Gene regulatory networks governing haematopoietic stem cell development and identity

JOHN E. PIMANDA*,1 and BERTHOLD GÖTtgENS*,2

1Lowy Cancer Research Centre and the Prince of Wales Clinical School, University of New South Wales, Sydney, Australia and 2Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK

ABSTRACT Development can be viewed as a dynamic progression through regulatory states which characterise the various cell types within a given differentiation cascade. To understand the progression of regulatory states that define the origin and subsequent development of haematopoietic stem cells, the first imperative is to understand the ontogeny of haematopoiesis. We are fortunate that the ontogeny of blood development is one of the best characterized mammalian developmental systems. However, the field is still in its infancy with regard to the reconstruction of gene regulatory networks and their interactions with cell signalling cascades that drive a mesodermal progenitor to adopt the identity of a haematopoietic stem cell and beyond. Nevertheless, a framework to dissect these networks and comprehend the logic of its circuitry does exist and although they may not as yet be available, a sense for the tools that will be required to achieve this aim is also emerging. In this review we cover the fundamentals of network architecture, methods used to reconstruct networks, current knowledge of haematopoietic and related transcriptional networks, current challenges and future outlook.

KEY WORDS: haematopoietic stem cell, gene transcription, regulatory network

Introduction

Transcription factor proteins respond to biological signals and regulate the transcription rates of genes. Gene regulatory networks (GRNs) describe the interactions between transcription factor proteins and the cis-regulatory modules (CRMs) in their target genes, often with a particular emphasis on regulatory genes including those that express spatially important signalling proteins (Alon, 2007a, Davidson, 2006). These intra- and intercellular networks control the temporal and spatial expression of proteins and govern the development and function of cells and the organism they constitute.

The gene regulatory state of a cell continually defines its identity throughout development. The prior state controls the current regulatory state of a cell which in turn determines its fate. Therefore, it is not surprising that the information processing power required to maintain a cell’s regulatory state is enormous. Approximate estimates of the number of regulatory functions operating in the human body at any given moment based on the number of genes that need to be regulated in a cell and the total number of cells in the body range in the order of 10^{16}-10^{20} (Istrail et al., 2007). These interactions however are not random and large GRNs contain a small set of recurring patterns called network motifs control basic logic operations and information flow (Alon, 2007b, Milo et al., 2002; Shen-Orr et al., 2002). These...
motifs are assembled to form larger subcircuits which determine which proteins are expressed, when, where and at what concentration to specify various body parts. The integrity of the developing organism is maintained by linking these subcircuits into a coherent GRN (Levine and Davidson, 2005; Levine and Tjian, 2003).

Development is a modular process that starts from a common template representing the undifferentiated state. This template is progressively refined as development proceeds to yield the various body parts with their distinct anatomical and functional properties (Davidson, 2006; Davidson and Erwin, 2006). GRNs that govern development are also inherently modular and hierarchical. The major structural domains and the transcriptional subcircuits that specify them are evolutionarily highly conserved. Indeed it has been proposed that conservation of the basic body plan in extant phyla in the animal kingdom is predicated by underlying conservation of GRN subcircuits or regulatory "kerne[l]s" (Davidson and Erwin, 2006). Following domain specification and patterning, functional properties of a body part are detailed by recruiting gene differentiation batteries that appropriately vary considerably between species.

The formation of blood, like the formation of other organ systems requires prior specification of the structural domain in the developing embryo from which it emerges. As detailed elsewhere in this supplement, these spatial domains are broadly conserved through evolution and the underlying regulatory kernels within large GRNs are themselves likely to be evolutionarily highly conserved. Therefore, although these networks may at first seem overwhelmingly complex, their analysis is facilitated by the modular architecture of both the organ systems and the regulatory circuitry as well as the use of a finite set of logic operations to control gene expression. In this review we cover the fundamentals of network architecture, methods used to reconstruct networks, current knowledge of haematopoietic and related transcriptional networks, current challenges and future outlook.

Fundamentals

Overview

At the cellular level, development represents a journey through successive regulatory states starting from a pluripotent stem cell and ending with fully differentiated mature cells (Waddington, 1957) (Fig. 1). Transcription factor concentrations in a cell are determined by events immediately preceding the current regulatory state and their activities determine the state that follows. The combination of transcription factors that are expressed in a cell will depend on the spatial context within the animal. Whether these factors are activated or not will in turn depend on regional cues supplied by signal transduction pathways. Transcription factors serve as vehicles to assemble activator and repressor protein complexes on DNA of target genes. If the regulatory proteins of a cell as a whole are considered an input, the output is determined by the DNA sequences of the CRMs and their accessibility to transcription factor complex binding. Therefore in terms of a hardwired regulatory circuit, the logic of the circuit ultimately resides within these physically disparate CRMs that can be strung out over tens or even hundreds of kilobases of DNA, up or downstream of a gene promoter or within an intron (Davidson, 2006).

Logic circuits are not unique to biology and the switches, gates and quantity controllers that control genetic information flow also underlie electronic circuits (Istrail et al., 2007). However, a fundamental difference between them is that although biological circuits are also hardwired in the sense that particular inputs result in predetermined outputs; biological circuits generate their own templates for successive regulatory events. Biological circuits can broadly be categorized as (i) developmental networks that underlie events during development that produce an animal from a fertilized egg and (ii) sensory networks that help maintain homeostasis within the fully formed animal.

Sensory transcription networks respond rapidly and make reversible decisions in response to environmental stimuli. Unlike signalling cascades, transcriptional networks can take up to one cell generation time to pass on a signal one step down a cascade (Alon, 2007b). As they operate on a short time-scale, long regulatory cascades are rare in signalling networks and most genes are regulated just one step away from their activator. Developmental transcription networks have different requirements and operate over longer time scales. Decisions often are largely irreversible and last long after the initial activator signal has disappeared as such developmental regulatory cascades are strung out over several cell generations (Alon, 2007b).

Cis-regulatory modules are information processing units and network anchors

A fundamental principle of developmental biology is that spatial patterning of gene expression is directly determined by the heritable cis-regulatory DNA sequence code (Davidson, 2006). When fragments of DNA representing cis-regulatory modules of genes are transferred into the genome, although they integrate randomly they still generate correct and predicted reporter ex-
pression that mirrors the developmental and spatial expression patterns of the endogenous genes they regulate (Delabesse et al., 2005; Gottgens et al., 2000; Gottgens et al., 2004; Sanchez et al., 1999; Silberstein et al., 2005). Genes are equipped with multiple, disparate CRMs, each several hundred base pairs long. As the transcription factor binding properties of each module varies, different modules are co-opted at specific stages of development to regulate the transcriptional output of a gene. Although, CRMs function as independent information processing units their outputs are communicated elsewhere. Protein complexes assembled on distant modules are brought into contact with each other and the basal transcription apparatus on the promoter by alternative looping of DNA to regulate transcriptional activity (Davidson, 2006). How accessibility at individual modules is coordinated at different stages of development and the precise determinants of this dynamic DNA conformation are not well understood (Fig. 2).

During specification of a body part, it may be required that a particular set or battery of genes needs to be expressed. Although the CRMs of these genes may indeed share target sites for critical transcription factors, no two CRMs of genes in a battery are exactly the same. The specific target sites in these modules vary in composition, spacing and order and respond to varying combinations of transcription factors that result in distinct gene outputs (Arnone and Davidson, 1997). These genomic regions anchor transcriptional networks by acting as information processors, which control basic logic operations and determine transcriptional output.

Information processing at CRMs can be reduced to suitable combinations of universal logic operations AND, OR and NOT (Istrail et al., 2007). If a CRM has sites for two transcription factors and its output is conditional on both sites being bound, then this module operates as an AND gate. Alternatively if the output is conditional on at least one of the inputs, then it acts as an OR gate. Although one input results in some activity many OR gates are additive in that the output is the sum of both inputs. The NOT gate acts as a switch; when the input is on, the output is off and vice versa.

Network Motifs

The wiring of transcriptional networks is not random. Analyses of GRNs in bacteria and yeast have shown that a small set of regulation patterns or network motifs are used repeatedly to regulate transcriptional output (Eichenberger et al., 2004; Lee et al., 2002; Milo et al., 2002). Where the CRMs act as information processing units that determine output of a particular gene, network motifs describe the patterns in which these individual outputs interact to regulate transcription over time.

The motifs found in biological networks are reviewed in detail elsewhere (Alon, 2007a, Alon, 2007b) and are summarized

**Fig. 2.** Distinct enhancers are used to regulate the expression of the same gene in different tissues. **(A)** The endoglin gene drawn to scale with the promoter (2) and -8kb (1), +7kb (3) and +9kb (4) enhancers marked in green and exons in brown. **(B)** Transgenic mice generated with Endoglin promoter constructs (Eng PlacZ) show little endothelial and no blood activity. Transgenic mice generated with the -8 enhancer and endoglin promoter (Eng -8/P lacZ) shows robust endothelial but no blood activity. Various combinations of the Eng P and Eng -8/P constructs together with the Eng +7 and Eng +9 enhancers target blood and endothelium.
here. The most basic interactions (simple regulation) are those between two genes, a transcription factor and its target gene. Following modification of a transcription factor by an activator, target gene concentrations could rise until it reaches a steady state level or feed-back and regulate itself either negatively (Lee et al., 2002; Rosenfeld et al., 2002) or positively (Kalir et al., 2005; Maeda and Sano, 2006). The kinetics of each of these responses with respect to the rate of expression of a target gene varies. Another commonly found network motif is the feed forward loop of which a number of types are described. This ‘three gene’ motif consists of a regulator X, that controls a second regulator Y, and gene Z that is controlled by both X and Y. The on-off kinetics of Z expression varies depending on the types of logic gates on the CRMs of both Y and Z. In a common coherent feed forward loop (Mangan and Alon, 2003), X positively regulates both Y and Z; an AND gate on Z results in slow activation and rapid inactivation of expression whereas an OR gate has the reverse response. In a common incoherent feed forward loop, X activates Z, but also represses Z by activating repressor Y (Mangan et al., 2003). This network motif utilizes both transcription factors that assemble repressor complexes on Z as well as microRNAs as repressor Y (O’Donnell et al., 2005). Double positive or double negative feedback loops are commonly found in developmental networks and describe the activation of both Y and Z by X following which Y and Z continue to positively or negatively regulate each other even after X is extinguished (Gardner et al., 2000; Johnston et al., 2005; Milo et al., 2004). This motif therefore ensures memory of X activity. It is possible to establish a series of differentiation steps by combining a series of these network motifs. Complex responses to signals can be mediated by single input modules or closely overlapping regulations which describe network motifs where either a single factor or groups of factors in varying combinations regulate sets of genes that must be activated or silenced for a particular functional response (Amit et al., 2007) (Fig. 3).

Experimental validation of these network motifs has mostly been performed in bacteria but data are being gathered in eukaryotic organisms. The ultimate challenge however, is to experimentally validate the complex networks that must operate in higher order metazoans. The methodology used to discover and experimentally validate one such motif, a variant of the fully connected triad that operates during blood stem cell development in the mouse embryo is detailed later in this review. In a fully connected triad, transcription factors X, Y and Z all positively regulate each other upon activation of one or more of its constituents a mechanism that maybe used more widely during development (Milo et al., 2002; Pimanda et al., 2007b).

**Cellular Memory**

As cells share the same genome and the gene regulatory state of a cell determines its identity and function, installing restricted chromatin states that play a role in expression or silencing of genes is essential for cells to adopt a more restricted phenotype. Gene expression patterns that are initially mandated by CRMs are retained long after the initial stimulus has been extinguished by locking in specific chromatin states in a cell. This is achieved by a variety of chromatin modifications including DNA methylation, histone acetylation and deacetylation, differential methylation of lysine residues of histones which install polycomb or trithorax group complexes and control accessibility of the transcription apparatus to promoters and distant CRMs (Allis et al., 2006; Davidson, 2006). Gene expression states can also be maintained by continuously functional sub-circuits that direct expression of downstream genes long after the initial trigger has ceased. Alternatively, transcriptionally driven memory subcircuits can be maintained by the community effect where all cells of a territory signal to one another and these inter-cellular feedback circuits maintain the regulatory state of cells within that territory (Davidson, 2006; Gurdon, 1988).

---

**Fig. 3. Different network motifs can be used to regulate the expression kinetics of developmentally important genes.** (A) (i) Simple negative regulation; the activation of X, up-regulates Y, which down regulates the activator (ii) Simple positive regulation; the activation of X, up-regulates Y, which further increases X. (B) Coherent feed-forward loops with different logic gates determine rates of target gene expression (i) shows an AND gate with delayed onset and rapid loss (ii) shows an OR gate with rapid onset and delayed loss of target gene expression. (C) Incoherent feed-forward loops, where the activator X up-regulates Z expression but also inhibits Z by up-regulating Y. These incoherent feed-forward loops act at different levels (i) transcription (ii) post-transcription (iii) post-translational. (D) Double positive and negative motifs. (i) Y and Z can continue to up-regulate each other after the activator X has been extinguished. (ii) The inhibition of Z can be initiated by the activation of X and maintained by Y.
Chromatin remodelling and activation of CRMs can be temporally linked to the expression of transcription factors that subsequently bind these modules (Bonifer and Hume, 2008). Occupancy of CRMs by pioneer transcription factors also helps maintain their ongoing accessibility and the adoption of inactive chromatin marks can follow the loss of expression of these factors (Tagoh et al., 2002; Xu et al., 2007). Chromatin condensation and inactivation of gene expression can also be achieved by tissue specific transcription factors recruiting histone bound co-repressors to gene promoters (Xu et al., 2007).

The Csfr1r surface protein is required for macrophage development and the csfr1r gene is silenced in the lymphoid lineage. However, whilst the gene is truly epigenetically silenced by inaccessibility of the transcription start site and promoter DNA methylation in T- cells, the promoter and intrinsic regulatory elements are maintained in a partially active chromatin conformation in B- cells. This has been shown to be due to the continuous expression in B- cells of Pu.1, an Ets transcription factor that binds to these regions of the Csfr1 locus. Csfr1r is not however expressed in B- cells owing to the active suppression of the promoter by Pax5, which is required to maintain B- cell identity (Tagoh et al., 2006). Therefore chromatin remodelling is a dynamic process and intermediate states can exist with a limited capacity for cells to transit between types.

**Robustness**

Establishing one developmental fate usually prohibits an alternate fate. This is important to maintain structural integrity in a given spatial domain. This is achieved by establishing multiple inter-cellular and intra-cellular regulatory feedback circuits that actively maintain a given state and repress alternate possibilities. A robust design would also safeguard against failure. Regulatory sub-circuits that specify various body parts have been selected in evolution to provide maximum insurance against the possibility of developmental failure (Cripps and Olson, 2002; Hinman et al., 2003; Pimanda et al., 2007a). Furthermore, the subcircuit elements of any developmental system may be of different ages and origins in evolution, selected and assembled piece-meal to best suit circumstance (Davidson and Erwin, 2006; Tokuoka et al., 2004). Robustness is also achieved by a degree of built in redundancy. CRMs, which constitute the nodes of a network often have multiple binding sites for transcription factors and utilize multiple related factors to drive gene expression.

**Network reconstruction**

Cell type specific transcription factor networks determine cellular response to external signals. Current strategies to reverse engineer such networks fall into two categories. The ‘bottom-up’ approach uses tissue specific CRMs of key regulators of a given tissue as anchor points to build block by block the regulatory interactions that operate in that tissue. The ‘top-down’ approach on the other hand attempts to reconstruct networks from genome-wide expression profiles of a particular tissue and predicted interactions between these expressed genes.

The ‘bottom-up’ approach

the most advanced haematopoietic stem cell (HSC) network identified in this way took the Scl/Tal1 gene as the starting point. Scl/Tal1 is required to specify blood development from mesodermal progenitors. Disruption of this gene during development results in a bloodless embryo and death. A 600 bp enhancer +19kb downstream of the transcription start site was found to be sufficient to target reporter gene expression to HSCs and endoderm in transgenic embryos (Gottgens et al., 2002; Rainis et al., 2005; Sanchez et al., 1999; Sanchez et al., 2001; Silberstein et al., 2005; Sinclair et al., 1999). In-vivo activity was shown to be dependent on two conserved Ets-binding sites and one conserved GATA- binding site which were bound by Fli1, Elf1, Erg and Gata2 in blood progenitors. The Scl +19 enhancer was therefore the first in vivo defined HSC cis-regulatory module of any gene required to specify HSCs. Using the Ets/Ets/Gata configuration of the Scl +19 enhancer as a template, a subsequent computational screen (Donaldson et al., 2005a) identified 67 clusters that were conserved between the human, mouse and dog genomes with the same spacing and orientation constraints as Scl +19. Three of these computationally identified clusters were situated in the Fli1, Hhex and Smad6 gene loci and showed HSC activity similar to the parent Scl +19 enhancer (Donaldson et al., 2005b, Pimanda et al., 2007a). Using bioinformatics, ChiP-seq and transgenic analyses, additional regulators of haematopoiesis have been progressively added to this nascent gene regulatory network (Chan et al., 2007; Chapman et al., 2003; Landry et al., 2008; Okuno et al., 2005; Wilson et al., 2009).

Therefore reverse engineering networks using a bottom-up approach involves a number of steps. The most important of which is the identification of tissue specific CRMs of genes required to specify that tissue. This is followed by the identification of transcription factor binding motifs within these modules that are important for their activity and the use of ChiP (Chromatin Immuno Precipitation) based techniques to identify transcription factors that are expressed in that tissue and are bound to these modules. From this point, one can either work up the developmental hierarchy by identifying tissue specific regulatory modules of transcription factors that are the drivers of the parent CRM or use bioinformatics tools to predict other target CRMs that respond in parallel with the parent CRM. When working up the developmental hierarchy, knowledge of the expression time course of various transcription factors during development of a particular tissue is extremely useful (Patterson et al., 2007; Patterson et al., 2005; Swiers et al., 2006). ChiP-chip analyses in appropriate cells can also be invaluable in identifying potential tissue specific CRMs in regulatory genes that can then be verified as truly targeting HSCs using transgenic assays.

The ‘top-down’ approach

this approach, which has been used successfully to determine regulatory networks in bacteria and yeasts, has recently been adopted to reverse engineer networks in B lymphocytes and macrophages (Basso et al., 2005; Gilchrist et al., 2006). Broadly, these approaches rely on genome-wide expression data sets in cells following various stimuli or experimental conditions as a starting point and use algorithms to infer gene-gene co-regulation. These methods aim to generate graphical representations of cellular networks where nodes represent genes and edges between them represent interactions, either between the encoded proteins or between encoded proteins and genes (Hartemink et al., 2001). Top down approaches have been variably successful in simple organisms and fall into the following broad categories:
optimization methods, which maximize a scoring function over alternative network models (Gat-Viks and Shamir, 2003; Gat-Viks et al., 2003); regression techniques, which fit the data to a priori models (Gardner et al., 2003); integrative bioinformatics approaches, which combine data from a number of independent experimental clues (Ideker et al., 2001); and statistical methods, which rely on a variety of measures of pair wise gene expression correlation (Butte and Kohane, 2000).

An early success in reconstructing mammalian regulatory networks using a ‘top-down’ approach was in human B- cells. Using a new method called ARACNe (algorithm for the reconstruction of accurate cellular networks), the authors show that a relatively small number of highly connected genes interact with most other genes in the cell, either directly or hierarchically, through highly connected sub-hubs (Margolin et al., 2006). The proto-oncogene MYC was identified as one of the largest hubs in the B-cell network (Basso et al., 2005). A top-down approach was also used to show that activating transcription factor 3 (ATF3) negatively regulates the Toll-like receptor transcriptional program in macrophages (Gilchrist et al., 2006). The authors analyzed waves of gene transcription following stimulation of macrophages with a Toll-like receptor 4 agonist and clustered genes based on their temporal patterns of expression. They hypothesized that a cluster of genes that included ATF3 (cluster 6) whose mRNA peaked at 1hr were co-regulated and share CRMs that were responsive to one or more members of this cluster. ATF binding sites were noticed to be over represented in these modules. Using Cytoscape, a network analysis and visualization tool to predict protein-protein interactions, ATF3 was predicted to act in concert with NF-kB and AP1. By scanning DNA sequences for ATF3/NF-kB and AP1 binding sites close to the transcription start sites, IL-6 and IL-12b were identified as target genes that are silenced following ATF3 binding.

There are approximately 200 transcription factor genes expressed in neural stem cells and possibly a similar number in HSCs (Akashi et al., 2003; Kummerfeld and Teichmann, 2006). To computationally predict and experimentally map every transcription factor-DNA interaction is a gargantuan task. Computational predictions are limited by the fact that DNA binding specificities of only a small fraction of transcription factors are sufficiently characterized to predict the sequences that they can and cannot bind although attempts are underway to redress this (Berger et al., 2006). The large data sets that inevitably will be generated can be filtered by focusing on phylogenetically conserved binding sites although phylogenetic footprinting will not identify sites that are functional but not conserved (Odom et al., 2007). The data sets can be further filtered by incorporating large scale experimental verification of chromatin accessibility and transcription factor binding events using ChIP-chip (Boyer et al., 2005) or ChIP-sequencing technologies (Marson et al., 2008). A major drawback however is the number of cells that are required to perform these genome-scale experiments, bearing in mind that the E11.5 mouse AGM (Aorta- Gonad- Mesonephros) is estimated to harbour approximately one HSC per embryo. Another limitation is the precise identification of tissue specific stem cells. The sensitivities of current flow cytometry protocols used to identify HSCs are approximately 50% (Weksberg et al., 2008) for mouse and 1% for human (Majeti et al., 2007).

**Regulatory networks in haematopoiesis**

To determine the regulatory state of a HSC, it would help to first know the identity of the cell or cells from which they arise and the site or sites from which they originate and not just reside. That the precursor cell is a derivative of the mesoderm and has a close association with the endothelial lineage is recognized and is detailed elsewhere in this supplement (Fig. 4). That the AGM is a site of HSC generation is not controversial and the recent demonstration that long-term repopulating HSCs in the AGM originate from VE-cadherin ‘CD45’ precursors is helpful in this regard (Taoudi et al., 2008). If the properties of HSCs are uniform, from a purely systems biology point of view, one could argue that wherever or from whatever precursor they arise, the chromatin state and transcription factors that are expressed (with allowance for factor redundancy) within these precursors are similar. However, molecular and phenotypic analysis has clearly shown intrinsic differences between HSC isolated from mouse embryos, juvenile, adult and aged mice (Dykstra et al., 2007; Kim et al., 2006; Sieburg et al., 2006; Yilmaz et al., 2006). Nevertheless, despite these differences, all these HSC populations share the hallmark properties of long-term self-renewal and multilineage differentiation capacity. Hence it is likely that many key aspects of

![Fig. 4. The ontogeny of blood development in a mouse model. Mesodermal cells from the posterior primitive streak (purple) of day 7 embryos establish extra-embryonic yolk sac endothelial plexus and blood. These cells circulate in the embryo after the establishment of cardiac contractions at day 8. Long-term repopulating blood stem cells originate from and around the dorsal aorta in the AGM region around day 10. These stem cells number is amplified in the fetal liver and placenta and populate the bone marrow around the time of birth. The bone marrow continues as the major hematopoietic organ in adult life.](image-url)
GRN architecture will be conserved among these various HSC populations.

We have reported that Gata2, Fli1 and Scl transcription factors, all of which are expressed in HSCs, cross-regulate each other in the AGM by binding CRMs that are active in the AGM and in LT-HSCs (Pimanda et al., 2007b) (Fig. 5). These transcription factors are required for normal haematopoiesis and are downstream of Bmp4, a morphogen that is required to specify the mesoderm and to specify haematopoiesis within the mesoderm (Maeno et al., 1996). This triad of transcription factors are components of a larger network in HSCs that we have constructed using a bottom-up approach as detailed above. The circuitry between Gata2, Fli1 and Scl and their CRMs conforms to a three gene network motif termed a fully connected triad (Pimanda et al., 2007b). This 3-way positive feed-forward loop, which shows a tendency to be locked into an ON state, is rare in the experimentally verified GRNs of bacteria and yeasts. An important feature of the fully connected triad is that following its activation, the members will perpetuate the expression of each other. Therefore the activating signal could be transient but sufficient to initiate a key transcriptional programme which will be retained by an ongoing dynamic interaction (see preceding section on ‘memory’). This has important implications for cells such as HSCs which emerge at one site (the AGM) and then travel to another (the fetal liver) for further amplification prior to taking up residency in the bone marrow.

The Gata2/Fli1/Scl triad is also intriguing as it provides a stepping stone to retrace transcriptional events that underlie HSC development (Fig. 6). The development of HSCs in the AGM is regulated by Notch-1 and is impaired in Notch-1 deficient mouse embryos (Kumano et al., 2003). Notch-1 and Gata2 are co-expressed in the haemogenic endothelium in the floor of the aorta at E10.5 and Notch1 binds the Gata2 promoter and acts as an upstream regulator of Gata2 expression (Robert-Moreno et al., 2005). Notch 1 itself is downstream of hedgehog and Vegf signalling (Gering and Patient, 2005; Lawson et al., 2002). However, because activation of the regulatory elements depends on “and” interactions between Gata2 and Fli1 (Scl likely plays a quantitative role once the Gata2/Fli1 complex is bound), activating Gata2 alone is unlikely to be sufficient to initiate the triad. Bmp4 signalling is required for the formation of the dorsal lateral plate mesoderm that gives rise to the dorsal aorta and the subsequent development of blood. Disruption of Bmp4 results in loss of Fli1 and Scl expression (Walmsley et al., 2002). The Gata2 promoter responds to Bmp4 which also initiates Fli1 expression (Oren et al., 2005). Activation of both Gata2 and Fli1 might therefore be sufficient to trigger the circuit that could then maintain expression of each other and Scl. As important as maintaining activity of the triad in HSCs is the abrogation of the triad to diminish concentrations of these factors as cells differentiate down.

Fig. 5. A fully connected triad of HSC transcription factors. Scl, Fli1 and Gata2 regulate each other by collectively binding the Scl+19, Fli1+12 and Gata2-3, haematopoietic stem cell enhancers. This triad is probably activated by Bmp4 signalling and the expression of these essential haematopoietic factors in HSCs can be maintained by each other in the absence of the activator.

Fig. 6. A bottom-up reconstruction of the haematopoietic stem cell regulatory network. The connections between each of these genes have been experimentally verified.
specific lineages. Interestingly, Gata1 has been shown to disrupt Gata2-positive auto-regulation by binding the Gata2-3 enhancer as cells differentiate down the erythroid-megakaryo-cytic lineages (Grass et al., 2003). Given the critical "and" interaction between Gata2 and FlI1, shutting down Gata2 may be sufficient to abrogate the triad and one would predict that maintenance of Scl expression in maturing erythroid cells would require an alternative erythroid enhancer. Interestingly, such an enhancer has indeed been identified and shown to depend on Gata1 and Scl as upstream inputs (Ogilvy et al., 2007).

Reconstruction of core transcriptional regulatory circuitry would not be complete without integrating regulatory RNAi. miRNAs appear to regulate expression of a significant proportion of all genes in a variety of mammalian cell types including haematopoietic cells (Garzon and Croce, 2008; Lewis et al., 2005; Stefani and Slack, 2008). The expression of miRNAs is not random. These genes themselves participate in cellular differentiation and are subject to tissue specific regulation (Dore et al., 2008). Incoherent feed-forward loops where X activates Z, but also represses Z by activating a repressor Y is a classic network motif used to generate pulse-like dynamics (Alon, 2007b, Basu et al., 2004). As detailed above, regulatory miRNA genes are ideally placed to act as 'Y' type repressors. Indeed by correlating genome-wide tri-methylation of the fourth Lysine residue of Histone H3 subunits (H3K4M3 chromatin marks) with known transcription start sites of primary and mature miRNA transcripts and ESTs (and not promoters of protein coding genes) a directory of transcription start sites of miRNA genes that are expressed in ES cells has been constructed and demonstrate overlapping occupancy of Oct4, Sox2, Nanog and Tcf3 on miRNA promoters (Marson et al., 2008). As promoter occupancy with Oct4, Sox2, Nanog and expression of protein coding genes in ES cells has already been mapped and because experimentally verified targets of certain miRNAs are known, it has been possible to wire miRNA modulation into the gene regulatory network in ES cells.

Regulatory networks and haematopoietic disorders

Diagnosis and classification of haematopoietic malignancies is largely based on morphology and immunophenotype. Although gene expression profiling of haematopoietic malignancies to risk stratify patients based on molecular signatures has been successful (Golub et al., 1999), it is unlikely to enter routine diagnostic and treatment paradigms unless less toxic and better treatment alternatives are available. If targeted therapy is the aim, differential gene expression profiles by themselves are not very informative. If the expression profiles can be organized into a hierarchical interactome that potentially guides the development of targeted therapy, the benefits of expression profiling would be more tangible. Comprehensive analysis of tissue regulatory networks provides a framework to understand perturbations that lead to various disease phenotypes. It also provides a sense of the most important regulatory genes. For example, MYC is one of the most connected hubs in the B- cell interactome (Li et al., 2003) and is the key oncogenic lesion in Burkitt’s lymphoma and a potential target for cancer therapy (Soucek et al., 2008). Clearly, if a recurrent cytogenetic lesion can be identified in a tumour either by conventional karyotyping or by comparative genomic hybridization arrays and the involved gene is either highly connected or is a critical mediator of an important hub within the normal interactome, this would be a target for therapy. Expression profiling has also been used to characterize the molecular signature arising from specific pharmacological interventions (Lamb et al., 2006). Again, it would be far more informative to characterize the networks that are perturbed in response to specific pharmacological interventions for it could guide therapy. By expression profiling a tumour and identifying the key networks that are deranged, specific interventions could be planned based on prior knowledge.

Network analysis of haematopoietic malignancies however will only be as good as the original template. The most informative blood cell template to date, the B- cell interactome, was constructed by integrating gene expression profiles from several stages of B- cell development with known protein-protein and protein-DNA interactions and known modulators of these interactions compiled from various databases (Mani et al., 2008). One drawback of this methodology is the large background population that is necessary for comparison. As dependency metrics like mutual information require a certain sample size to establish significance, this may pose a difficulty in situations where sample sizes are limited (Mani et al., 2008).

Challenges and future directions

By their very nature, HSCs are suspended in transit towards another more differentiated state. HSC purity and adequate cell numbers for in vitro analyses remain major issues for network reconstruction. In vitro in vivo culture and expansion of HSCs without altering their phenotype would be necessary unless fundamental technical advances in methodology permit gene expression and ChIP on very small numbers of cells. One possible substitute is a robust ES cell differentiation system in which LT-HSCs can be effectively identified.

To reconstruct networks, effective methods to identify protein-DNA interactions are required. The reliance on specific antibodies for ChIP is a hurdle. Methods to express tagged transcription factors at appropriate concentrations in ES cells to reconstruct the ES cell interactome (Kim et al., 2008; Wang et al., 2006) have been successful and could be adapted to HSCs derived from ES cells if the necessary protocols are developed. A limitation that is likely to be overcome in the near future is the lack of a comprehensive database of binding specificities of all the 2,500 or so known transcription factors. Protein binding microarray technology is rapidly increasing the number of known binding specificities (Berger et al., 2008; Berger et al., 2006). High throughput sequencing is also likely to have a major impact on our ability to document genome-wide transcription factor binding events using ChIP-seq technology. Transgenic evaluation of CRMs is currently regarded as necessary to determine potential activity in LT-HSCs. This is expensive, labour intensive and time consuming. The evaluation of CRMs in ES cell differentiation assays could be an acceptable alternative. As a clearer map of cell signalling cascades that drive developmental processes evolves, a future challenge will be to integrate haematopoietic transcriptional networks with
these cascades that activate the regulatory programme.

Acknowledgements

Work in the authors’ laboratories is supported by grants from the National Health and Medical Research Council of Australia; the Leukaemia Foundation, Australia; the Australian Research Council; the Viertel Foundation; the Ramaciotti Foundation, the Leukaemia Research Fund, UK; Cancer Research UK; the Leukemia and Lymphoma Society, the Newton Trust and the Kay Kendall Leukaemia Fund.

References


JOHNSTON, R.J., JR., CHANG, S., ETCHBERGER, J.F., ORTIZ, C.O. and


MAN migrates away as well as blood formation in zebrafish embryos. *Blood* 105: 3502-3511.


endothelium together with embryonic and adult haematopoietic progenitors.

*Development* 126: 3891-4827.


Further Related Reading, published previously in the *Int. J. Dev. Biol.*


**Dynamic expression of Endoglin, a TGF-beta co-receptor, during pre-circulation vascular development in chick**
Cantas Alev, Brendan A.S. McIntyre, Kanako Ota and Guojun Sheng

**Identification of hoxb1b downstream genes: hoxb1b as a regulatory factor controlling transcriptional networks and cell movement during zebrafish gastrulation**
Willem M.R. van den Akker, Antony J. Durston and Herman P. Spaink
*Int. J. Dev. Biol.* (2010) 54: 55-62

**Transcriptional regulation by Pax3 and TGFbeta2 signaling: a potential gene regulatory network in neural crest development**
Hiromichi Nakazaki, Yueh-Wei Shen, Beth Yun, Anvesh Reddy, Varun Khanna, Barbara Mania-Farnell, Shunsuke Ichi, David G. Mclone, Tadanori Tomita and C. Shekhar K. Mayanil
*Int. J. Dev. Biol.* (2009) 53: 69-79

**Epiblast-derived stem cells in embryonic and adult tissues**
Maria P. De-Miguel, Francisco Arnalich-Montiel, Pilar Lopez-Iglesias, Alejandro Blazquez-Martinez and Manuel Nistal
*Int. J. Dev. Biol.* (2009) 53: 1529-1540

**A history of Evo-Devo research in Spain**
Jaume Baguñà
*Int. J. Dev. Biol.* (2009) 53: 1205-1217

**A novel role of the glial fate determinant glial cells missing in hematopoiesis**
Cécile Jacques, Laurent Soustelle, István Nagy, Céline Diebold and Angela Giangrande
*Int. J. Dev. Biol.* (2009) 53: 1013-1022

**Reptile scale paradigm: Evo-Devo, pattern formation and regeneration**
*Int. J. Dev. Biol.* (2009) 53: 813-826

**The Hox Complex - an interview with Denis Duboule**
Michael K. Richardson
*Int. J. Dev. Biol.* (2009) 53: 717-723

**Genomic control of patterning**
Isabelle S. Peter and Eric H. Davidson
*Int. J. Dev. Biol.* (2009) 53: 707-716

5 yr ISI Impact Factor (2009) = 3.253