions in the network are based on Niwa *et al.* (2005) and Chen *et al.* (2009)). This hypothesis has been in fact recently confirmed (Wu *et al.*, 2010). Finally, we conjecture that the largely homogenous expression of *Sox2* and *Cdx2* across single ICM and TE cells, respectively, is likely to be ascribed to the cell location-dependent hippo signaling pathway. In outside cells, which develop into TE, the hippo pathway is inactive, leading to *Tead4*-induced *Cdx2* expression (Nishioka *et al.*, 2009). Conversely, the hippo pathway is active in inside cells developing into ICM, where it would control *Sox2*. Indeed, *Sox2* expression is absent when the hippo pathway is disrupted by knocking down its crucial kinase components *Lats1/2* (Lorthongpanich *et al.*, 2013). These observations argue for a regulation of *Cdx2* and *Sox2* based on cellular location, which might explain their homogenous expression levels when compared with other genes that are lower in the regulatory hierarchy. These examples illustrate how the use of a knowledge network that has been curated by experts in the field can be combined with expression data to generate plausible hypotheses. Specifically, the single-cell data highlight Sox2 in the ICM and Cdx2 in the TE, and data integration suggested specific functional hypotheses for these two genes. It goes without saying that such hypotheses then require experimental confirmation.

Conclusions

Single-blastomere transcriptomics has turned out to be extraordinarily helpful for developmental biology in general, and for early mouse embryology in particular. Bioinformatics tools are an absolute necessity in order to interpret and derive biological insight from the extensive datasets generated with current sequencing technologies. As highlighted above, in particular large datasets on single cells may lead to results whose interpretation is not straightforward. However, they also open the way for the formulation of new hypotheses in an unprecedented manner. In the years to come, we expect single-blastomere transcriptomics to become even more precise and to be followed by reliable single-blastomere epigenomics, and, in the long term, single-blastomere proteomics. Scientific progress will certainly not stop then. On the horizon are 'omics' investigations into subcellular compartments, with high resolution along the temporal axis, towards the ultimate goal of obtaining a mechanistic model of the earliest developmental steps—one of the true miracles of life.

Supplementary data

Supplementary data are available at http://molehr.oxfordjournals.org/.

Authors' roles

L.T., M.J.P. and G.F. conceived the outline of the review, analyzed and interpreted published data, drafted and revised the manuscript. All authors approved the final manuscript.

Funding

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, grants FU583/2-2 and BO2540/3-2).

Conflict of interest

None declared.

References

Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.

Ayub M, Hardwick SW, Luisi BF, Bayley H. Nanopore-based identification of individual nucleotides for direct RNA sequencing. *Nano Lett* 2013; **13**:6144–6150.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X *et al.* The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008;**26**:1146–1153.

Buettner F, Theis FJ. A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* 2012; **28**:i626–i632.

Casanova EA, Okoniewski MJ, Cinelli P. Cross-species genome wide expression analysis during pluripotent cell determination in mouse and rat preimplantation embryos. *PLoS One* 2012;**7**:e47107.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008; **133**:1106–1117.

Chen L, Yabuuchi A, Eminli S, Takeuchi A, Lu C-W, Hochedlinger K, Daley GQ. Cross-regulation of the Nanog and Cdx2 promoters. *Cell Res* 2009;**19**:1052–1061.

Deng Q, Ramsköld D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014;**343**:193–196.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**:14863–14868.

Gasch AP, Eisen MB. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol* 2002; **3**:RESEARCH0059.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**:R80.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ *et al.* Alternative expression analysis by RNA sequencing. *Nat Methods* 2010;**7**:843–847.

Guo G, Huss M, Tong GQ, Wang C. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev Cell* 2010;**18**:675–685.

Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 2010; **11**:422.

Hatem A, Kaya K, Çatalyürek ÜV. Microarray vs. RNA-Seq: a comparison for active subnetwork discovery. *ACM Conference on Bioinformatics, Computational Biology and Medicine*, 2012.

Huang D, Pan W. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics* 2006;**22**:1259–1268.

Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol* 2012;**8**:565.

Jedrusik A, Bruce AW, Tan MH, Leong DE, Skamagki M, Yao M, Zernicka-Goetz M. Maternally and zygotically provided Cdx2 have novel and critical roles for early development of the mouse embryo. *Dev Biol* 2010;**344**:66–78.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M *et al.* STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2009;**37**:D412–D416.

Kang M, Piliszek A, Artus J, Hadjantonakis AK. FGF4 is required for lineage restriction and salt-and-pepper distribution of primitive endoderm factors but not their initial expression in the mouse. *Development* 2013; **140**:267–279.

Keramari M, Razavi J, Ingman KA, Patsch C, Edenhofer F, Ward CM, Kimber SJ. Sox2 is essential for formation of trophectoderm in the preimplantation embryo. *PLoS One* 2010;**5**:e13952.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012; **40**:D841–D846.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al.* Human Protein Reference Database—2009 update. *Nucleic Acids Res* 2009;**37**:D767–D772.

Knight JM, Davidson LA, Herman D, Martin CR, Goldsby JS, Ivanov IV, Donovan SM, Chapkin RS. Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing. *Sci Rep* 2014;**4**:5453.

Krawchuk D, Honma-Yamanaka N, Anani S, Yamanaka Y. FGF4 is a limiting factor controlling the proportions of primitive endoderm and epiblast in the ICM of the mouse blastocyst. *Dev Biol* 2013;**384**:65–71.

Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* 2013;**22**:519–536.

Lorthongpanich C, Messerschmidt DM, Chan SW, Hong W, Knowles BB, Solter D. Temporal reduction of LATS kinases in the early preimplantation embryo prevents ICM lineage differentiation. *Genes Dev* 2013;**27**:1441–1446.

Luo J, Sladek R, Bader JA, Matthyssen A, Rossant J, Giguère V. Placental abnormalities in mouse embryos lacking the orphan nuclear receptor ERR-beta. *Nature* 1997;**388**:778–782.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;**9**:796–804.

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* 2014;**24**:496–510.

Martens L, Apweiler R. Algorithms and databases. *Methods Mol Biol* 2009; **564**:245–259.

Miyanari Y, Torres-Padilla ME. Control of ground-state pluripotency by allelic regulation of Nanog. *Nature* 2012;**483**:470–473.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008; **5**:621–628.

Nichols J, Smith A. Naive and primed pluripotent states. *Cell Stem Cell* 2009; **4**:487–492.

Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Schöler H, Smith A. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 1998;**95**:379–391.

Ning L, Liu G, Li G, Hou Y, Tong Y, He J. Current Challenges in the Bioinformatics of Single Cell Genomics. *Front Oncol* 2014;**4**:7.

Nishimoto M, Katano M, Yamagishi T, Hishida T, Kamon M, Suzuki A, Hirasaki M, Nabeshima Y, Katsura Y, Satta Y *et al. In vivo* function and evolution of the eutherian-specific pluripotency marker UTF1. *PLoS One* 2013;**8**:e68119.

Nishioka N, Inoue K, Adachi K, Kiyonari H, Ota M, Ralston A, Yabuta N, Hirahara S, Stephenson RO, Ogonuki N *et al.* The Hippo signaling pathway components Lats and Yap pattern Tead4 activity to distinguish mouse trophectoderm from inner cell mass. *Dev Cell* 2009;**16**:398–410.

Niwa H, Toyooka Y, Shimosato D, Strumpf D, Takahashi K, Yagi R, Rossant J. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* 2005;**123**:917–929.

Ohnishi Y, Huber W, Tsumura A, Kang M, Xenopoulos P, Kurimoto K, Oleś AK, Araúzo-Bravo MJ, Saitou M, Hadjantonakis AK *et al.* Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat Cell Biol* 2014; **16**:27–37.

Parisi S, Russo T. Regulatory role of Klf5 in early mouse development and in embryonic stem cells. *Vitam Horm* 2011;**87**:381–397.

Pfeiffer MJ, Esteves TC, Balbach ST, Araúzo-Bravo MJ, Stehling M, Jauch A, Houghton FD, Schwarzer C, Boiani M. Reprogramming of two somatic nuclei in the same ooplasm leads to pluripotent embryonic stem cells. *Stem Cells* 2013;**31**:2343–2353.

Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 2006;**4**:e309.

Robert C. Microarray analysis of gene expression during early development: a cautionary overview. *Reproduction (Cambridge, England)* 2010; **140**:787–801.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–140.

Rodriguez-Zas SL, Schellander K, Lewin HA. Biological interpretations of transcriptomic profiles in mammalian oocytes and embryos. *Reproduction* 2008;**135**:129–139.

Rossant J, Tam PP. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* 2009; **136**:701–713.

Rusnakova V, Honsa P, Dzamba D, Ståhlberg A, Kubista M, Anderova M. Heterogeneity of astrocytes: from development to injury—single cell gene expression. *PLoS One* 2013;**8**:e69734.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. A travel guide to Cytoscape plugins. *Nat Methods* 2012;**9**:1069–1076.

Som A, Harder C, Greber B, Siatkowski M, Paudel Y, Warsow G, Cap C, Schöler H, Fuellen G. The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. *PLoS One* 2010;**5**:e15165.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–D539.

Tabansky I, Lenarcic A, Draft RW, Loulier K, Keskin DB, Rosains J, Rivera-Feliciano J, Lichtman JW, Livet J, Stern JN *et al.* Developmental bias in cleavage-stage mouse blastomeres. *Curr Biol* 2013;**23**:21–31.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999;**96**:2907–2912.

Tan MH, Au KF, Leong DE, Foygel K, Wong WH, Yao MW. An Oct4-Sall4-Nanog network controls developmental progression in the pre-implantation mouse embryo. *Mol Syst Biol* 2013;**9**:632.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**:377–382.

Tang F, Barbacioru C, Nordman E, Bao S, Lee C, Wang X, Tuch BB, Heard E, Lao K, Surani MA. Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS One* 2011;**6**:e21208.

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. *Genome Res* 2011; **21**:2213–2223.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* 1999; **22**:281–285.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013;**31**:46–53.

Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol* 2007;**1**:8.

Van der Laan M, Pollard K, Bryan J. A new partitioning around medoids algorithm. *J Stat Comput Simul* 2003;**73**:575–584.

VerMilyea MD, Maneck M, Yoshida N, Blochberger I, Suzuki E, Suzuki T, Spang R, Klein CA, Perry AC. Transcriptome asymmetry within mouse zygotes but not between early embryonic sister blastomeres. *EMBO J* 2011;**30**:1841–1851.

Warsow G, Greber B, Falk SS, Harder C, Siatkowski M, Schordan S, Som A, Endlich N, Schöler H, Repsilber D *et al.* ExprEssence—revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol* 2010;**4**:164.

Wennekamp S, Hiiragi T. Stochastic processes in the development of pluripotency *in vivo*. *Biotechnol J* 2012;**7**:737–744.

Wu G, Gentile L, Fuchikami T, Sutter J, Psathaki K, Esteves TC, Araúzo-Bravo MJ, Ortmeier C, Verberk G, Abe K *et al.* Initiation of trophectoderm lineage specification in mouse embryos is independent of Cdx2. *Development* 2010;**137**:4159–4169.

Wu G, Han D, Gong Y, Sebastiano V, Gentile L, Singhal N, Adachi K, Fischedick G, Ortmeier C, Sinn M *et al.* Establishment of totipotency does not depend on Oct4A. *Nat Cell Biol* 2013;**15**:1089–1097.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* 2014; **11**:41–46.

Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res* 2010;**20**:804–815.

Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 2013;**500**:593–597.

Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 2013;**20**:1131–1139.

Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics* 2001;**17**:763–774.

Zernicka-Goetz M. Development: do mouse embryos play dice? *Curr Biol* 2013;**23**:R15–R17.

Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;**4**:Article17.

Zhou Q, Chipperfield H, Melton DA, Wong WH. A gene regulatory network in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 2007; **104**:16438–16443.