Developmental Biology 353 (2011) 440–447

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/developmentalbiology

Genomes & Developmental Control

Maximum parsimony analysis of gene expression profiles permits the reconstruction of developmental cell lineage trees

Anagha Joshi, Berthold Göttgens *

Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge, CB2 0XY, UK

ARTICLE INFO

Article history:
Received for publication 14 December 2010
Revised 7 February 2011
Accepted 16 February 2011
Available online 23 February 2011

Keywords:
Gene expression
Maximum parsimony
Development and differentiation
Haematopoiesis
Cell lineage trees

ABSTRACT

Spatiotemporal control of gene expression lies at the heart of generating several hundred distinct cell types required for the development of higher order animals. Different cell types within complex organs are often characterised by means of genome-wide gene expression profiling, but analogous information for early developmental as well as adult stem and progenitor cells is largely missing because their identity is commonly unknown or they are present in prohibitively small numbers. Here we show that maximum parsimony approaches previously used to reconstruct evolutionary trees from gene content of extant species can be adapted to reconstruct cellular hierarchies both during development and steady state homeostasis of complex mammalian tissues. Using haematopoiesis as a model, we show that developmental trees reconstructed from expression profiles of mature cells are not only consistent with current experimentally validated trees but also have predictive value in determining progenitor cell specific transcriptional programmes and lineage determining transcription factors. Subsequent analysis across diverse developmental systems such as neuronal development and endoderm organogenesis demonstrated that maximum parsimony-based reconstruction of developmental trees represents a widely applicable approach to infer developmental pathways as well as the transcriptional control mechanisms underlying cell fate specification.

Introduction

The generation of specialised cell types from multipotent progenitors underlies all metazoan development. Formation of the several hundred distinct mature cell types present in complex organisms occurs via a carefully orchestrated sequence of events starting from the pluripotent fertilised egg via the three germ layers and tissue-specific stem cells to even more restricted progenitors ultimately culminating in the generation of terminally differentiated mature cells that constitute the bulk of adult tissues. Maintenance and repair of adult tissues are similarly ensured through the continuous formation of mature cells from multipotent tissue-specific stem cells via clearly defined intermediate stages and cell-fate choices. Moreover, perturbation of this carefully balanced tissue homeostasis results in under- or over-production of relevant cell types and thus contributes to the development of degenerative diseases or cancer. A more thorough understanding of cellular differentiation trees is therefore of fundamental importance not just to the field of developmental biology but also to giving insight into the pathogenesis of major human diseases.

Cell differentiation or lineage trees have been studied using a range of techniques. The most comprehensive results have been generated by direct observation of developing embryos in the nematode *Caenorhabditis elegans* (Sulston et al., 1983). The generation of similarly complete cell lineage trees for more complex organisms is not only complicated by the vast increase in cell numbers, but also because embryonic development in mammals occurs *in utero* and can therefore not be observed directly *in situ*. Approaches such as injection of lineage tracers (Lawson and Pedersen, 1992), genetic single-cell labelling *in utero* (Tzouanacou et al., 2009) or analysis of somatic mutations (Salipante et al., 2010; Wasserstrom et al., 2008) have all been used successfully to clarify lineage relationships for specific aspects of mouse development. Moreover, continuous imaging of cultures stem and progenitor cells has provided important insights into embryonic stem and haematopoietic progenitor cell differentiation (Eilken et al., 2009; Rieger et al., 2009), but the technology cannot at present be adapted for the *in vivo* study of mammalian embryos or tissues. Our current understanding of cell lineage/differentiation trees during mammalian development remains incomplete. Thus, little is known for most tissues with regard to the cellular differentiation trees operating during tissue homeostasis, little is known for most tissues. Concerted cell biological and molecular studies have generated deep insights for some homeostatic tissues. Particularly for the haematopoietic system, where not only the haematopoietic stem cell (HSC), but also a plethora of intermediate progenitors with increasingly restricted lineage potential as well as more than 10 distinct mature cell types can all be isolated to near purity using combinations of cell surface markers (Bryder et al., 2006).

* Corresponding author. Fax: +44 1223 762670.
E-mail address: bg200@cam.ac.uk (B. Göttgens).

0012-1606/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ydbio.2011.02.013

Moreover, while some branchpoints during the specification of early multipotent progenitors are still the topic of active scientific debate (Adolfsson et al., 2005; Forsberg et al., 2006), powerful functional assays for both stem and progenitor cells have allowed the development of a broadly accepted view of the haematopoietic differentiation hierarchy, with a level of detail far exceeding the information available for any other adult stem cell system (Orkin and Zon, 2008).

The different cellular phenotypes along a developmental maturation pathway are commonly viewed as distinct gene expression states (Enver et al., 2009) with transitions between states mediated through interactions of extracellular signalling activities with the intracellular transcriptional machinery (Davidson, 2006). Transitions between gene expression states therefore involve the activation as well as the repression of specific gene subsets, and these changes can theoretically be monitored at the level of the whole genome using a number of gene expression profiling strategies. However, immature stem and progenitor populations are often inaccessible at acceptable high purities or frequencies. As a result, gene expression profiling information is often restricted to mature cell populations.

In evolutionary biology, the availability of completely sequenced genomes for a wide range of extant species has spurred the development of new computational approaches to simulate gains and losses of individual genes in presumed ancestral species. The resulting information has been used to reconstruct phylogenetic trees based on the maximum parsimony (MP) principle (Martens et al., 2008; Wapinski et al., 2007).

Here, we have explored maximum parsimony analysis of global gene expression patterns (Kluger et al., 2004) to reconstruct developmental trees from which the expression states of intermediate progenitors can be predicted, based on an analysis of gene expression profiles of mature cell types. We validated our approach using gene expression profiles from a range of mature haematopoietic cell types which resulted in the prediction of intermediate level progenitors and a differentiation tree consistent with experimental data. Moreover, loss or gain of gene expression could be directly correlated with the cellular function of individual haematopoietic cell types and, at a global level highly-correlated with lineage-specific transcriptional control mechanisms inferred from genome-wide transcription factor binding site analysis. Subsequent reconstruction of developmental trees for neuronal development and endoderm organogenesis demonstrated that maximum parsimony-based analysis of gene expression profiles represents a widely applicable approach to infer lineage trees and transcriptional control mechanisms for a wide range of developmental pathways.

Materials and methods

Gene expression data collection and discretization

Gene expression profiles for haematopoietic cell types (Chambers et al., 2007), neural development (Cahoy et al., 2008) and endoderm organogenesis (Sherwood et al., 2009) were discretized using two discretization methods—Constant cut-off and %X Max. For the constant cut-off method, the gene expression value in a given condition is set to 1 if its level of expression is greater than a constant value. We used 5 cut-offs—50, 100, 150, 200 and 250. For the X% Max method, a gene expression value in a given condition is set to 1 if its level of expression is in X% of the highest value (X = 30, 40, 50, 60, 70). Though each cut-off resulted in a different number of genes expressed in each cell type, the resultant tree structure was largely unaffected by the discretization method and the cut-off used (see Supplementary data, Section 2 for haematopoietic differentiation trees derived using two different discretization methods at various cut-offs). We used the tree reconstructed using a constant cut-off of 100 (mean of expression values across all conditions) for biological interpretation. The number of parsimony informative sites was 5320, 5122, and 1460 for the three datasets. The input files for tree reconstruction are provided in the Supplementary materials.

Inference of expression states and state changes

Expression states and state changes were mapped onto the branches of inferred cell differentiation trees using the PARS program from the PHYLIP package (Felsenstein, 1996). Haematopoietic stem cells were defined as the root of the tree for the haematopoiesis datasets while forebrain was selected as a root for the neural data and endoderm at day 8.5 for endoderm data.

Bootstrapping

Bootstrapping was performed using the SEQBOOT program from the PHYLIP package where 100 datasets were generated by randomly replacing a given discretized expression matrix. A consensus tree with a bootstrap confidence on each branch of the tree was reconstructed using the CONSENSE program from the PHYLIP package (Felsenstein, 1996).

Gene ontology

Gene set enrichment for each state change was calculated using the DAVID suite of programs (Huang et al., 2008).

Transcription factors

A list of transcription factors in mouse was obtained from DBD (Wilson et al., 2007).

State-change validation sets

ChIP-seq data: ChIP-seq data for validating the haematopoietic developmental tree was compiled from diverse publications (Wilson et al., 2010; Soler et al., 2010; Ouyang et al., 2009) and gene lists were derived from peak coordinates using the method described in (Wilson et al., 2010), if not provided in the original paper. ChIP-seq datasets for validating the endoderm developmental tree were obtained from Schmidt et al. (2010). For each transcription factor, enrichment of overlap of the candidate target gene set with each transition state gene set was calculated using a hypergeometric test.

Phenotypic data: Phenotypic data was obtained from the Jackson lab mouse genome informatics (MGI) database.

Normal and leukaemic progenitor datasets: 9 gene expression signatures (d-erythroid, differentiated, d-lymphoid, d-myeloid, r-myeolymphoid, s-erythroid, s-mpp, s-myelolymphoid and stem) were obtained from Ng et al. (2009). Enrichment (P value) of each signature gene set with respect to each state transition gene set was calculated using hypergeometric test.

Gene sets differentially expressed in acute myeloid leukaemia (AML) from two groups—high and low based on the percentage of leukaemia stem cells were obtained from (Somervaille et al., 2009) and enrichment for each state transition gene set was calculated using hypergeometric test.

Results

Reconstruction of the haematopoietic differentiation tree from maximum parsimony (MP) analysis of gene expression profiles

Recent genome-scale phylogenetic studies have illustrated the possibility of reconstructing evolutionary history based on gene content of extant species together with an to predict genomic content of ancestral species (Martens et al., 2008). Following the same

principle, we reasoned that it might be possible to reconstruct developmental lineage trees from the expression data of mature cell types. Just as for the phylogenetic trees, this analysis would not only highlight the differences between distinct mature cells but also be able to predict expression states of 'ancestral' progenitor cells. Analogous to phylogenetic tree reconstruction, we used the concept of parsimony, which assumes that the variation in gene expression between the different mature cell types is achieved by the minimum number of expression changes during differentiation.

To attempt developmental tree reconstruction from haematopoietic gene expression data, we first discretized the gene expression matrix from the recently published haematopoietic fingerprints compendium (Chambers et al., 2007) into a binary matrix (see Materials and methods for details). In the resulting matrix, each element a_{ij} denotes whether the gene i is active in the cell type j or not. Although discretization reduces information content of the data, this loss of information is at least in part compensated for by the fact that discretization can effectively reduce background noise, an inherent problem with the low signal to noise ratio found in many microarray datasets (Zhang et al., 2009). The haematopoietic fingerprints dataset contains gene expression profiles for 9 mature cell types as well as highly purified HSCs. 6093 genes were expressed in at least one cell type but not ubiquitously across all cell types, resulting in a discretized expression matrix of 10 columns (the 10 cell types) and 6093 rows (number of genes).

Using this dataset, a rooted tree was reconstructed using maximum parsimony approach from the 9 mature cell types and HSCs assuming that the HSCs are at the root of the tree. This approach resulted in the reconstruction of a haematopoietic differentiation tree (Fig. 1) that is in very good agreement with the experimentally defined consensus (Orkin and Zon, 2008). Importantly, not only were the major divisions into the myeloid and lymphoid arms as expected but also the prediction of intermediate precursors within these major partitions was in good agreement with the literature. To ensure the robustness of these results, we used a range of different discretization methods (see Materials and methods for details) which demonstrated that the resultant tree structure was independent of various discretization methods. Moreover we employed a standard bootstrap procedure (see Materials and methods for details) to estimate confidence at each branch point in the predicted tree. The maximum parsimony tree splits correctly into myeloerythroid and lymphoid lineages with 100% bootstrap confidence. Similarly, lineage separations within the myeloerythroid branch and the separation between B-cells and T-cells were obtained with 100% bootstrap confidence with the separation of NK cells and divisions within the T-cell lineage having somewhat lower levels of bootstrap confidence. Taken together, the reconstructed MP tree is consistent with the developmental history, thus suggesting that reconstruction of developmental lineage trees from expression data of mature cell types is indeed possible. Importantly, this is in marked contrast to trees reconstructed from the same expression dataset using traditional microarray analysis tools such as hierarchical clustering based on global distance-based measures, which failed to recapitulate the developmental history of haematopoietic differentiation (Chambers et al., 2007) Moreover, discretization of the expression data was not responsible for this apparent failure because clustering of discretized data using several dendrogram-based methods did not improve the grouping of related cell lineages (see Supplementary material).

Analysis of expression state changes

Cell differentiation trees, reconstructed using our new maximum parsimony-based approach, model the developmental state transitions at each branch in terms of genes activated or repressed during a given transition (Fig. 1). The differentiation tree reconstructed from the haematopoietic fingerprints dataset contains 17 branch points with 17 gene expression change events (numbered 1–17 in Fig. 1). We therefore wanted to explore whether these modelled gene expression state changes could inform the underlying biological mechanisms at play. One of the most notable observations was the extensive gene expression loss from the HSCs to the two multipotent progenitor states (branch 1, Fig. 1). Gene ontology analysis of the genes repressed at this first transition revealed an overrepresentation of genes involved in cell adhesion, cell–cell junction and endothelial development (Table 1). This is consistent with the hypothesis that cell–cell interactions may be more important for HSCs that are located in

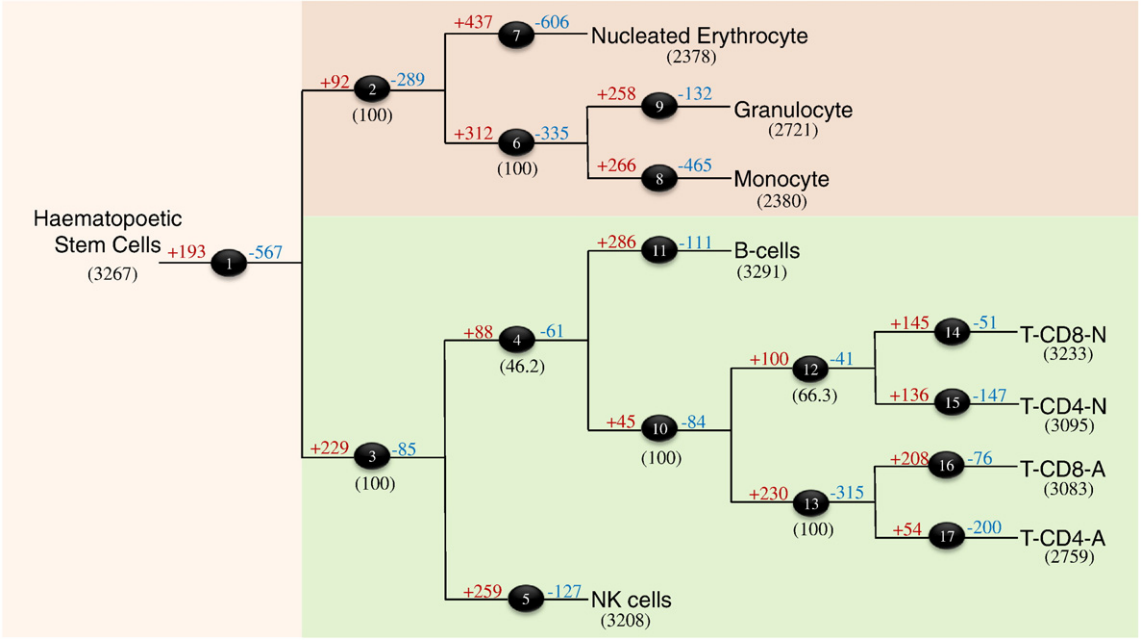

Fig. 1. Cell lineage differentiation tree encompassing 10 haematopoietic cell types with the total number of differentially expressed active genes in each cell type shown in brackets. Each branch, (numbered 1–17) represents a predicted developmental state transition. At each state transition, the number of genes predicted to be activated (red) or repressed (blue) as well as the bootstrap value for each branch (in brackets) is indicated. Supplementary Fig. 2 shows the same tree including the total number of genes predicted to be expressed at each node.

Table 1
Functional categories overrepresented for gene sets from inferred developmental state transitions during haematopoiesis along with p values.

Branch	On/off	Functional category	P value
1	Off	Cell junction	7.57E−06
		Focal adhesion	3.12E−04
		Cell adhesion	3.70E−04
		Biological adhesion	3.86E−04
		Blood vessel development	2.01E−03
		Vasculature development	2.52E−03
		Frizzled protein	6.75E−05
2	On	Defense response	2.08E−04
		Chemotaxis	9.62E−05
3	On	Immunoglobulin production	7.33E−05
4	On	Regulation of lymphocyte activation	1.58E−02
5	On	Natural killer cell mediated cytotoxity	1.38E−08
6	On	Response to molecule of bacterial origin	2.25E−06
		Response to lipopolysaccharide	1.01E−04
		Macrophage activation	6.48E−03
		Myeloid leukocyte activation	9.23E−03
		Inflammatory response	4.50E−07
		Defense response	3.07E−06
		Innate immune response	5.56E−05
		Regulation of interleukin-6 production	2.35E−06
7	Off	Cell death	3.02E−04
7	On	Erythrocyte development	2.02E−02
		Porphyrin metabolic process	8.92E−04
8	On	Blood circulation	1.61E−05
		Response to wounding	2.86E−04
8	Off	Cell cycle	1.09E−23
		DNA metabolic process	2.71E−20
10	On	T-cell receptor complex	1.15E−06
11	On	B-cell receptor signalling pathway	1.67E−05
16	On	Cytokine	4.50E−05

specific niches in contrast to the more mature blood cells which are predominantly in the circulation and also that HSCs share aspects of their transcriptional programmes with endothelial cells (Jaffredo et al., 2005; Silberstein et al., 2005). The prevalence of gene repression (567) over activation (193) at this transition was also consistent with the widely accepted notion that stem cells express a wide spectrum of genes various mature descendents with initial lineage specification decisions largely mediated through repressing alternative cell-fate gene expression programmes (the concept of "multi-lineage priming" (Hu et al., 1997).

Gene ontology analysis of the other modelled state transitions was also consistent with known cellular phenotypes (Table 1). For example, erythrocyte lineage-specific genes (branch 7) are overrepresented for erythrocyte development and heme related processes, NK cells (branch 5) with Natural killer cell cytotoxity, T-cells (branch 10) and B-cell (branch 11) with T-cell and B-cell receptor signalling respectively. Moreover, 8 out of 17 expression state-change gene sets are overrepresented ($P<0.05$) for 'hematopoietic system phenotype' as annotated by the Jackson lab mouse genome informatics (MGI) database. Taken together therefore, functional and phenotypic enrichment provides validation for the predicted state transitions.

Comparison of predicted expression states with normal and leukaemic haematopoietic progenitor cells

We next explored whether our predicted expression states and expression changes for intermediate cell types were relevant when compared with existing expression data for both normal and leukaemic stem/progenitor cells. A recent transcriptome analysis of HSCs and early multi-, bi- and unipotent progenitors (Ng et al., 2009) reported 9 gene expression signatures ranging from those characteristic for the most immature HSCs via multilineage progenitors to those affiliated with differentiation into the individual haematopoietic lineages. These experimentally obtained gene signatures provided an ideal test case to examine the relevance of the expression states for intermediate progenitors predicted by our parsimony-based developmental tree reconstruction. Comprehensive analysis of all gene signatures across all branch points within our MP tree demonstrated striking correlations between experimentally and computationally derived gene sets. For example, the HSC-specific genes from Ng et al (their 'stem' signature) showed statistically significant overlap with genes downregulated at our computationally predicted branch 1 but no other branchpoints. Similarly, the later erythroid signature from Ng et al. ('d-erythroid' signature) showed statistically significant upregulation at our branch 7 (mature erythroid differentiation) as well as statistically significant downregulation in the adjacent branch 6 (corresponding to maturing myeloid lineage cells). Hence, these results suggest that the expression states and state transitions predicted for intermediate stages within MP-reconstructed developmental lineage trees encapsulate important aspects of *in vivo* expression data.

To explore whether MP-reconstructed trees could also be exploited to interrogate gene expression profiles from malignant cells, we next analysed gene expression profiles from mouse models of acute myeloid leukaemia (AML) that have recently been classified into two groups based on the percentage of leukemic stem cells (Somervaille et al., 2009). It has long been thought that AMLs are sustained by a population of leukaemic stem cells that are phenotypically similar to normal HSCs. However, a recent study (Somervaille et al., 2009) showed that leukemic stem cells are actually more similar to blood progenitors located at intermediate levels within the haematopoietic hierarchy than to the stem cells at the top of the hierarchy. In agreement with this notion, we observed that the expression profiles from AML samples with a high proportion of leukaemia stem cells were not associated with the HSC expression state in our tree but instead showed statistically significant association with downregulation only during the later stages of myeloid differentiation (branches 6 and 8—Fig. 1). Therefore, analysis of both normal and malignant haematopoietic progenitor populations underlined the potential biological relevance of intermediate state transitions predicted by cell lineage tree reconstruction.

Interrogating transcriptional mechanisms underlying developmental state transitions

A number of transcription factors (TFs) have been demonstrated to function as key regulators of haematopoietic differentiation (Pimanda and Göttgens, 2010). Given that our reconstructed trees were based on expression profiles, we next focussed our analysis on the subsets of up/downregulated genes at inferred state transitions that encode transcriptional regulators. Gene ontology analysis showed that the transcription factor subsets at 14 of the 17 branches within our MP tree were overrepresented for 'Haematopoietic system phenotype', demonstrating that our analysis places TFs with known haematopoietic function into specific differentiation branch point. Moreover, overrepresentation of known regulators suggests that some of the TFs with unknown function are also likely to encode important haematopoietic regulators. For example, 33 TFs were activated at branch 7 (development of erythrocytes; see Fig. 1), and these included the erythroid master regulators Gata1 and Klf1 (see Supplementary Table 1 for full lists associated with all branches).

Given that our MP developmental lineage trees are reconstructed from gene expression data and that comparisons with experimental datasets had validated computationally predicted state transitions, we next explored whether MP developmental lineage trees would allow us to make inferences on the transcriptional mechanisms predicted to underlie haematopoietic cell state transitions. Genome-wide identification of transcription factor binding events by chromatin-immunoprecipitation coupled to high-throughput sequencing (ChIP-Seq) has the potential to link gene expression patterns to the binding of candidate upstream regulators. We therefore compared the gene sets

Table 2
Expression signatures with significant overlap (P value<0.001) with gene sets from inferred developmental state transitions during haematopoiesis. Detailed information on gene overlaps is provided in supplementary materials.

Branch	On / off	Expression signatures
1	Off	Stem
3	Off	Stem
7	Off	s-myelolymphoid, s-mpp, s-erythroid
7	On	d-erythroid
8	Off	Diff

obtained from all 17 state transitions in our MP tree (Fig. 1) with recently published candidate target genes obtained by ChIP-Seq analysis of 10 transcription factors in a multipotential haematopoietic progenitor cell line (Wilson et al., 2010) as well as the Gata1 and Klf1 transcription factors in erythroid cells (Soler et al., 2010). We also included in this analysis a 12 factor ChIP-Seq dataset from mouse embryonic stem (ES) cells (Ouyang et al., 2009) to explore whether candidate ChIP-Seq targets from a non-haematopoietic cell type may show specific associations with computationally inferred state transitions in our haematopoietic tree.

Gene set enrichment was calculated for the gene sets up- and downregulated for each of the 17 modelled state transitions and compared with the ChIP-Seq candidate targets from the four studies outlined in the previous paragraph. Table 2 summarises all statistically significant associations (P<0.001) found between transcription factor candidate targets and specific lineage transition events. Gata1 and Klf1 are widely recognised as the major drivers of terminal erythroid differentiation and specific association of both factors with genes upregulated at terminal erythroid differentiation was observed in our MP tree. Moreover, targets of five stem cell transcription factors (Gfi1b, Runx1, Erg, Meis1 and Fli1) were associated with genes extinguished at this same transition. Runx1, Erg, Meis1 and Fli1 are thought to encode transcriptional activators that are downregulated during erythroid differentiation, consistent with the downregulation of their candidate target genes. Gfi1b on the other hand is known to function as a transcriptional repressor (Vassen et al., 2005) with important functions in both HSCs and during erythroid differentiation (Khandanpour et al., 2010; Saleque et al., 2002). Association of Gfi1b ChIP-Seq targets in progenitor cells with genes downregulated during later erythroid differentiation therefore suggests that a substantial number of genes repressed by Gfi1b later on in differentiation are already bound in progenitors. Interestingly, ChIP-Seq candidate targets from the ES cell dataset also showed statistically significant overlaps, with a downregulation at transitions 6 and 8 being associated with targets for C-myc, E2f1, N-myc and Zfx. Given the known role of these factors in cell proliferation, it is likely that these overlaps reflect the significant loss of proliferation during the maturation of myeloid progenitors rather than any overlap with ES cell specific function. A notion further supported by the fact that we did not observe any statistically significant overlaps with the classical pluripotency factors such as Nanog or Oct4. All together, this analysis demonstrates that gene sets derived from computationally inferred state transitions show statistically significant overlaps with candidate target gene lists from ChIP-Seq studies that are consistent with the known biology of the haematopoietic system.

Maximum parsimony reconstruction of non-haematopoietic developmental lineage trees

Having demonstrated that MP lineage tree reconstruction generates information that corresponds well with experimentally obtained knowledge for the well-defined haematopoietic system, we next addressed how it would perform with much less characterised systems. We first assessed lineage tree reconstruction using an expression dataset for neural development. Neuroepithelial stem cells (NSCs) are thought to be capable of differentiating into neurons, astrocytes and oligodendrocytes. NSCs differentiate into glial-restricted precursor (GRP) cells and neuron-restricted precursor (NRP) cells. NRP cells can give rise to multiple populations of neurons, whereas GRP cells give rise to astrocytes and oligodendrocytes (Dietrich et al., 2006). Based on the hierarchical clustering of the expression data set across three cell types - neurons, oligodendrocytes and astrocytes, Cahoy et al. (2008) inferred that mature astrocytes and oligodendrocytes do not share a large cohort of common “glial” genes further questioning the concept of ‘glial cell’ class. In contrast to their observation, the differentiation tree reconstructed from the same gene expression data (Fig. 2) strongly supports the concept of ‘glial

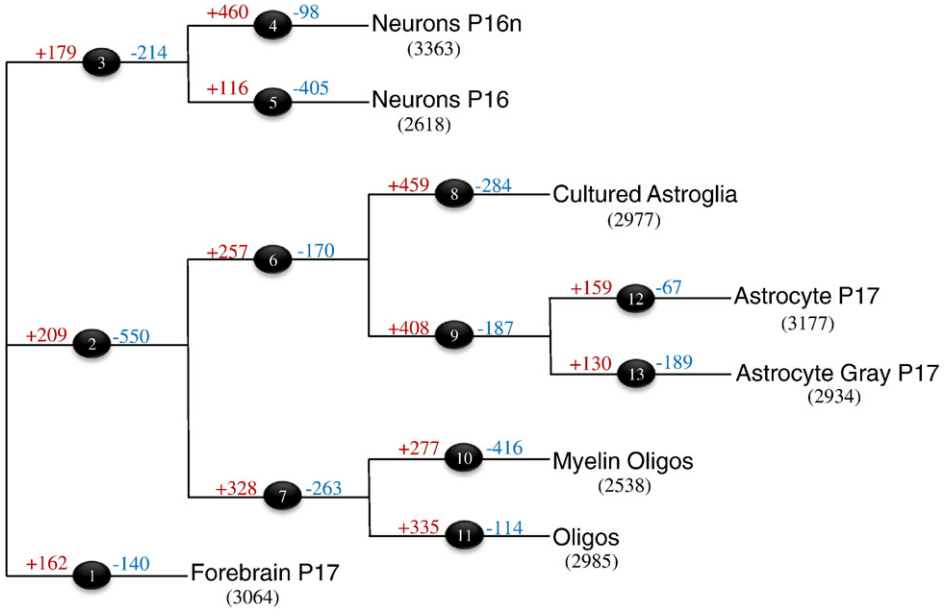

Fig. 2. Reconstruction of Neural differentiation from gene expression data using maximum parsimony. At each branch, numbers 1–13 represent predicted differentiation state transitions. At each state transition, the number of genes predicted to be activated (red) or repressed (blue) are indicated. The total number of differentially expressed active genes in each cell type shown in brackets. Supplementary Fig. 3 shows the same tree including the total number of genes predicted to be expressed at each node.

Table 3
List of candidate transcription factors important for a given state transition in the haematopoietic development tree. Shown are transcription factors with a significant overlap (P value<0.001) between its candidate target gene set and each state specific gene set. Detailed information on gene overlaps is provided in the Supplementary materials.

Branch	On / off	Transcription factors
1	Off	Suz12
3	Off	Gata1
7	Off	Erg, Fli1, Gfi1b, Meis1, Pu1, Runx1, Stat3
7	On	Gata1, Klf1
6	Off	C-myc, E2f1, N-myc, Zfx
8	Off	C-myc, E2f1, Erg, N-myc, Zfx

cell' class. Moreover, 10 out of 13 lineage-specific sets are overrepresented (P<0.05) for the 'neurological phenotype' from MGI with the highest enrichment for branch 2 with a p value of 1e−19 (Table 3).

Next we analysed a recently published expression dataset for embryonic development of definitive endoderm-derived organs which consists of expression profiles for immature definitive endoderm cells from embryonic day (E) 8.5 mouse embryos as well as E11.5 intestinal, pancreatic, liver, stomach, lung and oesophagus endodermal cells (Sherwood et al., 2009). During early separation into the different endodermal fates, the developing endoderm is thought to consist of expression domains arranged in an anterior to posterior sequence where adjacent subtypes show overlapping expression patterns (Sherwood et al., 2009). MP based analysis of the endodermal expression dataset resulted in a developmental tree that was largely arranged in an anterior to posterior sequence (Fig. 3a). We next compared the gene sets obtained from all 11 state transitions in the MP tree with candidate target genes obtained by ChIP-Seq analysis of adult liver tissue for two important transcriptional regulators of liver development and function (Cebpα and Hnf4) (Schmidt et al., 2010). Remarkably, this analysis showed much more highly significant overlaps between transcription factor targets and liver-specific genes than those of any other endodermal gene sets (Figs. 3b and c). In summary, these results demonstrate that Maximum Parsimony can be employed to reconstruct biologically informative cell differentiation trees using expression data from diverse developmental systems.

Discussion

Substantial research efforts have been invested into defining cell lineage trees, based on the premise that the developmental history of a cell critically influences its function within complex tissues. Meticulous observations of developing chicken embryos more than a century ago established the close developmental relationship

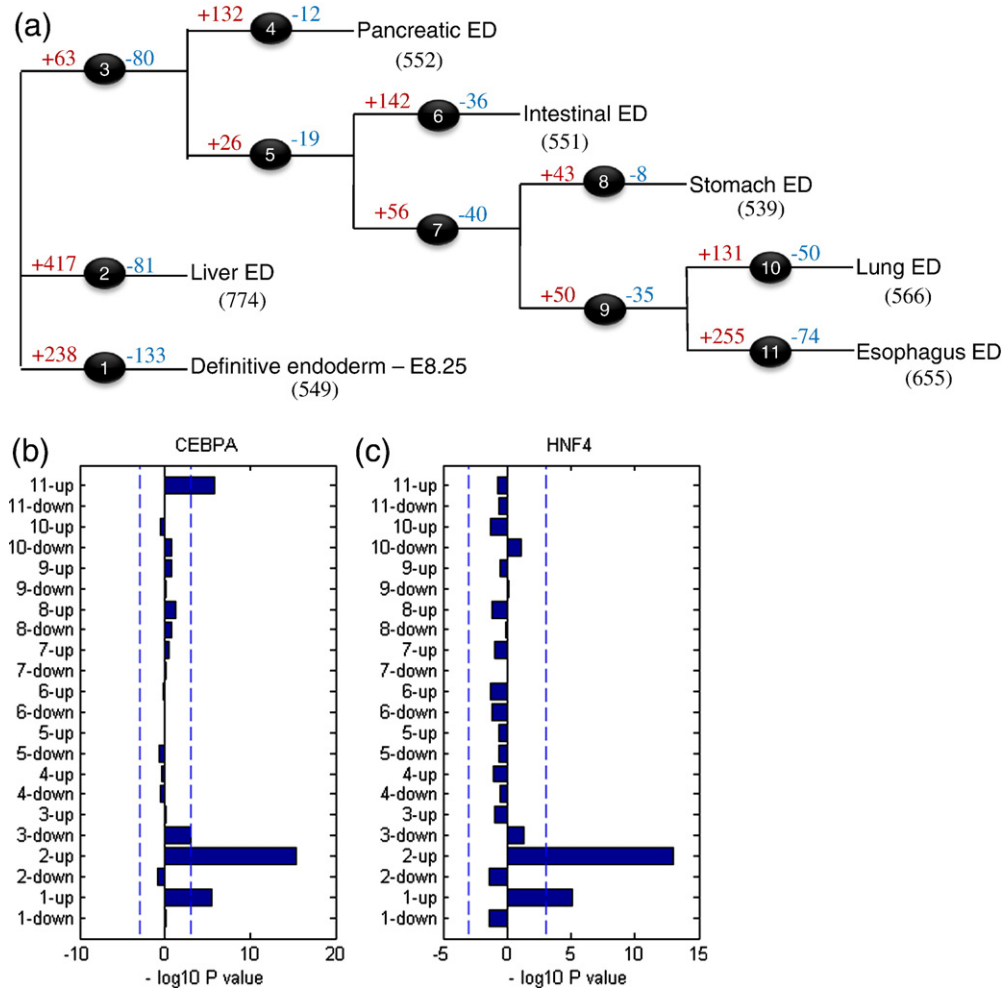

Fig. 3. a. A reconstructed developmental tree for early endoderm organogenesis using maximum parsimony. At each branch, numbers 1–11 represent predicted differentiation state transitions. At each state transition, the number of genes predicted to be activated (red) or repressed (blue) are indicated. ED—endoderm at E11.5. The total number of differentially expressed active genes in each cell type shown in brackets. b. and c. Overlap of gene sets at each developmental state transition with candidate target genes of Cebpα and Hnf4 transcription factors (Schmidt et al., 2010). p values were calculated using a hypergeometric test with Bonferroni correction (see Materials and methods for details).The dotted lines represent (p = 0.001). Supplementary Fig. 4 shows the same tree including the total number of genes predicted to be expressed at each node.

between blood and endothelial cells (His, 1900), which many years later was found to be reflected in transcriptional control mechanisms in both blood stem cells and endothelial cells (Chan et al., 2007; Silberstein et al., 2005). Similarly, the seminal studies by Sulston et al. not only provided the complete cell lineage tree for the nematode *C. elegans* (Sulston et al., 1983), but also made major contributions to other fields such as the study of programmed cell death (Ellis and Horvitz, 1986). Compared with the very detailed knowledge of nematode development, progress in delineating cell lineage tress for higher mammalian organisms has been severely hampered because of the obstacles in cell purity and frequency, tissue complexity, and *in utero* development. Consequently, only small segments of mammalian cell lineage differentiation trees have been defined thus far and the nature of most intermediate cell types remains completely obscure. In this paper we have shown that mammalian cell lineage trees can be approximated from gene expression profiles of mature cells; the type of information that is readily available for many mammalian tissues.

Due to the fragmented palaeontologic record, evolutionary biologists have long used comparisons of the anatomy, physiology and/or molecular sequence data of extant species to infer descent from common ancestors. More recently, comparative analysis of entire genomes was shown to allow reconstruction of phylogenetic trees based on inferring gene losses and gains and then using the principle of parsimony to reconstruct phylogenetic trees (Martens et al., 2008; Wapinski et al., 2007). Inspired by these recent studies, we explored whether comparative analysis of whole genome expression profiling data can be used in a similar fashion to reconstruct cell lineage trees. Analogous to the genome-wide reconstruction of phylogenetic tress that focuses on gene gains/losses but ignore smaller changes such as point mutations, we elected to reduce the complexity of gene expression datasets from continuous values to binary present/absent tables. This approach inevitably leads to the loss of potential information but makes the problem computationally tractable. Importantly, we were able to show that despite this "simplification", inferred gene gains/losses within our reconstructed cell lineage trees correlated well with the function of individual cell types when analysed based on gene ontology. Moreover, inferred expression changes displayed statistically significant correlations with candidate target gene lists obtained from recent ChIP-Seq studies. Within the reconstructed lineage differentiation tree for haematopoiesis for example, genes inferred as upregulated during erythroid development correlated with Gata1 candidate targets and genes downregulated with Gfi1b targets, which is consistent with the fact that both these genes are critical for erythroid differentiation but Gata1 commonly activates genes whereas Gfi1b is a known repressor. Other observations such as the significant overlap of Suz12 targets in embryonic stem cells with genes predicted to be turned off when blood stem cells differentiate provide intriguing clues that may link transcriptional control mechanisms in embryonic and adult stem cells.

We expect that parsimony-based analysis of expression datasets may find useful applications in addition to the reconstruction of developmental cell lineage trees. For example, tremendous research efforts are currently being invested into the development of protocols that allow the reprogramming of cellular phenotypes, mostly through ectopic expression of transcription factors to either mediate cell type specific differentiation from multipotent progenitors (forward programming) or the conversion of one type of mature cell into another (lateral programming) (Séguin et al., 2008; Zhou et al., 2008). For example, pushing pluripotent cells towards an early endodermal fate clearly represents an important first step in the development of clinically useful protocols for the production of pancreatic islet cells for type I diabetes therapy. Our analysis of expression data for different endodermal derivatives may help in identifying those genes that primarily specify the development of pancreatic endoderm and may instigate the development of new protocols for regenerative medicine applications.

Parsimony-based reconstruction of cell lineage trees from expression profiling datasets may also find future applications in cancer research. While descending from a primary lesion in a single cell, cancers have long been recognised as heterogeneous tissues consisting of multiple clones with overlapping but distinct patterns of secondary mutations (Nowell, 1976). Thus far, there have been no publications reporting comprehensive genome-scale analysis of expression differences resulting from this clonal heterogeneity. However, large-scale sequencing studies have already begun to reveal the spectrum of distinct mutations present within different metastases of the same primary tumour (Yachida et al., 2010). With the recent progress in generating expression profiles for single cells (Tang et al., 2010), it may soon be possible to generate expression data for several hundred cells of a given tumour. Subsequent maximum parsimony analysis could then be employed to identify expression changes that characterise the early events in tumorigenesis and may therefore provide relevant targets for the development of future therapies targeting the entire tumour rather than specific subclones.

Taken together, our study suggests that parsimony-based analysis of gene expression profiles has the ability to predict transcriptional states and developmental transitions within complex mammalian developmental systems. Importantly, intermediate cell types commonly not available for experimental analysis can be inferred and potentially novel regulators of cell fate decisions can be uncovered. Transcriptional regulation of cell fate choice is critical for normal organogenesis and tissue homeostasis, hence a better understanding of the underlying mechanisms also holds great promise for the development of effective cell-reprogramming protocols and new cancer therapies. An ability to predict transcriptional control mechanisms operating within experimentally inaccessible cell types therefore should benefit many diverse areas of biomedical research.

Supplementary materials related to this article can be found online at doi:10.1016/j.ydbio.2011.02.013.

Acknowledgments

We thank Dominic Schimdt for providing Cebpα and Hnf4 binding data, and Dean Griffiths, David Kent and Nicola Wilson for helpful comments. Work in the authors' laboratory is supported by the Medical Research Council, Leukaemia Lymphoma Research and the Leukemia and Lymphoma Society.

References

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C.T., Bryder, D., Yang, L., Borge, O., Thoren, L.A., Anderson, K., Sitnicka, E., Sasaki, Y., Sigvardsson, M., Jacobsen, S.E.W., 2005. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: a revised road map for adult blood lineage commitment. Cell 121, 295–306.

Bryder, D., Rossi, D.J., Weissman, I.L., 2006. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. Am. J. Pathol. 169, 338–346.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A., Thompson, W.J., Barres, B.A., 2008. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J. Neurosci. 28, 264–278.

Chambers, S.M., Boles, N.C., Lin, K.K., Tierney, M.P., Bowman, T.V., Bradfute, S.B., Chen, A.J., Merchant, A.A., Sirin, O., Weksberg, D.C., Merchant, M.G., Fisk, C.J., Shaw, C.A., Goodell, M.A., 2007. Hematopoietic fingerprints: an expression database of stem cells and their progeny. Cell Stem Cell 1, 578–591.

Chan, W.Y.I., Follows, G.A., Lacaud, G., Pimanda, J.E., Landry, J., Kinston, S., Knezevic, K., Piltz, S., Donaldson, I.J., Gambardella, L., Sablitzky, F., Green, A.R., Kouskoff, V., Göttgens, B., 2007. The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl−/− phenotype. Blood 109, 1908–1916.

Davidson, E., 2006. The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Academic Press, Amsterdam [Netherlands]; Boston [Ma.].

Dietrich, J., Han, R., Yang, Y., Mayer-Proschel, M., Noble, M., 2006. CNS progenitor cells and oligodendrocytes are targets of chemotherapeutic agents in vitro and in vivo. J. Biol. 5, 22.

Eilken, H.M., Nishikawa, S., Schroeder, T., 2009. Continuous single-cell imaging of blood generation from haemogenic endothelium. Nature 457, 896–900.
Ellis, H.M., Horvitz, H.R., 1986. Genetic control of programmed cell death in the nematode *C. elegans*. Cell 44, 817–829.
Enver, T., Pera, M., Peterson, C., Andrews, P.W., 2009. Stem cell states, fates, and the rules of attraction. Cell Stem Cell 4, 387–397.
Felsenstein, J., 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Meth. Enzymol. 266, 418–427.
Forsberg, E.C., Serwold, T., Kogan, S., Weissman, I.L., Passegué, E., 2006. New evidence supporting megakaryocyte–erythrocyte potential of Flk2/Flt3+ multipotent hematopoietic progenitors. Cell 126, 415–426.
His, W., 1900. Lecithoblast und angioblast der wirbelthiere. Histogenetische studien, Abhandlungen der Sächsischen Akademie der Wissenschaften zu Leipzig.
Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., Enver, T., 1997. Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. 11, 774–785.
Huang, D.W., Sherman, B.T., Lempicki, R.A., 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57.
Jaffredo, T., Nottingham, W., Liddiard, K., Bollerot, K., Pouget, C., de Bruijn, M., 2005. From hemangioblast to hematopoietic stem cell: an endothelial connection? Exp. Hematol. 33, 1029–1040.
Khandanpour, C., Sharif-Askari, E., Vassen, L., Gaudreau, M., Zhu, J., Paul, W.E., Okayama, T., Kosan, C., Möröy, T., 2010. Evidence that Growth factor independence 1b (Gfi1b) regulates dormancy and peripheral blood mobilization of hematopoietic stem cells. Blood 116, 5149–5161.
Kluger, Y., Tuck, D.P., Chang, J.T., Nakayama, Y., Poddar, R., Kohya, N., Lian, Z., Ben Nasr, A., Halaban, H.R., Krause, D.S., Zhang, X., Newburger, P.E., Weissman, S.M., 2004. Lineage specificity of gene expression patterns. Proc. Natl. Acad. Sci. USA 101, 6508–6513.
Lawson, K.A., Pedersen, R.A., 1992. Clonal analysis of cell fate during gastrulation and early neurulation in the mouse. Ciba Found. Symp., 165, pp. 3–21. discussion 21–26.
Martens, C., Vandepoele, K., Van de Peer, Y., 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. Proc. Natl. Acad. Sci. USA 105, 3427–3432.
Ng, S.Y., Yoshida, T., Zhang, J., Georgopoulos, K., 2009. Genome-wide lineage-specific transcriptional networks underscore ikaros-dependent lymphoid priming in hematopoietic stem cells. Immunity 30, 493–507.
Nowell, P.C., 1976. The clonal evolution of tumor cell populations. Science 194, 23–28.
Orkin, S.H., Zon, L.I., 2008. Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132, 631–644.
Ouyang, Z., Zhou, Q., Wong, W.H., 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc. Natl. Acad. Sci. 106, 21521–21526.
Pimanda, J.E., Göttgens, B., 2010. Gene regulatory networks governing haematopoietic stem cell development and identity. Int. J. Dev. Biol. 54, 1201–1211.
Rieger, M.A., Hoppe, P.S., Smejkal, B.M., Eitelhuber, A.C., Schroeder, T., 2009. Hematopoietic cytokines can instruct lineage choice. Science 325, 217–218.
Saleque, S., Cameron, S., Orkin, S.H., 2002. The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. Genes Dev. 16, 301–306.
Salipante, S.J., Kas, A., McMonagle, E., Horwitz, M.S., 2010. Phylogenetic analysis of developmental and postnatal mouse cell lineages. Evol. Dev. 12, 84–94.
Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., Odom, D.T., 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328, 1036–1040.
Séguin, C.A., Draper, J.S., Nagy, A., Rossant, J., 2008. Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells. Cell Stem Cell 3, 182–195.
Sherwood, R.I., Chen, T.A., Melton, D.A., 2009. Transcriptional dynamics of endodermal organ formation. Dev. Dyn. 238, 29–42.
Silberstein, L., Sánchez, M., Socolovsky, M., Liu, Y., Hoffman, G., Kinston, S., Piltz, S., Bowen, M., Gambardella, L., Green, A.R., Göttgens, B., 2005. Transgenic analysis of the stem cell leukemia +19 stem cell enhancer in adult and embryonic hematopoietic and endothelial cells. Stem Cells 23, 1378–1388.
Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Stadhouders, R., Palstra, R., Stevens, M., Kockx, C., van IJcken, W., Hou, J., Steinhoff, C., Rijkers, E., Lenhard, B., Grosveld, F., 2010. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev. 24, 277–289.
Somervaille, T.C., Matheny, C.J., Spencer, G.J., Iwasaki, M., Rinn, J.L., Witten, D.M., Chang, H.Y., Shurtleff, S.A., Downing, J.R., Cleary, M.L., 2009. Hierarchical maintenance of MLL myeloid leukemia stem cells employs a transcriptional program shared with embryonic rather than adult stem cells. Cell Stem Cell 4, 129–140.
Sulston, J.E., Schierenberg, E., White, J.G., Thomson, J.N., 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev. Biol. 100, 64–119.
Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., Surani, M.A., 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468–478.
Tzouanacou, E., Wegener, A., Wymeersch, F.J., Wilson, V., Nicolas, J., 2009. Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis. Dev. Cell 17, 365–376.
Vassen, L., Fiolka, K., Mahlmann, S., Möröy, T., 2005. Direct transcriptional repression of the genes encoding the zinc-finger proteins Gfi1b and Gfi1 by Gfi1b. Nucleic Acids Res. 33, 987–998.
Wapinski, I., Pfeffer, A., Friedman, N., Regev, A., 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449, 54–61.
Wasserstrom, A., Adar, R., Shefer, G., Frumkin, D., Itzkovitz, S., Stern, T., Shur, I., Zangi, L., Kaplan, S., Harmelin, A., Reisner, Y., Benayahu, D., Tzahor, E., Segal, E., Shapiro, E., 2008. Reconstruction of cell lineage trees in mice. PLoS ONE 3, e1939.
Wilson, D., Charoensawan, V., Kummerfeld, S.K., Teichmann, S.A., 2007. DBD—taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. 36, D88–D92.
Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E., Pimanda, J.E., de Bruijn, M.F.T.R., Göttgens, B., 2010. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 7, 532–544.
Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Iacobuzio-Donahue, C.A., 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467, 1114–1117.
Zhang, Z., Li, T., Ding, C., Ren, X., Zhang, X., 2009. Binary matrix factorization for analyzing gene expression data. Data Min. Knowl. Discov. 20, 28–52.
Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J., Melton, D.A., 2008. In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. Nature 455, 627–632.