A Bioinformatics Approach Toward Identification of Genes involved in Hematopoiesis and Leukemia

Twyla T. Pohar[1], Hao Sun[2], Sandya Liyanarachchi[3], S. James S. Stapleton[4] and Ramana V. Davuluri[5]

Keywords: database, data-mining, EST, hematopoiesis, leukemia, statistical analysis, gene expression.

1 Introduction.

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. This complex, biological process is under stringent control, with various extra- and intra-cellular stimuli that result in the activation of downstream signaling cascades. Ultimately, these signaling cascades converge at the level of gene expression where positive and negative modulators of transcription delineate the pattern of gene expression [1].

In order to elucidate the mechanisms involved in such regulation, it is important to identify the genes, which play integral roles in the hematopoietic process. Our in-house developed Hematopoiesis Promoter Database, HemoPDB, includes experimentally defined genes, which were manually curated from published literature [2]. Characterization of the gene expression profiles of these genes may allow us to efficiently identify previously unannotated genes via a combination of statistical analysis and data-mining of expressed sequence databases. We have designed a data-mining pipeline in order to manage the data obtained from dbEST [3] to determine the UniGene [4] clusters likely representing genes expressed preferentially in hematopoietic tissues and organs.

2 Results.

We downloaded all of the ESTs via dbEST: ftp://ncbi.nlm.nih.gov/repository/dbEST to create a standardized gene expression database for each sequence according to organ/tissue. Our automated pipeline then determines the genomic coordinates of each sequence by its alignment to the genome by BLAT [5]. These coordinates are then associated with the corresponding UniGene cluster. We then perform a genome-wide statistical analysis by applying a binomial test to compare the proportion of organ/tissue-specific ESTs expressed in each cluster vs. the entire EST population. The imposed criterion of a p-value, which conforms to a specific level of statistical significance, allows us to determine the clusters that are hematopoietic-related. We intend to substantiate these data obtained

[1] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 570A, Columbus, Ohio, USA. E-mail: pohar-2@medctr.osu.edu

[2] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 570B, Columbus, Ohio, USA. E-mail: sun.143@osu.edu

[3] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 570B, Columbus, Ohio, USA. E-mail: liyanarachchi-1@medctr.osu.edu

[4] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, Dept. of Physics, 420 West 12th Ave. Room 570, Columbus, Ohio, USA. E-mail: stapleton.41@osu.edu

[5]Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 524, Columbus, Ohio, USA. E-mail: davuluri-1@medctr.osu.edu

by comparing them to the gene expression profiles of experimentally characterized hematopoietic genes.

We will present the development of our automated pipeline, in addition to results obtained from this genome-scale analysis. We are hopeful that this study will allow the identification of key genes in the normal and malignant hematopoietic environments.

3 Figures.

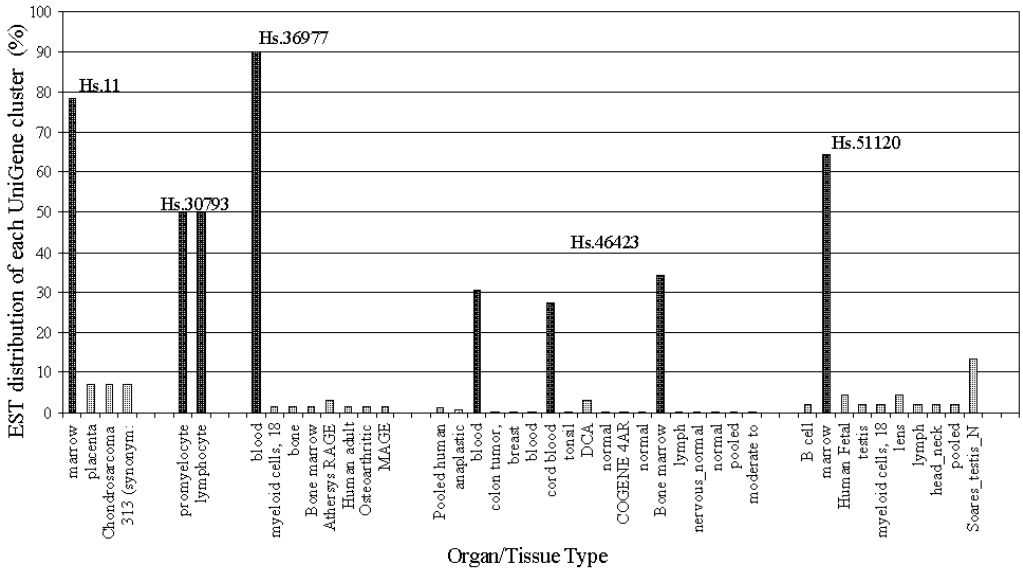

Figure 1. Sample EST percentage distribution of 5 UniGene clusters obtained via our automated method, which identifies hematopoietic-related genes. The highest percentages, conveyed by black bars, are consistent with hematopoietic organ/tissue types.

4 References.

[1] Barreda, D.R. and Belosevic M. 2001. Transcriptional regulation of hemopoiesis. *Developmental and Comparative Immunology* 25: 763-789.

[3] Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags". 1993. *Nat Genet.* 4(4): 332-3.

[5] Kent, W.J. and Brumbaugh, H. 2002. BLAT--the BLAST-like alignment tool. *Genome Res.* 12: 656-664.

[2] Pohar, TT, Sun, H. and Davuluri, RV. 2004. HemoPDB: An information resource of transcriptional regulation in blood cell development. *Nuceleic Acids Research* 32: D86-D90.

[4] 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. and Wagner, L. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31: 28-33.