Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development

Ian J. Donaldson\(^1\), Michael Chapman\(^1\), Sarah Kinston\(^1\), Josette Renée Landry\(^1\), Kathy Knezevic\(^1\), Sandie Piltz\(^1\), Noel Buckley\(^2\), Anthony R. Green\(^1\) and Berthold Göttgens\(^1,\)*

\(^1\)Department of Haematology, Cambridge Institute for Medical Research, Cambridge University, Hills Road, Cambridge CB2 2XY, UK and \(^2\)Schools of Biochemistry and Microbiology and Biomedical Sciences, University of Leeds, Leeds LS2 9JT, UK

Received October 27, 2004; Revised December 16, 2004; Accepted January 5, 2005

The development of blood has long served as a model for mammalian cell type specification and differentiation, and yet the underlying transcriptional networks remain ill defined. Characterization of such networks will require genome-wide identification of cis-regulatory sequences and an understanding of how regulatory information is encoded in the primary DNA sequence. Despite progress in lower organisms, genome-wide computational identification of mammalian cis-regulatory sequences has been hindered by increased genomic complexity and cumbersome transgenic assays. Starting with a well-characterized blood stem cell enhancer from the SCL gene, we have developed computational tools for the identification of functionally related gene regulatory sequences. Two candidate enhancers discovered in this way were located in intron 1 of the Fli-1 and PRH.Hex genes, both transcription factors previously implicated in controlling blood and endothelial development. Subsequent transgenic and biochemical analysis demonstrated that the two computationally identified enhancers are functionally related to the SCL stem cell enhancer. The approach developed here may therefore be useful for identifying additional enhancers involved in the control of early blood and endothelial development, and may be adapted to decipher transcriptional regulatory codes controlling a broad range of mammalian developmental programmes.

INTRODUCTION

A major challenge of the post-genome era will be the integration of genomic sequences with transcription factor function and expression to decipher the gene regulatory networks which lie at the heart of development (1,2). Progress towards this goal will require identification of gene regulatory regions within the genomic landscape and an understanding of the way in which specific patterns of expression are encoded within the primary DNA sequence. In silico approaches, which identify gene regulatory regions with specific biological activity, are proving increasingly successful in ‘lower’ organisms such as yeast or flies (3–7). However, the higher biological complexity of mammals is associated with increasingly intricate mechanisms of gene regulation (2), and genome-wide identification of mammalian regulatory regions with predicted biological activity demonstrated through in vivo transgenic analysis has not been described.

Haematopoiesis has long served as a paradigm for adult stem cell systems and markers are available for mature progeny of all haematopoietic lineages as well as for a plethora of progenitors with varying potentials. Studies over the last 20 years have established that transcriptional control is central to the specification and subsequent differentiation of haematopoietic stem cells (HSCs) (8,9). With many of the key transcription factors known, haematopoiesis provides a powerful cellular system for the analysis of mammalian gene regulatory networks. The key missing ingredient, particularly for the stem and progenitor cell stages, is a set of experimentally validated gene regulatory regions together with a molecular understanding of their biological activity.

Here, we describe new computational tools for the identification of mammalian enhancers based on evolutionarily conserved patterns of transcription factor binding sites. Starting with a well-characterized blood stem cell enhancer, we demonstrate that this approach permits identification of...
enhancers with predicted biological activity in vitro and in vivo. Moreover, molecular and transgenic analysis of these new enhancers allowed us to develop a nascent transcription factor network controlling formation of early blood and endothelium. The approach presented here is therefore likely to allow the characterization of transcriptional regulatory codes controlling a broad range of mammalian developmental programmes.

RESULTS

To identify transcriptional control regions active during early embryonic haematopoietic development, we have developed a strategy that is based on a haematopoietic stem cell enhancer (+19 enhancer) identified downstream of the SCL transcription factor, a key regulator of haematopoietic stem cell formation (10). Transgenic reporter studies have demonstrated that the SCL +19 element (numbering of enhancers refers to the distance in kilo base pair from the transcriptional start site) targets expression to haemangioblasts, HSCs and endothelium (11–13). Subsequent molecular analysis identified two conserved Ets sites and one conserved GATA site (Fig. 1A), all of which were critical for enhancer activity (13). We now developed a new suite of bioinformatic analysis tools to generate human genome-wide catalogues of potential transcription factor binding sites. The human genome contains >47 million potential Ets binding sites (GGAA/T) and >18 million GATA binding sites (GATA). To identify the subset of sites likely to be functional, we excluded motifs located in repetitive DNA and included only those for which the core motif was conserved in whole genome alignments of human, mouse and rat, and was located within an 8 bp block of sequence identity. The latter criterion was based on our previous comparative analysis of multiple mammalian SCL loci, which demonstrated that functionally significant binding sites are typically situated in blocks of extended sequence identity (14). These approaches reduced the number of sites to ~625 000 Ets and 189 000 GATA sites (Fig. 1B).

To further enrich for putative regulatory elements with sequence properties similar to those of the SCL +19 enhancer, we developed a new program (TFBScluster) and used it to determine the locations of binding site clusters reflecting the combination of sites found in the SCL +19 core enhancer (at least two conserved Ets sites and one conserved GATA site within a 50 bp sequence window). Clusters found in coding exons or >100 kb away from the first or last exons of the nearest swissprot (15) gene were excluded from further analysis. This resulted in the identification of 1555 clusters (Fig. 1B), 1053 of which were located in regions of the genome with high regulatory potential scores (16). Regulatory potential scores are based on short alignment patterns in regulatory regions and neutral DNA with a threshold for regulatory potential obtained from experimentally verified erythroid gene regulatory regions. As with all other filtering steps, regulatory potential scores were used to increase the likelihood that binding sites were located in gene regulatory regions, but their use may also remove some ‘true’ functional binding sites from further analysis (false negatives). The computational tools developed combine the rationales behind two types of existing programs, namely, identification of conserved transcription factor binding sites [i.e. rVista (17)] and genome-wide identification of transcription factor binding site clusters [i.e. (5)].

Enhancer activity is frequently dependent on the precise spacing and orientation of its constituent binding sites, and an understanding of these constraints can help in silico identification of bona fide transcriptional control regions (7). The SCL +19 enhancer contains two Ets sites and one GATA site, all in the forward orientation and separated by 1.5–2.5 turns (17–24 bp) of the double helix in all four mammalian species studied (Fig. 1A). Previous studies have demonstrated that this spatial arrangement of the constituent binding sites is critical for enhancer activity (13). We, therefore, extracted only those clusters with a spatial organization similar to the SCL +19 enhancer (two Ets followed by one GATA site, all in the forward orientation and each separated by one to three turns of the double helix). This resulted in the identification of 67 clusters, which were localized to 64 genes (Fig. 1B; for localization procedure see Materials and
These 64 genes included the blood stem cell marker CD34 as well as several other membrane receptors expressed during early blood and endothelial development. Interestingly, the gene list contained 18 transcription factors suggesting that the earlier mentioned strategy may be able to identify those cis-regulatory sequences that constitute the nodes of transcription factor networks. Two of the 18 transcription factors encoding the Fli-1 and PRH (also known as Hex) genes had been linked with blood formation, although their position within a regulatory hierarchy remained ill defined (13,18–20). Comparative genomic sequence analysis demonstrated that both the Fli-1 and PRH loci contain multiple non-coding peaks of elevated sequence identity (non-coding homology peaks) (Fig. 2). Indeed, the region of the Fli-1 locus shown in Figure 2 contains more than 60 peaks of homology, all of which are conserved to the same degree as Fli-1 coding exons. However, only the peak situated 12 kb downstream of the start of exon 1 (Fli-1 +12 element) contained an Ets/Ets/GATA cluster similar to the SCL +19 enhancer. Similarly within the PRH locus, only the peak situated in intron 1 (PRH +1 element) contained an analogous Ets/Ets/GATA cluster. The sequence of both the Fli-1 +12 and PRH +1 regions was highly conserved in human/mouse/rat alignments, and revealed the presence of several conserved Ets and GATA sites (Fig. 2, blue shading) in addition to the Ets/Ets/GATA clusters (Fig. 2, yellow shading).

To assess enhancer activity of the Fli-1 +12 and PRH +1 homology peaks, 500 and 600 bp fragments were subcloned into luciferase reporter constructs, the activity of which was analysed following stable transfection into haematopoietic cell lines. A 604 bp fragment of the SCL +19 enhancer increased mean luciferase activity 21-fold in a multipotent progenitor cell line (416B), but was inactive in the T-cell line BW5147. Similarly, mean luciferase activity in 416B cells was increased 64-fold and 27-fold by the Fli-1 +12 and PRH +1 elements, whereas both were inactive in BW5147 T-cells (Fig. 3A). Importantly, deletion of the Ets/Ets/GATA cluster from the enhancer fragment abrogated enhancer activity in 416B cells. Taken together, these data demonstrated that genome-wide computational analysis had identified functional enhancers with predictable biological activity in vivo.

To compare the biological activity of the Fli-1 +12, PRH +1 and SCL +19 enhancers in vivo, we generated transgenic mice carrying a lacZ reporter transgene driven by the SV40 minimal promoter and the respective enhancers. As previously described, the SV40 promoter alone does not give rise to any consistent pattern of β-galactosidase expression (21), whereas the SCL +19 enhancer directed expression to a subset of haematopoietic cells in E11.5 fetal liver as well as to endothelium and endocardium (Fig. 3B) (13). Two out of three transgenic lines carrying the Fli-1 enhancer construct displayed strong expression of the lacZ reporter during embryonic development (Fig. 3B). Wholemount analysis of E11.5 embryos revealed staining in the vasculature, the heart and also the fetal liver, the latter being the major haematopoietic organ at this stage of development. Histological sections demonstrated β-galactosidase expression in a minority of haematopoietic cells in fetal liver, in the endocardium and in the endothelium (Fig. 3B and data not shown). The PRH +1 enhancer gave rise to β-galactosidase expression, detectable by wholemount analysis, in the fetal liver and heart at E11.5 in four of five embryos showing β-galactosidase expression (Fig. 3B). In addition to the liver and heart staining, variable ectopic expression in the brain and spinal cord was also observed in two of these embryos indicating that the PRH fragment may be susceptible to position effects. Histological sections demonstrated β-galactosidase expression in a subset of fetal liver haematopoietic cells, in the endocardium and in rare endothelial cells in all four embryos with fetal liver and heart staining (Fig. 3B and data not shown).

The expression domains of both Fli-1 and PRH include developing endothelium and a subset of blood cells, including haematopoietic progenitors (22,23). Moreover, the PRH +1 enhancer identified in this study lies within a larger 3000 bp fragment shown previously to direct expression to yolk sac blood islands in E7.5 mouse embryos (23). The Fli-1 +12 and PRH +1 enhancers, therefore, target expression to a sub-domain of the normal expression pattern of their corresponding genes. Of particular note, our results demonstrate that both enhancers give rise to a pattern of tissue-specific expression very similar to that generated by the SCL +19 enhancer. The genome-wide strategy described here has therefore resulted in the identification of enhancers with overlapping activity in vivo. With the exception of the one GATA and two Ets sites, the sequences of the SCL +19, Fli-1 +12 and PRH +1 enhancers showed little overall similarity (http://hscl.cimr.cam.ac.uk/supplementary_don04.html) emphasizing the necessity to use motif-based tools rather than sequence homology searches for the computational identification of gene regulatory regions with predicted biological activity.

To integrate the new Fli-1 and PRH enhancers into the regulatory network controlling blood formation, it was important to identify transcription factors that bind the enhancers in vivo. Previous studies have shown that the SCL +19 enhancer is bound by GATA-2 and by the Ets factors Fli-1 and Elf-1 (13). Chromatin immunoprecipitation experiments were therefore performed using 416B cells to identify transcription factors binding to the Fli-1 +12 and PRH +1 enhancers in vivo. The Fli-1 +12 enhancer was enriched in immunoprecipitates generated by Elf-1, Fli-1 and GATA-2 antibodies, demonstrating that this enhancer was occupied by all three factors. The PRH +1 enhancer was enriched in immunoprecipitates generated using Fli-1 and GATA-2 antibodies with a more modest enrichment obtained with an Elf-1 antibody (Fig. 4A). These results therefore place the two new enhancers into the same regulatory network as the SCL +19 enhancer, and have allowed us to assemble a nascent transcriptional network for early embryonic blood and endothelial development (Fig. 4B).

DISCUSSION

The data presented in this study demonstrate the utility of a genome-wide computational approach for the identification of cis-regulatory elements with defined biological activity.
This is a task of considerable magnitude given the size of the human genome and the complexity of transcriptional regulation in mammals (2). Including introns and flanking sequences 50 kb either side of the gene, the Fli-1 locus alone contains more than 90 non-coding homology peaks of human/mouse/rat, all of which represent candidate regulatory elements based on their level of sequence conservation (as high or higher than Fli-1 coding exons) and at least some of these are likely to play a role in the overall regulation of Fli-1 expression. Nevertheless, the current study allowed
us to rapidly home in on individual homology peaks, which, through subsequent molecular and transgenic analysis, were integrated into an endothelial/blood stem-progenitor transcriptional network, with nine experimentally verified direct interactions including, for example a previously unrecognized positive feedback loop involving Fli-1. Two candidate Enhancers were chosen for functional validation in the current study, both associated with transcription factor genes previously implicated in controlling early blood and endothelial development. In view of this somewhat biased nature of choosing candidate elements for functional validation, further transgenic analyses of additional candidate enhancers will be required to assess the full potential of the current search parameters in identifying new enhancers active in developing blood and endothelial cells. Pioneering studies in Drosophila suggest that further refinement of our computational approach will be possible (5–7). Accordingly, detailed structural and functional comparisons of the Fli-1 +12, PRH +1 and SCL +19 Enhancers are likely to identify additional common features, which could be incorporated into future algorithms. Moreover, identification of false positive binding site clusters (i.e. those lacking Enhancer activity) will permit the inclusion of stringent negative criteria (24). Interestingly, deletion of the SCL +19 Enhancer from the endogenous SCL locus in ES-cells appears to have little or no effect on the levels of SCL expression or the ability of ES-cells to differentiate into blood either in vitro or in vivo (25). This observation serves to underline the likely high degree of redundancy of mammalian regulatory elements, a point well illustrated by extensive deletion analyses of α- and β-globin regulatory elements (26–29). Dissection of transcription factor networks controlling development is currently most advanced in flies, worms and sea urchins. However, none of these three model organisms lends themselves to deleting specific Enhancer sequences from the germline in order to study loss of function or possible redundancy of gene regulatory elements. Enhancer redundancy may therefore not be

Figure 3. The Fli-1 +12 and PRH +1 elements function as haematopoietic/endothelial Enhancers. (A) The Fli-1 +12, PRH +1 and SCL +19 elements enhance the expression of a luciferase reporter gene linked to the SV40 minimal promoter in a multipotent progenitor cell line (416B), but not a T-cell line (BW5147). Histograms represent mean luciferase values and standard deviations of eight independent pools of stably transfected cells for each construct in each cell line. Deletion of the Ets/Ets/GATA cluster sequences abrogates Enhancer activity in 416B cells (see SV/luc/Fli-1 +12M and SV/luc/PRH +1M). (B) Analysis of F0 mouse embryos transgenic for SV/luc/Z reporter constructs containing the indicated Enhancers. The figure shows representative E11.5 X-gal stained embryos as well as histological sections of the fetal liver and heart demonstrating expression of all three constructs in haematopoietic cells and endocardium.

Figure 4. Integration of the Fli-1 +12 and PRH +1 Enhancers into the transcription factor network controlling early blood and endothelial development. (A) The Fli-1 +12 and PRH +1 Enhancers are bound by Ets and GATA factors in haematopoietic progenitor cells. Chromatin immunoprecipitation was performed in 416B haematopoietic progenitor cells using antibodies against Elf-1, Fli-1 and GATA-2. Semi-quantitative PCR followed by Southern blotting showed enrichment for the Fli-1 +12 Enhancer in Elf-1, Fli-1 and GATA-2 immunoprecipitates, and enrichment for the PRH +1 Enhancer in Fli-1, GATA-2 and (to a lesser extent) Elf-1 immunoprecipitates. Immunoprecipitates obtained using rabbit IgG served as negative controls. PCR amplification of two control regions (exons 6 of LMO2 and SCL) showed no significant enrichment with Elf-1, Fli-1 and GATA-2 immunoprecipitates when compared with the IgG control sample (data not shown). (B) A nascent transcriptional network controlling early blood and endothelial development. Red lines indicate new links identified in the current study.
confined to vertebrates, and should be considered as a potential caveat whenever a given regulatory element is designated a network node connecting two transcription factors. Of note, redundancy of regulatory elements, and their subsequent partial specialization, likely represent important mechanisms by which organisms can evolve additional levels of biological complexity.

Two other recently described strategies promise to provide complementary approaches to genome-wide identification of candidate cis-regulatory elements. Firstly, chromatin immunoprecipitation combined with genomic arrays should allow a description of transcription factor binding patterns within a given cell type (30,31). Secondly, genome-wide sampling of regions of open chromatin (32,33) might provide a way of concentrating motif searches on the portion of the genome involved in regulating gene expression. However, both of these approaches require purification of significant numbers of cells, a particular problem when studying stem cell systems or early developmental programmes. Deciphering regulatory information encoded in the primary DNA sequence circumvents the need for pure cell populations and has the potential to provide information on cis-regulatory elements operating in all mammalian cell types regardless of the amount of biological material available. The approach used in the current study therefore represents a widely applicable strategy to characterize mammalian gene regulatory networks.

MATERIALS AND METHODS

Identification of conserved binding site clusters

Genome-wide positions of consensus binding sites were identified in individual and in aligned genomes available from Genome Bioinformatics at the University of California (Santa Cruz) using a program developed by our group called TFBSSearch, which examines a single sequence or global alignment for conserved non-overlapping binding sites (14). Consensus sequences for the factors used in this study were obtained from the literature (for relevant references, see http://hscl.cimr.cam.ac.uk/supplementary_don04.html). Obtaining transcription factor consensus binding sites is greatly facilitated by the TRANSFAC (http://www.gene-regulation.com/pub/databases.html#transfac) and JASPAR databases (http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl). To help identify binding sites located in blocks of extended sequence conservation, the IUPAC code ‘N’ was added to the start and end of the consensus sequence, where ‘N’ could represent any nucleotide as long as it was conserved across the alignment. The tools developed for this manuscript can be adapted to use position weight matrices (PWMs) as well as IUPAC consensus sequences. However, the published PWMs for GATA and ETS factors are largely based on in vitro experiments and introduce flexibility into the core binding sites. As this added degeneracy has not been validated by in vivo studies, only IUPAC consensus sequences were used for the dataset presented in this manuscript.

A new program (TFBScluster, http://hscl.cimr.cam.ac.uk/supplementary_don04.html) was developed to identify binding site clusters containing a defined number of conserved sites within a user-defined sequence window. Any overlapping clusters are merged together, which can result in clusters larger than the user defined window size. We have developed a suite of integrated Perl scripts to determine whether a cluster is localized to annotate genes, and to aid subsequent analysis (when ‘localizing’ clusters to genes, only clusters situated in introns and up to 100 kb 5’ and 3’ flanking sequences are considered. Therefore, some clusters will not be ‘localized’ to any gene, and some genes can have several clusters). To identify putative functions of localized genes, external references (via db_xref from Ensembl download files) for each transcript are parsed to find SWISS-PROT accessions (http://www.ebi.ac.uk/swissprot/).

Functional analysis of candidate enhancers

Candidate enhancer sequences were amplified by PCR from human DNA and subcloned into the reporter plasmid pG12-promoter (Promega). Cell lines were grown and transfected as described (13). At least eight independent pools of stably transfected cells were assayed for each construct using a microplate luminometer as described previously (13). Deletions in the Fli-1 and PRH enhancer constructs were generated by PCR and confirmed by sequencing. Chromatin immunoprecipitation assays were performed in the haematopoietic progenitor cell line 416B as described (13). Primer sequences used to amplify the Fli-1 and PRH/Hex enhancers were Fli-1E-Fw TGGTCTGCAAGGTAAGGAA, Fli-1E-Rev TATGTTTGCTCACAAGATCC, PRHE-Fw GACCCCTTCCGTACATACAGA and PRHE-Rev AAGT CACATCCACAGAGGA.

Transgenic mice were generated using pronuclear injection into CBA3C57Bl/6 fertilized mouse oocytes which were allowed to divide into two cells prior to implantation into the oviducts of pseudopregnant CD1 female mice. For analysis of embryonic enhancer activity, recipient females were sacrificed and embryos analyzed at E11.5 days of gestation (F0 analysis). Embryos were stained with X-gal for β-galactosidase activity as described (21). For genotyping, DNA from yolk sac was prepared and tested for the presence of the lacZ transgene by PCR with an internal myogenin control.

ACKNOWLEDGEMENTS

We are grateful for many stimulating discussions with members of our laboratories and also with D. Vetrie, P. Couttet and D. Bentley (Wellcome Trust Sanger Institute). Work in the authors’ laboratory is funded by the Leukaemia Research Fund, Wellcome Trust, BBSRC, Cambridge MIT Institute, an IBM SUR equipment grant and the Canadian Institute of Health Research.

REFERENCES


