D86–D90 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh056

HemoPDB: Hematopoiesis Promoter Database, an information resource of transcriptional regulation in blood cell development

Twyla T. Pohar, Hao Sun and Ramana V. Davuluri*

Division of Human Cancer Genetics, Comprehensive Cancer Center, Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, Columbus, OH 43210, USA

Received August 15, 2003; Revised and Accepted September 23, 2003

ABSTRACT

Hematopoiesis describes the process of the normal formation and development of blood cells, involving both proliferation and differentiation from stem cells. Abnormalities in this developmental program yield blood cell diseases, such as leukemia. Although, in recent years, extensive molecular research in normal hematopoietic development has characterized transcription factors and their binding sites in the target gene promoters, the information generated is highly fragmented. In order to integrate this important regulatory information with the corresponding genomic sequences, we have developed a new database called Hematopoiesis Promoter Database (HemoPDB). HemoPDB is a comprehensive resource focused on transcriptional regulation during hematopoietic development and associated aberrances that result in malignancy. HemoPDB (version 1.0) contains 246 promoter sequences and 604 experimentally known *cis*-regulatory elements of 187 different transcription factors, with links to published references. Orthologous promoters from different species are linked with each other and displayed in the same database record, accompanied by a visual image of the promoters and corresponding annotations of *cis*-regulatory elements. HemoPDB may be searched for the promoter of a specific gene, transcription factors and target genes, and genes that are expressed in a certain cell type or lineage, through a user-friendly web interface at http://bioinformatics.med.ohio-state.edu/HemoPDB. Links to the documentation and other technical details are provided on this website.

INTRODUCTION

Hematopoiesis is the process by which mature blood cells of distinct lineages (e.g. red, white and lymphoid cells) are produced from pluripotent hematopoietic stem cells (HSCs) (1). This highly orchestrated process involves a complex interplay between the intrinsic genetic processes of blood cells and their environment. This interplay determines whether HSCs, progenitors and mature blood cells remain quiescent, proliferate, differentiate, self-renew or undergo apoptosis (2). Hence, there are various extracellular and intracellular stimuli that result in the activation of specific downstream signaling cascades. Ultimately, all signal transduction pathways converge at the level of gene expression where positive and negative modulators of transcription delineate the pattern of gene expression. Transcription factors (TFs), therefore, represent a nodal point of hematopoietic control through the integration of the various signaling pathways and subsequent modulation of the transcriptional machinery (3).

TFs are sequence-specific DNA-binding proteins with a variety of functions that include: (i) folding of the DNA molecule into distinct domains; (ii) the initiation of DNA replication; and (iii) the control of gene transcription (4). In order to understand the process of hematopoietic regulation, it is important to identify and characterize the TFs that positively and negatively regulate important genes for the normal development of cells in the various hematopoietic lineages. The past several years have yielded a great deal of research in the identification of TF and TF binding sites important to hematopoiesis. Although the majority of such protein-coding genes in mammals have been annotated, the genomic mapping of promoters and *cis*-regulatory elements associated with these coding sequences continues to be a formidable challenge.

A portion of published research findings have been added to publicly available databases: GenBank (5), TRANSFAC (6) and MPromDb (http://bioinformatics.med.ohio-state.edu/MPromDb). Although these resources provide a wealth of information, they are not specific to promoter annotation and hematopoietic regulation. Consequently, we have developed a publicly available, web-based resource, HemoPDB (http://bioinformatics.med.ohio-state.edu/HemoPDB). HemoPDB is composed of integral regulatory information specific to hematopoiesis, including TFs, target genes and promoter annotation. To our knowledge, this is the only database of its kind, with the exception of EpoDB (7), which is a database of genes expressed exclusively in red blood cells.

*To whom correspondence should be addressed at 420 West 12th Avenue, TMRF 524, Columbus, OH 43210, USA. Tel: +1 614 688 3088; Fax: +1 614 688 4006; Email: Davuluri-1@medctr.osu.edu

SIGNIFICANCE

In the past few years, a number of studies have focused on the dominant role of TFs in normal hematopoietic differentiation. It has become evident that TFs are important regulators of hematopoiesis, from the analyses of mice deficient in TF proteins and the characterization of chromosomal breakpoints in human leukemia (8). Disruption of the expression, sequence, structure of critical TFs or their associated regulatory proteins may upset the delicate balance between proliferation and differentiation and lead to leukemogenesis (9). A growing number of TFs (Supplementary table 1) that regulate a wide range of hematopoietically relevant genes and their disruptions in blood-related cancers have been discovered. Further, the availability of various mammalian genomes (human, mouse and rat) provides an excellent opportunity to integrate the existing information. An integrated resource of transcriptional regulation in hematopoiesis may contribute to the elucidation of the complex regulatory mechanisms of hematopoiesis and events that lead to malignancy progression.

DATA ACQUISITION

The majority of the data housed in HemoPDB are curated manually. A comprehensive literature review is performed periodically to ensure inclusion of recently published data. The remaining data, limited to TFs, binding sequences, target genes and references, are acquired from the aforementioned public databases via a data-mining pipeline. TF classes and gene descriptions are obtained manually for all entries, regardless of source.

DATABASE ORGANIZATION

The promoter and *cis*-regulatory sequences, corresponding attributes and annotation data are stored in a MySQL relational database. HemoPDB is structured as a set of relationships between several entities. We implemented a unique ID generating mechanism to establish the relationship between database tables. In addition, we created a local database system which functions as an annotation server, where a unique UniGene (10) cluster ID defines each gene. This method provides an efficient mechanism for data retrieval, in addition to allowing the consistent usage of conventional, widely used gene symbols. The graphical and textual presentation is accomplished by querying the data from respective tables.

The two largest tables are 'BindingSiteInfo', which stores the TF and binding site data, and 'PromoterInfo', which contains the coordinate information and relative position to the transcription start site (TSS) for every promoter. 'GeneInfo' utilizes each UniGene ID to efficiently annotate the promoters and respectively store their gene annotation.

In addition to these tables, the database includes several non-coordinate tables corresponding to the TF and target gene information, e.g. 'FactorInfo' stores the attributes for each TF, such as functional class and uniquely generated ID. 'Reference' provides efficient acquisition of the published citation corresponding with each *cis*-regulatory element. 'HMHomology' stores the information for human–mouse homology data, including gene symbols and promoter IDs for each orthologous gene pair. 'CpGScore' maintains the CpG score of all the promoters collected in HemoPDB.

Currently, HemoPDB uses MySQL as the database server and Jboss as the HTTP application server. It runs on Red Hat Linux Enterprise Edition 7.2. All the analysis for HemoPDB was written in Perl, while the client interface was all written in Java.

ANNOTATIONS

Mapping promoter and first exon sequences to the genome

The sequence upstream of the TSS to which RNA polymerase binds and accomplishes the initiation reaction for transcription defines the 'promoter region'. The sequence downstream of the TSS, however, may also be inherent to initiation. This entire region, i.e. the flanking region of the TSS, is obtained via the utilization of experimentally characterized exons, promoters and full-length cDNAs from either GenBank or DBTSS (11), a database of full-length human cDNAs obtained via an oligo-capping method. Composite queries to GenBank, such as {'homo sapiens'(ORGN) AND ['5′UTR'(FKEY) OR 'promoter'(FKEY) OR 'exon'(FKEY)]}, are performed through Entrez (12). The records are then parsed with a set of Perl scripts in order to obtain the first exon (TSS to donor site) sequences. The full-length mRNA sequences were downloaded from DBTSS and aligned against the genome using BLAT (13) to determine the first exon coordinates of the corresponding genes.

Annotating the promoter with gene information and *cis*-regulatory elements

Each promoter and first exon is associated with a GenBank accession ID and gene symbol. We use this information to obtain the corresponding promoter annotation from the UniGene database. Each binding sequence is then extended an additional 80 bp downstream, to ensure accuracy of the data. These extended binding sequences are then mapped to the corresponding genomic sequence (human: NCBI Build 30 or mouse: MGSC V3).

Once the promoter and binding sequences have been mapped to their respective genome, a SQL procedure is utilized to efficiently associate each binding site (*cis*-regulatory element) with its corresponding promoter, provided its genomic position falls within the predefined coordinates of the promoter sequence.

DATA ACCESS AND VISUALIZATION

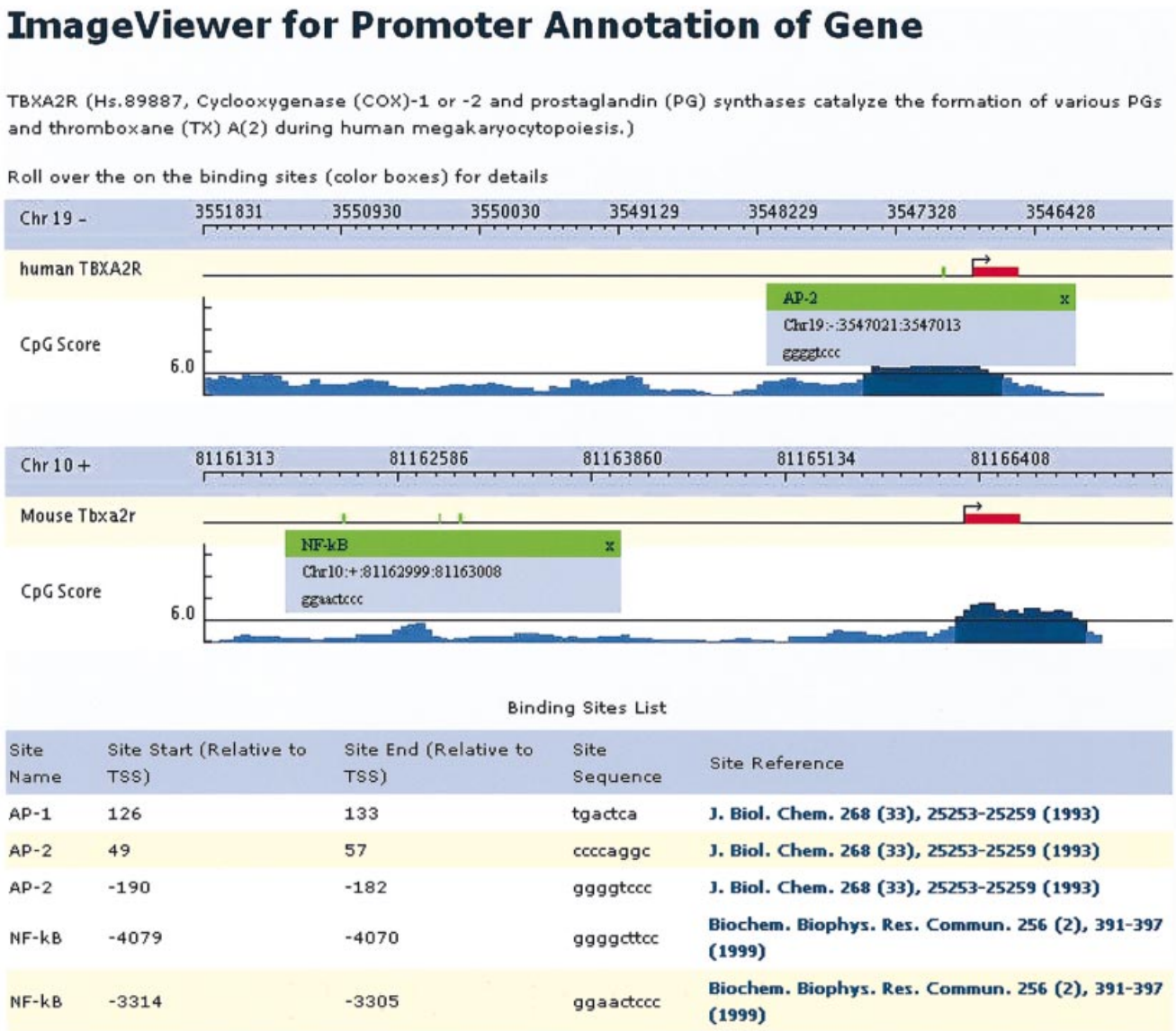
HemoPDB has an efficient interactive web interface, which provides selective attributes for TFs and hematopoietic-specific genes in textual and graphical form. HemoPDB may be accessed via the database link on the OSU HCG Bioinformatics home page (http://bioinformatics.med.ohio-state.edu) or the direct link mentioned previously. A user may query a TF of interest via TF name; a gene of interest via gene symbol, GenBank accession or UniGene ID, lineage or cell type within which it is expressed. Currently, the available query options for species include human and mouse.

You are searching *Transcription Factor (Binding Site) Name* with *PU.1*

Factor Name	Factor Species	Target Gene	Factor Functional Class	links
PU.1	Human	CD53	Ets-type	G
PU.1	Human	CD68	Ets-type	G
Pu.1	Human	ITGAM	Ets-type	G
PU.1	human	ITGAM	Ets-type	G
PU.1	human	RUNX1	Ets-type	G

Records 1-5 of 5 First Previous Next Last

Figure 1. Screen shot of the TF record for PU.1. Fields contain TF name, species, target gene, TF functional class and links to NCBI. Links to both target gene promoter annotation and GenBank are highlighted in blue text.



ImageViewer for Promoter Annotation of Gene

TBXA2R (Hs.89887, Cyclooxygenase (COX)-1 or -2 and prostaglandin (PG) synthases catalyze the formation of various PGs and thromboxane (TX) A(2) during human megakaryocytopoiesis.)

Roll over the on the binding sites (color boxes) for details

Binding Sites List

Site Name	Site Start (Relative to TSS)	Site End (Relative to TSS)	Site Sequence	Site Reference
AP-1	126	133	tgactca	J. Biol. Chem. 268 (33), 25253-25259 (1993)
AP-2	49	57	ccccaggc	J. Biol. Chem. 268 (33), 25253-25259 (1993)
AP-2	-190	-182	ggggtccc	J. Biol. Chem. 268 (33), 25253-25259 (1993)
NF-kB	-4079	-4070	ggggcttcc	Biochem. Biophys. Res. Commun. 256 (2), 391-397 (1999)
NF-kB	-3314	-3305	ggaactccc	Biochem. Biophys. Res. Commun. 256 (2), 391-397 (1999)

Figure 2. Sample output for the TBXA2R gene. The header includes the gene symbol (UniGene ID, role in hematopoietic development). The visual module displays species of interest (on top) and a graphical representation of promoter annotation. The TSS is indicated by a 90° arrow, the first exon by a red rectangle. Each small green rectangle represents a *cis*-acting element, position relative to TSS. Mouse-over options provide TF name, genomic position and binding sequence. The corresponding binding site reference is highlighted in blue text and is linked to its PubMed record.

Gene Symbol	Linage Cell Type	Gene Description	Species	Genome View for Promoter Annotation	links
FLI1	Erythroid, Myeloid, Lymphoid	Friend leukemia integration 1 (Fli-1) is a member of the Ets family of transcriptional activators that has been shown to be an important regulator during megakaryocytic differentiation. Interaction of Fli-1 with GATA-1, a well-characterized, zinc finger transcription factor is critical for both erythroid and megakaryocytic differentiation. (Mol Cell Biol. 2003 May;23(10):3427-41.)	human	Chr11:128591668-128598653	U , G
IL1A	Lymphoid (B, specifically)	IL-1 alpha has negative regulatory role in early stages of B lymphopoiesis.	human	Chr2:113447835-113454790	U , G
SIL	pre-HSC, mature lymphoid and megakaryocytic cells	SCL encodes basic helix-loop-helix bHLH TF with an essential role in specifying HSC SCL and GATA-1 interact in a transcriptional complex with the LIM domain protein LMO-2AJ243474	human	Chr1:47137241-47143379	U , G
CD53	lymphoid, myeloid	tetraspanin protein mostly expressed in to the lymphoid-myeloid lineage	human	Chr1:110512830-110518978	U , G

Figure 3. Screen shot output for a lymphoid query in the lineage or cell type field. Representative genes that are known to be expressed in this particular lineage are displayed. In addition, its corresponding role in hematopoiesis, species and promoter annotation, and the respective links are provided.

The HemoPDB database management and information presentation are implemented via an in-house-developed Java application framework called Genome Visualization Tool Kit (GDVTK). The visual presentation is in the form of an image map of regulatory regions including interactive contextual menus for easy navigation. A web interface to HemoPDB has been developed using J2EE technology (JSP and Java Servlet).

The ability to access information about each TF, such as target gene, binding site positions relative to the TSS, binding site sequence and functional class, is particularly useful for those who are interested in the characteristics, binding preference or mechanism of a specific TF. Each TF (binding site) query acquires the corresponding TF record, which displays the TF name, species, target gene and TF functional class. Links to the target gene promoter annotation and GenBank record allow integration of relevant target gene information (Fig. 1).

Alternatively, information about a specific gene: lineage or cell type within which it is expressed, genomic position, annotated *cis*-regulatory elements, description of its role in hematopoietic development and literature references, offers the user an opportunity to view regulatory information at the transcriptional level in textual and graphical forms. The visual module provides depiction of the promoter with corresponding TSS, exon 1 position and *cis*-elements relative to the TSS. The mouse-over option is a user-friendly feature that contains the respective TF name with genomic position and binding sequence. The CpG score is depicted as a histogram, where the line of score 6.5 is used to represent the cut-off value to determine whether the promoter is CpG or non-CpG related. The textual data provides information in static form, which includes TF name, binding site positions and sequence, and respective binding site reference, with link to PubMed (14) (Fig. 2).

The characterization of lineage and cell type specificity of a gene, during the various developmental stages of hematopoietic differentiation, is especially useful for pattern delineation of transcriptional regulation. Each record displays every gene in the database that has been experimentally characterized to be present or expressed in the particular lineage or cell of interest. In addition, each gene is presented with its corresponding descriptive role in hematopoiesis and a link to its promoter annotation (Fig. 3). Since the promoters of orthologous genes are interlinked, HemoPDB serves as a platform for comparative genomics of transcriptional regulation in hematopoiesis.

FUTURE DIRECTIONS

The long-term goal of this project is to contribute to the understanding of hematopoietic transcriptional regulation and corresponding aberrances that lead to malignancy. As more data are published, we will continue to incorporate the annotations into the content of HemoPDB. This database will provide the foundation to further develop new features linked to the database such as promoter-specific and TF-specific databases explicitly for hematopoietic regulation. Similar resources, which combine analogous features, such as the *Arabidopsis* Gene Regulatory Information Server (AGRIS) (15), are appropriately designed to integrate these regulatory data and are suitable models for HemoPDB.

High-throughput technologies, such as DNA microarrays, are facilitating large-scale comparisons of gene expression in normal versus cancer cells. However, there are still many

unanswered questions concerning the regulatory roles that TFs play in both normal and malignant development. Future plans include the integration of the data from genome-scale gene expression analysis such as EST, SAGE and microarray projects in addition to the incorporation of relative expression information acquired from UniGene clusters and links to the SOURCE database (16).

In conclusion, HemoPDB provides integral, hematopoiesis-specific, transcriptional regulatory information in a sensible and easily accessible way. It provides the foundation to comprehend and analyze gene-specific through genome-scale data. We are hopeful its implementation will contribute toward the elucidation of the complex process of hematopoiesis.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank Saranyan, Sang Gook and Zhouhai for assistance with the web design and implementation. R.V.D. is a V foundation scholar; HemoPDB is partly supported by the V Foundation for Cancer Research and an institutional seed grant from the American Cancer Society.

REFERENCES

1. Godin,I.E., Garcia Porrero,J.A., Coutinho,A., Dieterlin-Lievre,F. and Marcos,M.A.R. (1993) Para-aortic splanchnopleura from early mouse embryos contains B1a cell progenitors. *Nature*, **364**, 67–70.
2. Krause,D.S. (2002) Regulation of hematopoietic stem cell fate. *Oncogene*, **21**, 3262–3269.
3. Barreda,D.R. and Belosevic,M. (2001) Transcriptional regulation of hemopoiesis. *Dev. Comp. Immunol.*, **25**, 763–789.
4. van Oostveen,J., Bijl J., Raaphorst,F., Walboomers,J. and Meijer,C. (1999) The role of homeobox genes in normal hematopoiesis and hematological malignancies. *Leukemia*, **13**, 1675–1690.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
6. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
7. Stoeckert,C.J.,Jr, Salas,F., Brunk,B. and Overton,G.C. (1999) EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Res.*, **27**, 200–203.
8. Lenny,N., Westendorf,J.J. and Hiebert,S.W. (1997) Transcriptional regulation during myelopoiesis. *Mol. Biol. Rep.*, **24**, 157–168.
9. Gomes,I., Sharma,T.T., Edassery,S., Fulton,N., Mar,B.G. and Westbrook,C.A. (2002) Novel transcription factors in human CD34 antigen-positive hematopoietic cells. *Blood*, **100**, 107–119
10. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
11. Suzuki,Y., Yamashita,R., Nakai,N. and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
12. Geer,R.C. and Sayers,E.W. (2003) Entrez: Making use of its power. *Brief. Bioinform.*, **4**, 179–184.
13. Kent,W.J. and Brumbaugh,H. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
14. McEntyre,J. and Lipman,D. (2001) PubMed: bridging the information gap. *CMAJ*, **164**, 1317–1319.
15. Davuluri,R.V., Sun,H., Palaniswamy,S.K., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: *Arabidopsis* Gene Regulatory Information Server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.
16. Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D., Brown,P.O. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.