4416–4428 *Nucleic Acids Research, 2006, Vol. 34, No. 16* *Published online 26 August 2006*
doi:10.1093/nar/gkl578

Hematopoietic gene promoters subjected to a group-combinatorial study of DNA samples: identification of a megakaryocytic selective DNA signature

Yehonathan Hazony[1],*, Jun Lu[2], Cynthia St. Hilaire[2] and Katya Ravid[2],*

[1]College of Engineering, Boston University, Boston, MA, USA and [2]Department of Biochemistry, Boston University School of Medicine, 715 Albany Street, K225, Boston, MA 02118, USA

Received March 5, 2006; Revised July 10, 2006; Accepted July 24, 2006

ABSTRACT

Identification of common sub-sequences for a group of functionally related DNA sequences can shed light on the role of such elements in cell-specific gene expression. In the megakaryocytic lineage, no one single unique transcription factor was described as linage specific, raising the possibility that a cluster of gene promoter sequences presents a unique signature. Here, the megakaryocytic gene promoter group, which consists of both human and mouse 5′ non-coding regions, served as a case study. A methodology for group-combinatorial search has been implemented as a customized software platform. It extracts the longest common sequences for a group of related DNA sequences and allows for single gaps of varying length, as well as double- and multiple-gap sequences. The results point to common DNA sequences in a group of genes that is selectively expressed in megakaryocytes, and which does not appear in a large group of control, random and specific sequences. This suggests a role for a combination of these sequences in cell-specific gene expression in the megakaryocytic lineage. The data also point to an intrinsic cross-species difference in the organization of 5′ non-coding sequences within the mammalian genomes. This methodology may be used for the identification of regulatory sequences in other lineages.

INTRODUCTION

Functional regulatory elements for gene expression reside in the genome in the form of short subsequences. In the majority of the identified cases, these regulatory elements appear in promoter regions upstream of gene coding sequences. They are often recognized by transcriptional factors which activate/suppress gene transcription. In some other cases, regulatory elements appear in the 3′-untranslated region of a gene, and modulate the stability and translatability of the transcribed message through protein factors, or through RNA molecules, such as micro RNAs, as reviewed by Bartel (1). A recent study by Xie *et al*. has shown the feasibility and significance of genome-wide analysis to extract common regulatory elements conserved in several species (2). From a different perspective, common-sequence analysis may also be applied to identify common mechanisms governing the expression of co-regulated genes.

This article presents a case study on the megakaryocytic promoter group. Megakaryocytes are hematopoietic cells that give rise to platelets, as discussed by Ravid *et al*. and Shivdasani (3,4). During the differentiation of megakaryocytes, lineage-specific/selective activation of genes takes place as reviewed by Kaluzhny (5). The following genes are selectively co-expressed in megakaryocytes: the platelet factor 4 (PF4) (6), glycoprotein IIb (GPIIb) (7), glycoprotein-V (GPV) (8–10), glycoprotein VI (GPVI) (11) and c-Mpl (12,13). The whole change in gene expression profile is important for megakaryocyte/platelet development and function (3). However, the mechanism by which these genes are all selectively expressed in megakaryocytes is not fully understood. In other lineages of the hematopoietic system, often genes expressed in the same lineage share a common mechanism of control, such as common DNA binding sites to PU.1 factor in genes expressed in the myeloid lineage as reviewed by Friedman (14). No unique, tissue-selective transcription factor has been identified in the megakaryocytic lineage. It is then reasonable to hypothesize that there is a common mechanism underlying the megakaryocyte-specific gene regulation, but this might rely on a, yet unidentified, unique combination of common sequences. In support of this contention, all megakaryocyte-expressing genes are regulated by DNA binding sites to Ets

*To whom correspondence should be addressed. Tel: +1 617 638 5053; Fax: +1 617 638 5054; Email: ravid@biochem.bumc.bu.edu
*Correspondence may also be addressed to Yehonathan Hazony. Tel: +1 617 353 3270; Email: hazony@bu.edu

and GATA-1 transcription factors (6–13). However, it is not clear whether these are the only factors regulating specific gene expression in this lineage. Thus, the application of the presented computational platform on this group of promoters may help in identifying new regulatory sites and confirming already described ones.

A vast volume of work is reported in the literature of Bioinformatics concerning analysis of DNA sequences, focusing primarily on statistical methods (15–17), and various statistical-scoring techniques, expressed in statistical matrices [see also (18–22), http://www.npaci.edu/Press/98/111898_webeng.html]. The methodology discussed in the open literature commonly utilizes web-based platforms that enable the user to develop applications based on computer resources out of their control (15).

The development of 'standalone workstations' running higher-level graphic and command interfaces to the commonly used BLAST suite of software was recently described by Buisine and Chalmers (23). However, the implementation reported is restricted to workstations running under the Unix operating system, and limited to the capabilities of the BLAST software.

Another platform available for the development of customized applications is based on the BioMoby system (http://www.biomoby.org). A tool built using BioMoby, called Taverna (http://taverna.sourceforge.net) offers an 'open source' platform employing Web-based 'Grid Computing'. It provides a graphic-user interface that enables one to assemble whole process lines using services provided by servers scattered over the internet without having to develop any software.

The computational platform utilized for the current study differs from the two customizable systems described above. It focuses on a PC-based implementation that evolved in response to the specific study described in this paper. A battery of algorithms were embedded in an interactive network of graphic display and control screens, and interfaced with a server-based database. These algorithms were designed to extract common subsequences from a group of DNA samples, extended to mono-gap sequences with a varying gap size. The results for gap sizes varying between 0 and 10 elements are utilized to identify common, double-gap segments, which are further extended to identify long common multi-gap sequences.

The hardware platform consists of a Personal-Computer (PC)-based Workstation, which can be used in stand-alone mode or as part of a server-controlled, distributed network of PCs. Such platforms are commonplace in many research and process laboratories. The software platform was originally developed for the design, implementation, operation and teaching of manufacturing-process control networks, as described by Hazony (24,25). It is particularly adept for iterative development of customized applications, providing a problem solving paradigm in which both problem specification and problem solution are concurrently refined through an iterative process. The present report describes such a customized application. However, the platform can be easily adapted for other purposes.

Finally, the false-positive-identification (FPI) investigation, reported below, is based on two different methods: one, using searches of randomized samples, and the second, utilizing available bench-mark batches of 18 406 samples, each of 5000 elements upstream of genes (in relation the ATG switch), extracted from the entire human genome (http://genome-archive.cse.ucsc.edu/goldenPath/gh16/bigZips/), and a corresponding batch of 14 102 samples, each of 5000 elements upstream of genes, extracted from the entire murine genome (http://genome-archieve.cse.ucsc.edu/goldenPath/mm4/bigZips/).

The focus of this study has been to describe an application of a computational platform that allows examination of the hypothesis that genes selectively expressed in one hematopoietic lineage share a unique combination of regulatory elements. Using the megakaryocyte as a model system, our study identified a cluster of sequences that are unique to gene promoters selectively active in this lineage. Within these sequences, Ets and GATA binding sites were recognized, imbedded in clusters of sequences highly selective to megakaryocyte promoters, and for which mammalian binding factors have not been identified yet. The findings validate our unbiased search method for conserved regulatory elements within a related group of genes, as Ets and GATA binding sites have been previously described to regulate genes specifically expressed in megakaryocytes. This method may be also used for searching other groups of genes within and out of the hematopoietic lineage.

MATERIALS AND METHODS

Hardware

The hardware platform used for this work consists of a cluster of contemporary PCs running under the Windows-XP operating system and supported by a Windows-2000 Server running on a separate machine. The communication is provided via Ethernet, which provides also rapid access to the internet. Each machine is equipped with a 3 GHz Pentium4 processor and 1 Gb of high-speed memory (RAM). One PC was upgraded to a RAM value of 3 Gb to accommodate the extensive-temporary memory required for the full FPI runs described above. All the work described in this article was executed on this upgraded machine. However, some of the exploratory studies benefited from a 'production-line' mode of operation afforded by the server-controlled cluster of personal computers.

Software

The software platform used for the present work is an adaptation of a server-controlled, distributed network of PCs, developed for the design, implementation, operation and teaching of manufacturing-process control. This is an intranet-based configuration, connected to the internet for the purpose of database exchanges over the web as well as for remote access to the system. The system is based on a computational paradigm that has evolved in past several decades in the context of CAD/CAM and Process Control. In this computing paradigm the end user of the methodology is totally divorced from the software via several layers of 'user interfaces' and tools for software automation (24,25). This is accomplished employing a 'network' of utilities such as textual and/or graphic control panels, popup windows, pull-down command lists, database services, mathematical tools, rule-based algorithms, and graphic data mining and displays. Software customization is commonplace in this domain in the form of specialized controllers, machines and processes and the end user is the beneficiary of the customization but not necessarily its developer.

The specific platform used is equipped with extensive tools for the study and implementation of process customization, permitting the dual role of developer/user, where the user can toggle back and forth, under software control, between the two roles during application development. This capability was utilized during the development of the software employed in the final version of the current study. The system infrastructure, as well as its algorithmic components, are implemented in the APL2 programming language, which is available for most of the popular workstations and mainframe computers.

The Megakaryocyte Group, Determination of Group-Combinatorial Sequence Incidences (GCSI) and the algorithms used

A total of nine megakaryocyte gene promoters (MegaKP) were analyzed, including five human (hPF4, hGPV, hGPVI, hGPIIB, hMpl) and four mouse (mPF4, mGPIIb, mGPV, mMpl) sequences. The focus of analysis was on 5 kb upstream to the ATG (sequences derived from GenBank) in genes and species for which megakaryocyte-specific expression has been previously demonstrated to be driven by the 5′ non-coding region (6–13). Intron and 3′ non-coding regions were not included in this computational search, as they have not been demonstrated to be crucial for promoter activity for any of the above genes. Moreover, this selection allowed comparison of sequences equal in length and gene topography. The goal has been to apply an algorithm to identify within these segments sequences spaced by different gaps, which might be selectively appearing in the megakaryocyte group, as compared with random and other specific sequences.

DNA segments are built on a combinatorial arrangement of four elements represented by the letters A, C, G and T, and on the complementary nature of the AT and CG pairing. Two DNA segments of the same length are considered complementary if the elements of one segment are complementary to the respective elements of the second segment in reverse order, i.e. ACTG is complementary to CAGT. A comprehensive group-combinatorial scan of nine DNA samples, associated with the MegaKP group, was performed, searching for the common presence of all complementary pairs of segments of length *N*, and the results are shown in the first row of the table depicted in Figure 1.

A rapidly-converging algorithm is employed, based on a ‘tree-pruning’ approach, which reduces the search space to a manageable size, while guarantying the completeness of the search. The algorithm performs a concurrent-comprehensive search for sequences of length *N* that are common to a group of related DNA samples. It starts scanning a full combinatorial list for $N = 4$, and eliminates entire tree branches based on the absence of some short sequences from this initial list. The reduced list serves as a starting list for the derivation of the reduced combinatorial list for sequences of length of $N + 1$. The iterative process proceeds until reaching an *N* value for which the reduced list is empty. A conceptually-similar algorithm was reported by Brazma *et al.* (19). The merit of such an algorithm is emphasized when the numeric values in the zero-gap row (Figure 1A) are compared with the corresponding combinatorial values of $\sim 0.5 \times 4^N$. The combinatorial-search algorithm used retains also the upstream-starting location of each individual occurrence of each member of the recorded-segment list.

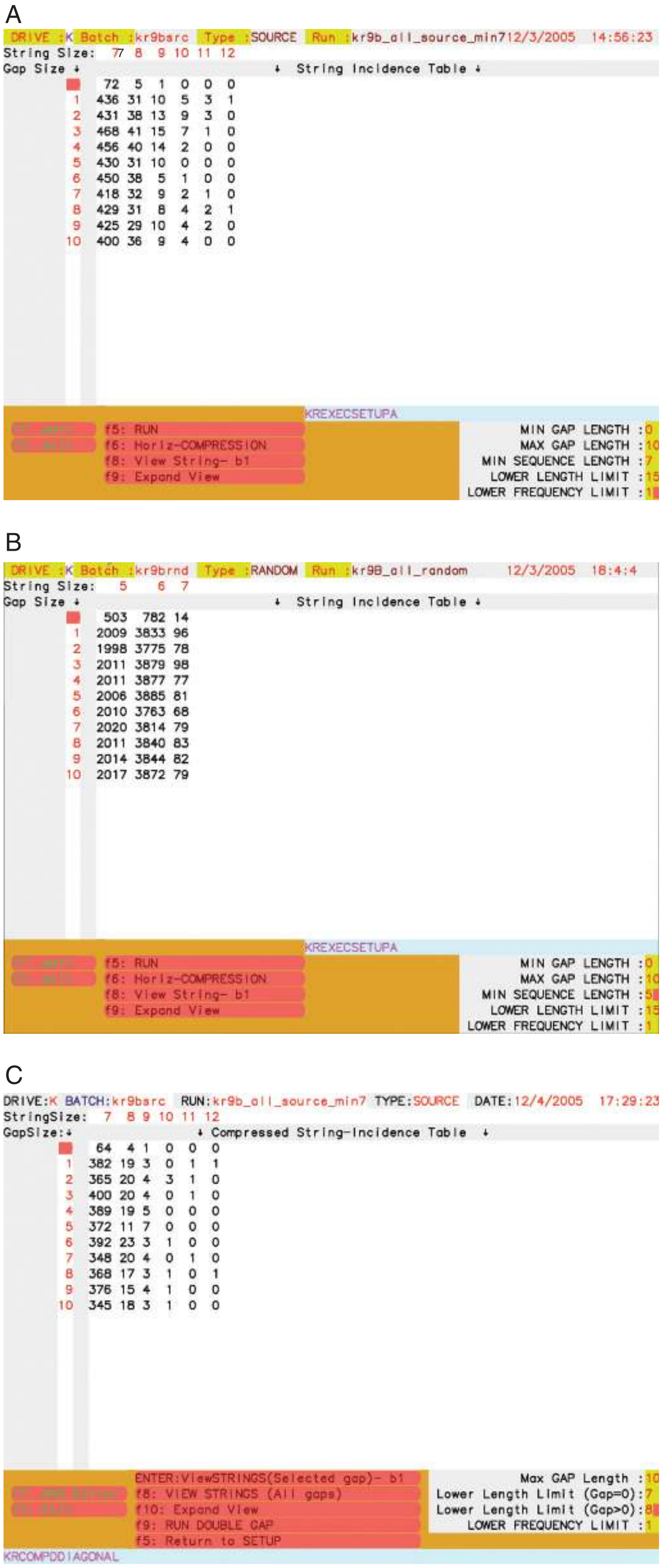
The results of a similar search algorithm for sequences that include a single ‘gap’ are also shown in the corresponding rows in the table depicted in Figure 1A, for gaps of lengths between 1 and 10. The presence of a gap in a segment implies that the search algorithm ignores the elements occupying the respective gap locations. The numbers shown in the table represent the number of incidences of at least one of each complementary pair of length *N*, where *N* represents the number of non-blank elements in a sequence. The header of the table shown in Figure 1A indicates the length *N* of sequences recorded in a particular column of the table, and the gap-size column on the left indicates the single-gap size recorded in the particular row of the table.

The total number of sequence-searches employed to reach a certain column in a row, corresponding to a particular gap (including 0), is obtained by summing up all preceding numbers on the left. The tree-pruning algorithm used guarantees a complete search. A direct link to the database is embedded in the algorithm creating a structured archive that includes all the data needed for a customized data-mining process. These include data labels, sequence-incidence records and the specific location for each occurrence.

Figure 1A shows also the display/control panel of the workstation, which includes the relevant experimental control parameters, i.e. (i) Minimum sequence length—this determines which is the first *N* value recorded in the database supporting the display. This parameter does not affect the execution of the algorithm but defines which of the data are stored for further use and (ii) Lower frequency limit—this parameter sets a minimal value for the number of occurrences of a particular sequence in the sample being investigated to qualify as a score in the GCSI table. The actual value used in the present study is 1, meaning that the sequence scanned occurs at least once in each of the samples belonging to the selected group.

Electromobility shift assay

Y10L8057 megakaryocytic cells (6) were cultured in F-12 media (Invitrogen) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin at 37°C. To induce differentiation, cells were washed and resuspended in IMDM media (Invitrogen) supplemented with 10% FBS, 1% penicillin/streptomycin and 25 ng/ml thrombopoietin and cultured for 2–3 days. As control, we used mammary epithelial cells (NMuMG, ATCC CRL #1636) grown as instructed by ATCC. Cells were washed 2× with ice cold 1× phosphate-buffered saline (PBS), resuspended in lysis solution (10 mM Tris–HCl, pH 7.6, 10 mM NaCl, 3 mM $MgCl_2$, 0.5% NP-40), incubated on ice for 5 min, centrifuged at 500× *g* for 5 min and then washed with 1 ml lysis buffer. Nuclei were resuspended in freezing buffer (50 mM Tris–Hcl, pH 8.3, 5 mM $MgCl_2$, 0.1 mM EDTA, 40% glycerol) at $\sim 10^7$ nuclei/100 µL and snap frozen. An equal volume of NBP solution [20 mM Hepes, pH 7.9, 0.4 M NaCl, 15 mM $MgCl_2$, 0.2 mM EDTA, 25% glycerol, 0.5 µg/mL DTT supplemented with 1× miniComplete® proteinase inhibitors cocktail (Roche)] was added to the frozen nuclei and the extraction stirred gently with a cut pipette tip and incubate on ice for 20 min. The extraction was then centrifuged at 14K r.p.m. for 30 min at 4°C and supernatant recovered. Protein concentration determined using Bio-Rad protein assay kit. Extract aliquots are stored at −80°C. Electromobility shift assay (EMSA) was used to determine binding of nuclear proteins to oligomers, as described in Lu *et al*. (6). Briefly, samples contained 20 µg of nuclear extracts, 5 µg poly(dI–dC) (Amersham Biosciences), 1× binding buffer (10 mM HEPES, pH 7.9, 1 mM DTT, 0.1% Triton-X, 0.5% glycerol) and 5 µl of labeled oligo diluted 1:10 and incubated 0.5–1 h at room temperature. Protein–DNA complexes were resolved on a 4% native polyacrylamide gel [4% acrylamide, 1× TBE (90 mM Tris–Borate, 2 mM EDTA, pH 8.0), 0.6% ammonium persulfate, 0.06% TEMED] in 1× TBE buffer at 130 V for ∼2.5 h. Gel was fixed in solution of 10% methanol, 10% acetic acid and dried on a gel dryer (Biorad) and analyzed by autoradiography. The wild-type oligo used was 5′-AGCTACTTGNGAGGCCGAGGCAG***GAGAAT***TGCTTGAA-3′. The mutated oligo was (note italic bases) 5′-AGCTACTTGNGAGGCCGAGGCAG***CATGTA***TGCTTGAA-3′. As control to test the nuclear extracts, assays were also performed using an Sp1 probe with the consensus site in bold) **5′–C**CGACTGCGG**CCCCGCCCC**TTAGAACGTTGTGACGTAGGAGCATTCCACG-3′. To anneal oligos 10 µg of each was boiled for 5 min in 10 mM Tris, pH 8.0, and 10 mM $MgCl_2$ followed by gradual cooling to 4°C. Annealed oligos were diluted 1:5 and labeled with $[\gamma\text{-}^{32}P]$ATP using T4 polynucleotide kinase (New England Biolabs) according to manufacturers protocol. Sample volume was brought to 100 µl with STE (10 mM Tris–HCl, pH 7.8, 10 mM NaCl, 1 mM EDTA) and then run through a Sephadex G-50 column 2× before incubation as above.



```
DRIVE :K Batch :kr9barc  Type :SOURCE  Run :kr9b_all_source_min712/3/2005  14:56:23
String Size:  77  8   9  10  11  12
Gap Size ↓                              ↓ String Incidence Table ↓
            72   5   1   0   0   0
        1  436  31  10   5   3   1
        2  431  38  13   9   3   0
        3  468  41  15   7   1   0
        4  456  40  14   2   0   0
        5  430  31  10   0   0   0
        6  450  38   5   1   0   0
        7  418  32   9   2   1   0
        8  429  31   8   4   2   1
        9  425  29  10   4   2   0
       10  400  36   9   4   0   0

                                KREXECSETUPA
        f5: RUN                                       MIN GAP LENGTH :0
        f6: Horiz-COMPRESSION                         MAX GAP LENGTH :10
        f8: View String- b1                      MIN SEQUENCE LENGTH :7
        f9: Expand View                           LOWER LENGTH LIMIT :15
                                               LOWER FREQUENCY LIMIT :1
```

```
DRIVE :K Batch :kr9brnd  Type :RANDOM  Run :kr9B_all_random      12/3/2005  18:4:4
String Size:    5     6   7
Gap Size ↓                              ↓ String Incidence Table ↓
            503   782  14
        1  2009  3833  96
        2  1998  3775  78
        3  2011  3879  98
        4  2011  3877  77
        5  2006  3885  81
        6  2010  3763  68
        7  2020  3814  79
        8  2011  3840  83
        9  2014  3844  82
       10  2017  3872  79

                                KREXECSETUPA
        f5: RUN                                       MIN GAP LENGTH :0
        f6: Horiz-COMPRESSION                         MAX GAP LENGTH :10
        f8: View String- b1                      MIN SEQUENCE LENGTH :5
        f9: Expand View                           LOWER LENGTH LIMIT :15
                                               LOWER FREQUENCY LIMIT :1
```

```
DRIVE:K BATCH:kr9barc  RUN:kr9b_all_source_min7 TYPE:SOURCE  DATE:12/4/2005  17:29:23
StringSize:  7  8 9 10 11 12
GapSize:↓                   ↓ Compressed String-Incidence Table  ↓
          64   4 1   0   0   0
       1  382 19 3   0   1   1
       2  365 20 4   3   1   0
       3  400 20 4   0   1   0
       4  389 19 5   0   0   0
       5  372 11 7   0   0   0
       6  392 23 3   1   0   0
       7  348 20 4   0   1   0
       8  368 17 3   1   0   1
       9  376 15 4   1   0   0
      10  345 18 3   1   0   0

            ENTER:ViewSTRINGS(Selected gap)- b1              Max GAP Length :10
            f8: VIEW STRINGS (All gaps)         Lower Length Limit (Gap=0):7
            f10: Expand View                    Lower Length Limit (Gap>0):8
            f9: RUN DOUBLE GAP                     LOWER FREQUENCY LIMIT :1
            f5: Return to SETUP
KRCOMPDDIAGONAL
```

Figure 1. (**A**) GCSI table for the 9-sample MegaKP group. The rows correspond to single-gap sequences. Gap lengths vary between 0 and 10 elements, as indicated by the left-most column labeled as 'Gap Size'. The columns correspond to sequence length excluding the gap, as indicated by the header row labeled as 'String Size'. For example, the entry of the table corresponding to gap-size of 4 and string-size of 9 contains 14 incidences of sequences, each one consisting of two shorter sub-sequences separated by a gap of four elements, while the sum of non-gap elements is nine. The top row includes key-wards serving for storage and retrieval of these data in the database. The bottom part of the figure includes lists of 'commands' on the left and list of operational parameters on the right to illustrate the format of pursuit with this application. (**B**) The GCSI table for the randomly scrambled 9 samples of the MegaKP Group. The layout of this figure is the same as in (A). The contents of the figure changed in several ways. The entry in the 'Type' field in the top row has changed automatically to 'RANDOM' reflecting the fact that the derivation process started with a randomizing algorithm. The 'MIN SEQUENCE LENGTH' parameter was changed to 5, resulting in the first column of the GCSI table containing 5-element sequences (excluding the gap). Finally, the longest common sequences found are of lengths of seven elements (excluding the gap), depicting the impact of the sample-randomization process. (**C**) The compressed-GCSI table for the 9-sample MegaKP group. This figure varies from Figure 1A in the layout of the bottom part, including the lists of commands available and the list of operational parameters. Furthermore, the numeric contents of the Table have changed as a result of the application of two compression algorithms designed to remove redundancies in the table (see text). The 'Lower-Length Limit' parameters will be used later to create a 'pre-selected list' of sequences.

RESULTS

Group-combinatorial sequence incidences search of randomly scrambled samples

DNA sequences may be scrambled by re-indexing the elements using randomized-index vectors. Results obtained by applying the GCSI search to randomly scrambled samples are shown in Figure 1B. Repeated application of this method generates tables of varying contents but similar patterns. A most significant feature of the table is the entries in the last column on the right, which in this particular case correspond to string size of $N = 7$. This last column may have few more entries or be eliminated altogether in different randomization runs. Similarly, the numerical values in the preceding column, corresponding to string length of six, may vary and the listing of the sequences would differ owing to the randomization process involved.

The significance of the data summarized in this table is comparable with Figure 1A. The absence of entries in columns 8–12 in Figure 1B contrasts with the structure displayed in Figure 1A, suggesting that in the quest for sequences unique to the particular group at hand (MegaKP) it is more promising to focus on string lengths >7. A comparison of the magnitude of the entries in this column, in both tables, shows significantly larger incidence values for the non-scrambled table. The gross ratio between these values suggests FPI indicators of the order of 25% for the sequences recorded in column 7 of Figure 1B. Notice that in contrast to Figure 1A, the leading column in the GCSI table shown in Figure 1B corresponds to $N = 5$. This selection is controlled by the 'min sequence length' parameter defined in the parameter column shown at the bottom-right of the screen. This selection does not affect the search speed but rather the retrieval speed of archived data.

Compressed GCSI tables

Inspection of the contents of the GCSI table reveals substantial redundancy owing to two sources: (i) Horizontal redundancy—any row in the table includes all combination of sub-sequences derived from a longer sequence present on the right side of the row and (ii) Back-Diagonal redundancy—any sequence present at the top-right of the table spawns sequences of same length but larger gaps, which appear along back-diagonals of the table, extending from top-right to bottom left.

Two algorithms are employed to remove the redundancy from the data shown in Figure 1A, resulting in the compressed GCSI table presented in Figure 1C. A similar table is shown in Figure 2 for the 5-sample human subgroup of the MegaKP group, followed by the table corresponding to the 4-sample murine subgroup depicted in Figure 3. Note that the first column shown in Figures 2 and 3 corresponds to ($N = 7$) as in Figure 1A and C.

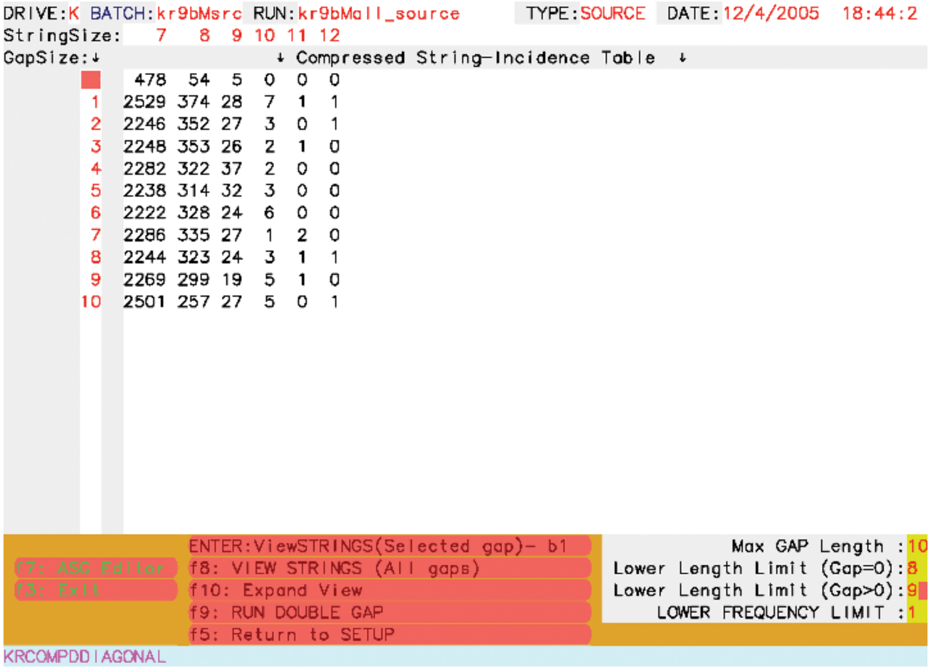
False-positive identification (FPI) indicators

The FPI indicators are used to sort out unique sequences as candidates for further studies. These indicators are derived using two different approaches: the first based on the comparison with results obtained for 'randomized samples', as discussed above. This method is refined (optionally) by scanning a pre-selected list of sequences against a large number, i.e. 1000, of randomized samples. In this case the FPI value is determined by the ratio between the number of randomized samples, which included the particular string, and the number of samples used. The second approach is based on 'randomly selected samples' out of large databases and used as reference batches. Four reference batches were used for the present study: (i) an 18 406-sample batch of 5000 upstream elements each, taken from upstream regions of genes (in relation to the ATG translation start) in the human genome (http://genome-archive.cse.ucsc.edu/goldenPath/gh16/bigZips/); (ii) A 1000-sample-batch extracted from the above batch taken at an equidistance sample separation of 18; (iii) A 14 102-sample batch of 5000 upstream elements each, taken from upstream regions of genes (in relation to the ATG translation start) in the murine genome (http://genome-archive.cse.ucsc.edu/goldenPath/mm4/bigZips/) and (iv) A 1000-sample-batch extracted from the above batch (3) taken at an equidistance separation of 14. While both methods are available within the platform developed for this research, a conceptual difference between the two methods should be emphasized since they answer two different questions.

The procedure applied to the second approach consists of the following steps: (i) A candidate-sequence list is derived from the data associated with a compressed GCSI table (Figures 1C, 2 and 3). Note that each entry in the list consists of a complementary pair. (ii) Sequences of low interest to the present study, i.e. all mono- and dual-element repetitive strings, may be eliminated for computational efficiency. This optional step was made available in response to experience-based intuitive suggestions by molecular biologists. (iii) Each entry in the derived-sequence list is scanned against an available reference-sample batch, and each reference sample which had at list one match of either of the complementary pair is marked and (iv) The number of matching-reference samples is summarized and compared with the total number of samples in the respective reference batch, yielding an FPI indicator. Only those of the candidate sequences complying with a preset FPI Cap are retained.

The FPI-indicator table for the 9-sample MegaKP group is depicted in Figure 4 derived using an FPI scan batch of 1000 human samples (#2) and for an FPI cap of 10%. It was derived using a candidate list obtained from the data associated with the compressed-GCSI table shown in Figure 1C. The selection was made by automatically including in the list all entries of the columns above and including a pre-determined minimal-N value. This value is set by the 'Lower-Length Limit' parameter included in Figures 1C and 2 above. The table depicted in Figure 4 was obtained using a minimum-sequence length of 7 for the row corresponding to a gap of zero, and a minimum length of 8 for gaps of length between 1 and 10. These restrictions and the selection rules outlined above were applied for computational efficiency and for clarity of the displays.

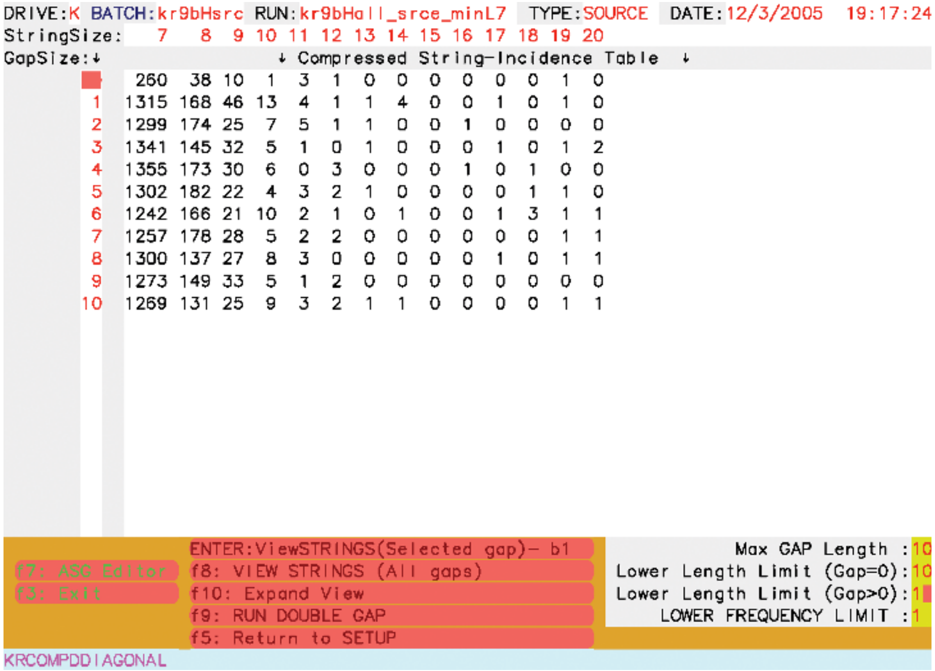
The number of members of the reference batch, which scanned positive for a specific sequence, is shown in the third column of Figure 4. The FPI indicator assigned to a particular sequence is determined by the ratio between this value and the number of samples in the specific reference batch employed. The fourth column in Figure 4 includes the total



```
DRIVE:K BATCH:kr9bMsrc RUN:kr9bMall_source      TYPE:SOURCE  DATE:12/4/2005   18:44:2
StringSize:    7    8   9  10  11  12
GapSize:↓             ↓ Compressed String-Incidence Table  ↓
            478   54   5   0   0   0
        1  2529  374  28   7   1   1
        2  2246  352  27   3   0   1
        3  2248  353  26   2   1   0
        4  2282  322  37   2   0   0
        5  2238  314  32   3   0   0
        6  2222  328  24   6   0   0
        7  2286  335  27   1   2   0
        8  2244  323  24   3   1   1
        9  2269  299  19   5   1   0
       10  2501  257  27   5   0   1

                                         Max GAP Length :10
                            Lower Length Limit (Gap=0):8
                            Lower Length Limit (Gap>0):9
                              LOWER FREQUENCY LIMIT :1
```

Figure 2. The Compressed-GCSI Table for the 4-sample murine component of the MegaKP group. This figure varies from Figure 1C in the numeric contents of the Table and in the values of the operational parameters in the bottom right column.



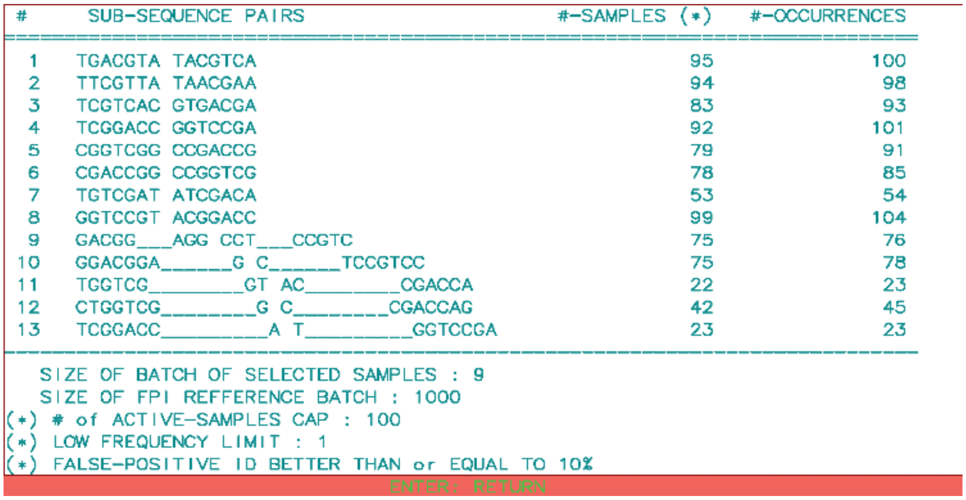
```
DRIVE:K BATCH:kr9bHsrc RUN:kr9bHall_srce_minL7  TYPE:SOURCE  DATE:12/3/2005   19:17:24
StringSize:    7    8   9  10  11  12  13  14  15  16  17  18  19  20
GapSize:↓             ↓ Compressed String-Incidence Table  ↓
            260   38  10   1   3   1   0   0   0   0   0   0   1   0
        1  1315  168  46  13   4   1   1   4   0   0   1   0   1   0
        2  1299  174  25   7   5   1   1   0   0   1   0   0   0   0
        3  1341  145  32   5   1   0   1   0   0   0   1   0   1   2
        4  1355  173  30   6   0   3   0   0   0   1   0   1   0   0
        5  1302  182  22   4   3   2   1   0   0   0   0   1   1   0
        6  1242  166  21  10   2   1   0   1   0   0   1   3   1   1
        7  1257  178  28   5   2   2   0   0   0   0   0   0   1   1
        8  1300  137  27   8   3   0   0   0   0   0   1   0   1   1
        9  1273  149  33   5   1   2   0   0   0   0   0   0   0   0
       10  1269  131  25   9   3   2   1   1   0   0   0   0   1   1

                                         Max GAP Length :10
                            Lower Length Limit (Gap=0):10
                            Lower Length Limit (Gap>0):1
                              LOWER FREQUENCY LIMIT :1
```

Figure 3. The compressed-GCSI table for the 5-sample human component of the MegaKP group. This figure varies from Figures 1C and 2 in the numeric contents of the table and in the values of the operational parameters in the bottom right column.

number of occurrences of the scanned sequence that had a score in the third column. A comparison between the numbers in the third and fourth columns indicates that only a fraction of these sequences had multiple occurrences in the same sample. It is based on the data associated with the compressed-GCSI table shown in Figure 1C, and on an FPI Cap of 10%. Inspection of the numbers appearing in the third column of Figure 4 indicates that setting up an FPI Cap of 2% would eliminate the entries in the table altogether.

An FPI-indicator table for the 4-samples murine component of the MegaKP group, as depicted in Figure 5, is based on the data associated with the compressed-GCSI table shown in Figure 2. It was derived using the FPI-reference batch of 1000 murine samples and an FPI Cap of



```
#      SUB-SEQUENCE PAIRS                        #-SAMPLES (*)   #-OCCURRENCES
==============================================================================

 1    TGACGTA TACGTCA                                    95            100
 2    TTCGTTA TAACGAA                                    94             98
 3    TCGTCAC GTGACGA                                    83             93
 4    TCGGACC GGTCCGA                                    92            101
 5    CGGTCGG CCGACCG                                    79             91
 6    CGACCGG CCGGTCG                                    78             85
 7    TGTCGAT ATCGACA                                    53             54
 8    GGTCCGT ACGGACC                                    99            104
 9    GACGG___AGG CCT___CCGTC                            75             76
10    GGACGGA______G C______TCCGTCC                      75             78
11    TGGTCG________GT AC________CGACCA                  22             23
12    CTGGTCG________G C________CGACCAG                  42             45
13    TCGGACC_________A T_________GGTCCGA                23             23
------------------------------------------------------------------------------
   SIZE OF BATCH OF SELECTED SAMPLES : 9
   SIZE OF FPI REFFERENCE BATCH : 1000
(*) # of ACTIVE-SAMPLES CAP : 100
(*) LOW FREQUENCY LIMIT : 1
(*) FALSE-POSITIVE ID BETTER THAN or EQUAL TO 10%
                                 ENTER: RETURN
```

Figure 4. The FPI indicator table for the 9-sample MegaKP group. This was analyzed with an FPI-reference batch of randomly selected 1000 human samples, each of 5000 bases, and an FPI Cap of 10%. The FPI values are determined by the ratio between the entrees in the third column and the FPI reference batch size.

1%. The 20 sequences so obtained, shown in Figure 5, are considered as unique to the murine component of the MegaKP group within the FPI limit of 1%. A comparison between columns 3 and 4 in the table indicates that all but one of these incidences has more than one occurrence per sample. In other words, raising the 'minimal frequency parameter' from one to two would eliminate 19 out of the 20 entries in this table.

A similar table is shown for comparison for the 5-sample human component of the MegaKP group. It was derived using a dual-stage FPI scan, involving (i) the application of the 1000-sample human-reference batch (#2) and applying an FPI Cap of 1%, resulting in a reduced 59-member list, as shown in Figure 6 and (ii) further refinement of this 59-sequence list is obtained by scanning it against the larger reference batch of 18 406 human samples (#1) and an FPI Cap of 0.01%.

An FPI-indicator table for the 5-samples human component of the MegaKP group, as depicted in Figure 6, is based on the data associated with the compressed-GCSI table shown in Figure 3. The 13 sequences so obtained, shown in Figure 6, are considered as unique to the human component of the MegaKP group, within the FPI limit of 0.01%. A comparison between columns 3 and 4 in the table indicates that none of these incidences has more than one occurrence per sample. In other words, raising the 'minimal frequency parameter' from one to two would eliminate all entries in this list.

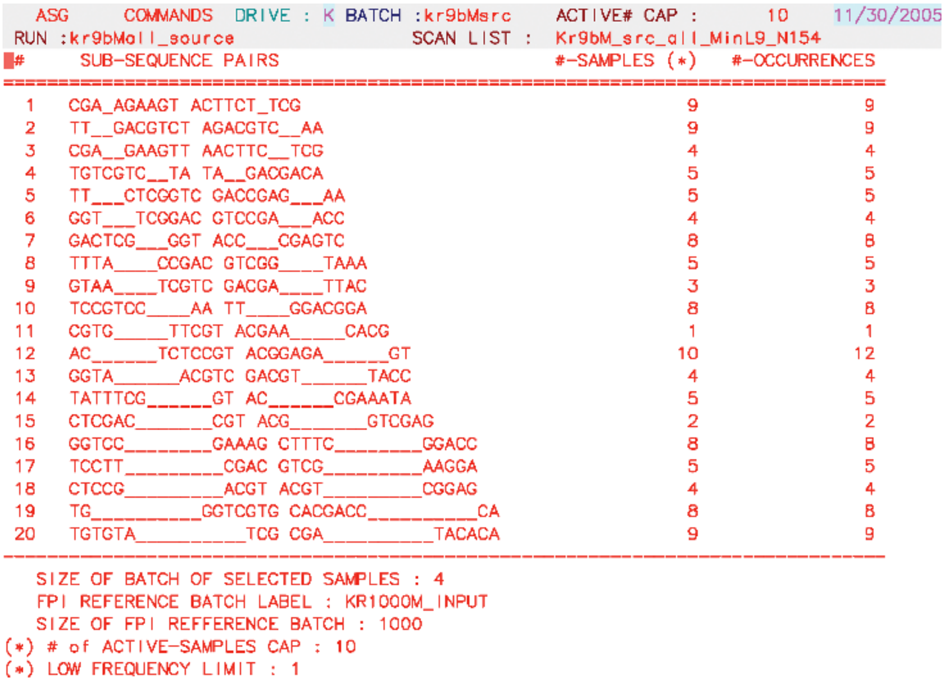
Sequence-location diagrams, multi-gap sequences

The sequence-location data, associated with each of the recorded incidences of the sequences listed in Figure 7, are summarized graphically in Figure 8A for the promoter of the human *hGPV* gene. The zero-reference point, corresponding to the location of the ATG switch, is on the right end of the Location axis, and the 'upstream' direction is pointing to the left. The sequential numbers in the left column refer to the sequences listed in Figure 7. However, the number of samples used in this display is too large for the sequential first column to be readable. For better readability the number of samples participating in such a display has to be reduced as is illustrated in Figure 8B. All the elements of a particular sequence (excluding the gap) are marked in Figure 8A and B by vertical-line segments. All sequences are represented in reverse order to comply with the 'upstream' convention, namely that the element nearest to the ATG switch is on the right of the segment and the furthest one is on the left.

Figures 7 and 8 are included to illustrate the capabilities of the computational platform in its present stage of development. The combined representation of all sequences, selected for *hGPV* as an example, is shown at the top row of Figure 8A, representing a multi-gap sequence, extending over a range of 258 elements, including multiple gaps. The range covered spans between the limits of −1565 and −1823 with respect to the ATG switch. The elements corresponding to the multiple gaps in the combined sequence are denoted by the character 'N'. Exploration of the relevant data for all other members of the human component of the MegaKP group yields a distribution of these conserved sequences over the upstream promoter regions (Supplementary Figures A-1–B-6). Hence, the megakaryocyte group of gene promoters contains a distinctive DNA signature that consists of a unique combination of sequences (as in Figure 8C) that are spaced to different degrees in the various megakaryocytic genes.

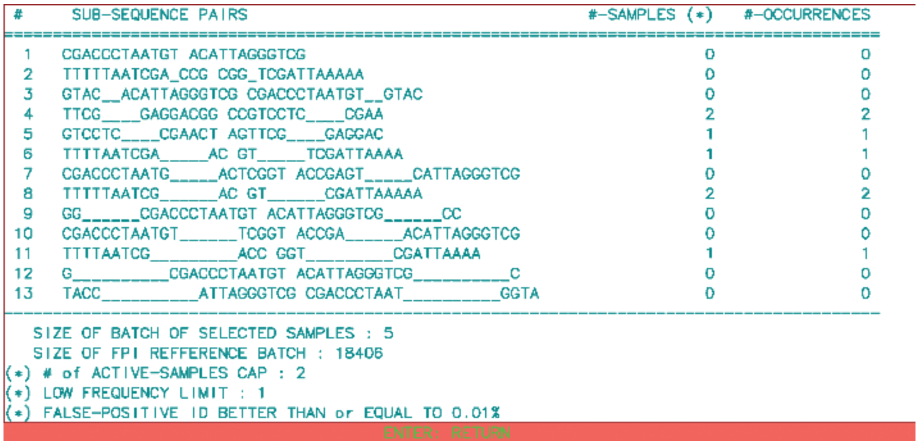
Identification of putative transcription factor binding sites within the newly identified sequences that are conserved in the MegaKP group

The cluster of conserved human sequences in all the megakaryocyte group of genes was further subjected to a search against a database of known transcription factor binding sites. Several known mammalian and non-vertebrate transcription factors (for which the mammalian homologs have not been characterized yet) were identified (Figure 8C). These include GATA-1, Ets binding sites and Pax binding



```
  ASG      COMMANDS  DRIVE : K BATCH :kr9bMsrc     ACTIVE# CAP :      10    11/30/2005
 RUN :kr9bMall_source                    SCAN LIST :  Kr9bM_src_all_MinL9_N154
 #     SUB-SEQUENCE PAIRS                            #-SAMPLES (*)    #-OCCURRENCES
==================================================================================
  1    CGA_AGAAGT ACTTCT_TCG                               9                9
  2    TT__GACGTCT AGACGTC__AA                             9                9
  3    CGA__GAAGTT AACTTC__TCG                             4                4
  4    TGTCGTC__TA TA__GACGACA                             5                5
  5    TT___CTCGGTC GACCGAG___AA                           5                5
  6    GGT___TCGGAC GTCCGA___ACC                           4                4
  7    GACTCG___GGT ACC___CGAGTC                           8                8
  8    TTTA____CCGAC GTCGG____TAAA                         5                5
  9    GTAA____TCGTC GACGA____TTAC                         3                3
 10    TCCGTCC____AA TT____GGACGGA                         8                8
 11    CGTG_____TTCGT ACGAA_____CACG                       1                1
 12    AC______TCTCCGT ACGGAGA______GT                    10               12
 13    GGTA______ACGTC GACGT______TACC                     4                4
 14    TATTTCG______GT AC______CGAAATA                     5                5
 15    CTCGAC_______CGT ACG_______GTCGAG                   2                2
 16    GGTCC________GAAAG CTTTC________GGACC               8                8
 17    TCCTT_________CGAC GTCG_________AAGGA               5                5
 18    CTCCG_________ACGT ACGT_________CGGAG               4                4
 19    TG__________GGTCGTG CACGACC__________CA             8                8
 20    TGTGTA__________TCG CGA__________TACACA             9                9
----------------------------------------------------------------------------------
   SIZE OF BATCH OF SELECTED SAMPLES : 4
   FPI REFERENCE BATCH LABEL : KR1000M_INPUT
   SIZE OF FPI REFFERENCE BATCH : 1000
(*) # of ACTIVE-SAMPLES CAP : 10
(*) LOW FREQUENCY LIMIT : 1
```

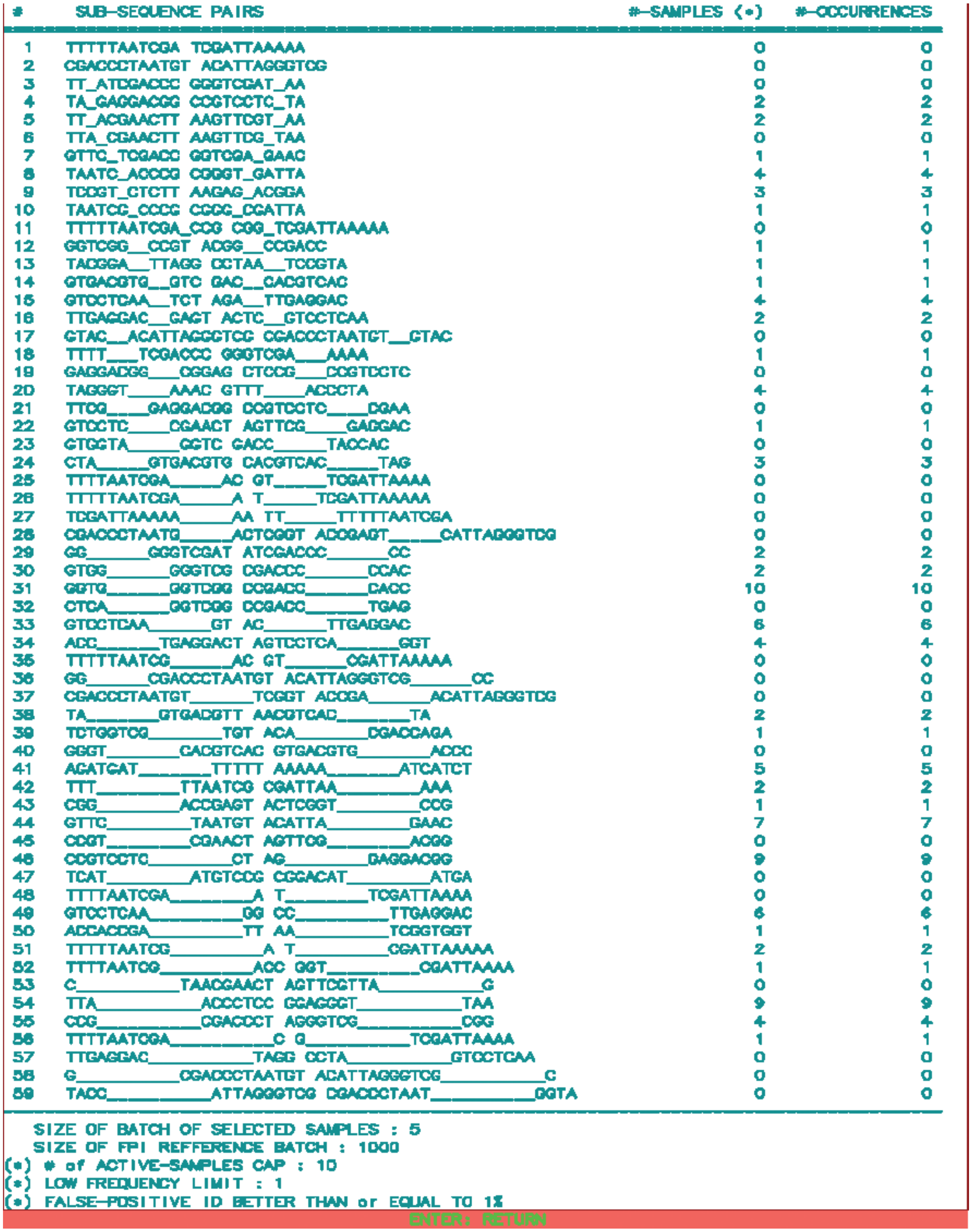
Figure 5. The False-Positive-Identification (FPI) indicator table for the 4-samples murine component of the MegaKP group, with an FPI-reference batch of 1000 murine samples and an FPI Cap of 1%.



```
 #     SUB-SEQUENCE PAIRS                                     #-SAMPLES (*)    #-OCCURRENCES
============================================================================================
  1    CGACCCTAATGT ACATTAGGGTCG                                     0                0
  2    TTTTTAATCGA_CCG CGG_TCGATTAAAAA                               0                0
  3    GTAC__ACATTAGGGTCG CGACCCTAATGT__GTAC                         0                0
  4    TTCG____GAGGACGG CCGTCCTC____CGAA                             2                2
  5    GTCCTC____CGAACT AGTTCG____GAGGAC                             1                1
  6    TTTTAATCGA_____AC GT_____TCGATTAAAA                           1                1
  7    CGACCCTAATG_____ACTCGGT ACCGAGT_____CATTAGGGTCG               0                0
  8    TTTTTAATCG______AC GT______CGATTAAAAA                         2                2
  9    GG______CGACCCTAATGT ACATTAGGGTCG______CC                     0                0
 10    CGACCCTAATGT______TCGGT ACCGA______ACATTAGGGTCG               0                0
 11    TTTTAATCG_________ACC GGT_________CGATTAAAA                   1                1
 12    G__________CGACCCTAATGT ACATTAGGGTCG__________C               0                0
 13    TACC__________ATTAGGGTCG CGACCCTAAT__________GGTA             0                0
--------------------------------------------------------------------------------------------
   SIZE OF BATCH OF SELECTED SAMPLES : 5
   SIZE OF FPI REFFERENCE BATCH : 18406
(*) # of ACTIVE-SAMPLES CAP : 2
(*) LOW FREQUENCY LIMIT : 1
(*) FALSE-POSITIVE ID BETTER THAN or EQUAL TO 0.01%
                                 ENTER: RETURN
```

Figure 6. A refined FPI scan for the human component of the MegaKP group. A pre-selected list was scan against the 1000 sample human-scan-batch to produce the list shown in Figure 7, which then was scanned versus the 18 406-sample human-scan-batch and an FPI Cap of 0.01%.

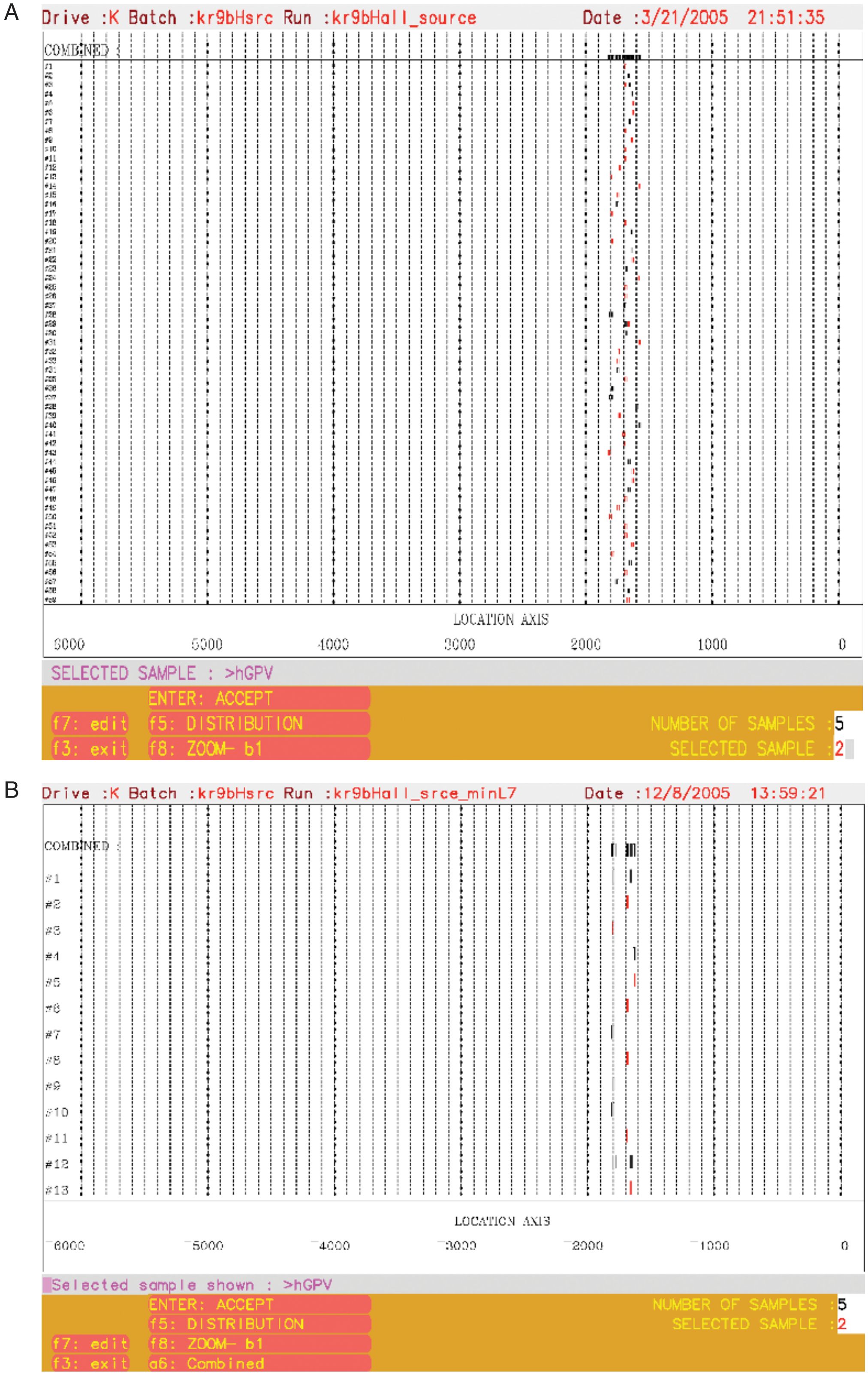
sites, which bind to Pax, a protein that also recognizes Ets binding sites and alter Ets transactivating properties (26–28); part of the mammalian Lymphocyte enriched DNA-binding protein (Lyf-1) binding site, the *Drosophila melangaster* Bicoid (Bcd) binding site, and the yeast *Saccharomyces Cerevisiae* heat shock factor, HSF, recognition site (29–31). The Ets family of transcription factors has a DNA-binding domain in common that binds a core GGA(A/T) DNA sequence. Ets-1 was first described as the cellular homolog of v-Ets. Ets-2 was subsequently described as a closely related protein that contains the highly conserved Ets DNA-binding domain. Our earlier studies and those of other indicated that Ets as well as GATA binding sites are important for activation of gene promoters that are uniquely expressed in megakaryocytes (please refer to the Introduction). Moreover, it has been reported that the 5′ non-coding region in the human GPV contains closely spaced GATA and Ets sites that are functionally active (32). The same sites were identified in our computational quest as shown in Figure 8C. Hence, our current findings highlight the value of the search method employed here to identify regulatory regions within any other groups of genes that is uniquely expressed in one lineage. In addition, the search identified new conserved sequences surrounding the Ets and GATA sites, and these sequences are unique to the megakaryocyte group. The only variable found between these clusters in all



```
#     SUB-SEQUENCE PAIRS                                               #-SAMPLES (*)   #-OCCURRENCES
====================================================================================================
 1    TTTTTAATCGA TCGATTAAAAA                                                  0               0
 2    CGACCCTAATGT ACATTAGGGTCG                                                0               0
 3    TT_ATCGACCC GGGTCGAT_AA                                                  0               0
 4    TA_GAGGACGG CCGTCCTC_TA                                                  2               2
 5    TT_ACGAACTT AAGTTCGT_AA                                                  2               2
 6    TTA_CGAACTT AAGTTCG_TAA                                                  0               0
 7    GTTC_TCGACC GGTCGA_GAAC                                                  1               1
 8    TAATC_ACCCG CGGGT_GATTA                                                  4               4
 9    TCCGT_CTCTT AAGAG_ACGGA                                                  3               3
10    TAATCG_CCCG CGGG_CGATTA                                                  1               1
11    TTTTTAATCGA_CCG CGG_TCGATTAAAAA                                          0               0
12    GGTCGG__CCGT ACGG__CCGACC                                                1               1
13    TACGGA__TTAGG CCTAA__TCCGTA                                              1               1
14    GTGACGTG__GTC GAC__CACGTCAC                                              1               1
15    GTCCTCAA__TCT AGA__TTGAGGAC                                              4               4
16    TTGAGGAC__GAGT ACTC__GTCCTCAA                                            2               2
17    GTAC__ACATTAGGGTCG CGACCCTAATGT__GTAC                                    0               0
18    TTTT___TCGACCC GGGTCGA___AAAA                                            1               1
19    GAGGACGG___CGGAG CTCCG___CCGTCCTC                                        0               0
20    TAGGGT____AAAC GTTT____ACCCTA                                            4               4
21    TTCG____GAGGACGG CCGTCCTC____CGAA                                        0               0
22    GTCCTC____CGAACT AGTTCG____GAGGAC                                        1               1
23    GTGGTA_____GGTC GACC_____TACCAC                                          0               0
24    CTA_____GTGACGTG CACGTCAC_____TAG                                        3               3
25    TTTTAATCGA_____AC GT_____TCGATTAAAA                                      0               0
26    TTTTTAATCGA_____A T_____TCGATTAAAAA                                      0               0
27    TCGATTAAAAA_____AA TT_____TTTTTAATCGA                                    0               0
28    CGACCCTAATG_____ACTCGGT ACCGAGT_____CATTAGGGTCG                          0               0
29    GG______GGGTCGAT ATCGACCC______CC                                        2               2
30    GTGG______GGGTCG CGACCC______CCAC                                        2               2
31    GGTG______GGTCGG CCGACC______CACC                                       10              10
32    CTCA______GGTCGG CCGACC______TGAG                                        0               0
33    GTCCTCAA______GT AC______TTGAGGAC                                        6               6
34    ACC______TGAGGACT AGTCCTCA______GGT                                      4               4
35    TTTTTAATCG______AC GT______CGATTAAAAA                                    0               0
36    GG______CGACCCTAATGT ACATTAGGGTCG______CC                                0               0
37    CGACCCTAATGT______TCGGT ACCGA______ACATTAGGGTCG                          0               0
38    TA_______GTGACGTT AACGTCAC_______TA                                      2               2
39    TCTGGTCG_______TGT ACA_______CGACCAGA                                    1               1
40    GGGT_______CACGTCAC GTGACGTG_______ACCC                                  0               0
41    AGATCAT_______TTTTT AAAAA_______ATCATCT                                  5               5
42    TTT________TTAATCG CGATTAA________AAA                                    2               2
43    CGG________ACCGAGT ACTCGGT________CCG                                    1               1
44    GTTC________TAATGT ACATTA________GAAC                                    7               7
45    CCGT________CGAACT AGTTCG________ACGG                                    0               0
46    CCGTCCTC________CT AG________GAGGACGG                                    9               9
47    TCAT________ATGTCCG CGGACAT________ATGA                                  0               0
48    TTTTAATCGA________A T________TCGATTAAAA                                  0               0
49    GTCCTCAA_________GG CC_________TTGAGGAC                                  6               6
50    ACCACCGA_________TT AA_________TCGGTGGT                                  1               1
51    TTTTTAATCG_________A T_________CGATTAAAAA                                2               2
52    TTTTAATCG_________ACC GGT_________CGATTAAAA                              1               1
53    C__________TAACGAACT AGTTCGTTA__________G                                0               0
54    TTA__________ACCCTCC GGAGGGT__________TAA                                9               9
55    CCG__________CGACCCT AGGGTCG__________CGG                                4               4
56    TTTTAATCGA__________C G__________TCGATTAAAA                              1               1
57    TTGAGGAC__________TAGG CCTA__________GTCCTCAA                            0               0
58    G__________CGACCCTAATGT ACATTAGGGTCG__________C                          0               0
59    TACC__________ATTAGGGTCG CGACCCTAAT__________GGTA                        0               0
----------------------------------------------------------------------------------------------------
   SIZE OF BATCH OF SELECTED SAMPLES : 5
   SIZE OF FPI REFFERENCE BATCH : 1000
(*) # of ACTIVE-SAMPLES CAP : 10
(*) LOW FREQUENCY LIMIT : 1
(*) FALSE-POSITIVE ID BETTER THAN or EQUAL TO 1%
                                   ENTER: RETURN
```

Figure 7. An FPI Scan of the 5-sample human component of the MegaKP group. A pre-selected list was scanned against the 1000-sample human scan batch and an FPI Cap of 1%. Note that the majority of the entries in the third column of the table are zeros, indicating a better than 0.1% FPI values, which prompted the more refined scan shown in Figure 6.

the genes analyzed was the spacing between the conserved sequences (indicated in Figure 8C by N). In some gene promoters, such as GPV, the clusters are in close proximity, while in other genes the distances are greater (see Supplementary Material online). While Ets and GATA sequences were already confirmed binding sites in genes expressed in megakaryocytes, the sequence GCTACTTGNGAG GCCGAGGCAGGAGAATTGCTTGAA (Figure 8C) is the longest stretch that is perfectly identical in all megakaryocytic genes and it has never been tested for protein binding. Here, we used EMSA (Figure 9) to demonstrate clear binding of megakaryocytic nuclear proteins to this oligomer (oligo) as well as to a similar oligo in which the putative yeast heat shock factor binding site was mutated (putative HSF site shown in Figure 8; AGAAT). As also shown in Figure 9, no significant binding was detected in nuclear extracts prepared from epithelial cells (which was previously tested for binding to a consensus Sp1 site as detailed in the legend to Figure 9). Two complexes were noted in nuclear megakaryocytic cells incubated with the



A

Drive :K Batch :kr9bHsrc Run :kr9bHall_source Date :3/21/2005 21:51:35

SELECTED SAMPLE : >hGPV

NUMBER OF SAMPLES :5

SELECTED SAMPLE :2

B

Drive :K Batch :kr9bHsrc Run :kr9bHall_srce_minL7 Date :12/8/2005 13:59:21

Selected sample shown : >hGPV

NUMBER OF SAMPLES :5

SELECTED SAMPLE :2

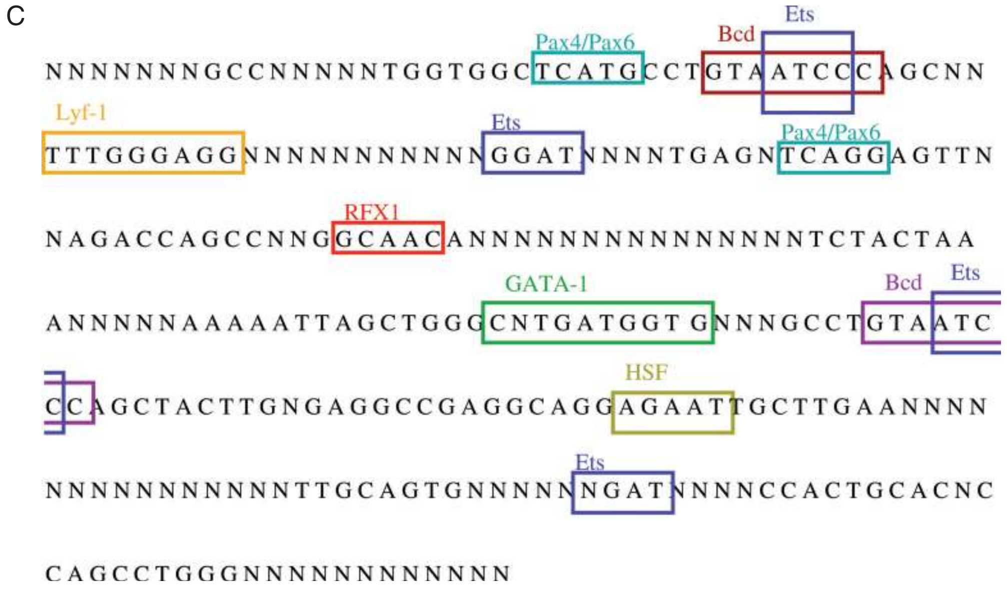

C
Pax4/Pax6
Bcd
Ets
NNNNNNNGCCNNNNNNTGGTGGCTCATGCCTGTAATCCCAGCNN
Lyf-1
Ets
Pax4/Pax6
TTTGGGAGGNNNNNNNNNNNGGATNNNNTGAGNTCAGGAGTTN
RFX1
NAGACCAGCCNNGGCAACANNNNNNNNNNNNNNNNTCTACTAA
GATA-1
Bcd
Ets
ANNNNNAAAAATTAGCTGGGCNTGATGGT GNNNGCCTGTAATC
HSF
CCAGCTACTTGNGAGGCCGAGGCAGGAGAATTGCTTGAANNNN
Ets
NNNNNNNNNNNTTGCAGTGNNNNNNGATNNNNCCACTGCACNC
CAGCCTGGGNNNNNNNNNNNN

Figure 8. (**A**) A location diagram for the hGPV sample and the 59-member list shown in Figure 7, based on an FPI Cap of 1%. The 59 individual sequences are referred to by the index vector shown in the right column. The location of the ATG switch is at the right of the location scale (bottom), and the 'upstream' direction is pointing to the left. Note the overlapping location of all these sequences, as depicted on the 'COMBINED' line at the top of the picture. This multi-gap sequence is 258-element long, as is detailed in panel (C). (**B**) A location diagram for the hGPV sample and the 13-member list shown in Figure 6, based on an FPI Cap of 0.01%. (**C**) The cluster of combined-unique sequences shown in (A) for the promoter of the hGPV gene and identification of putative transcription factor binding sites. The number of participating segments is 59 and the FPI Cap used is 1%. The gaps are indicated by 'N' elements. The leading non-N element (first 'G' in left of top row) is located at position -1830 upstream from the ATG switch, and the last non-N element (rightmost 'G' in bottom row) is located at position −1551 upstream from the ATG switch. This 258-element, multi-gap sequence contains sequences that are uniquely common to all the megakaryocyte expressing genes. This stretch was also searched for transcription factor binding sites using the web-based TFSEARCH program, developed by Yutaka Akiyama (http://www.cbrc.jp/research/db/TFSEARCH.html) searching in all matrices (vertebrate, arthropod, plant and yeast) with a homology threshold of 90% or more. The program searches highly correlated sequence fragments against TFMATRIX transcription factor binding site profile database in the 'TRANSFAC' databases developed by Heinemeyer *et al*. (33). Core nucleotides of Ets transcription factors were also identified using the 'find' function on MacVector6.5.3™ (Accelrys, San Diego, CA).

wild-type oligo and these were competed by the cold oligo or the mutated one. This suggested that the main binding is not via this site (AGAAT), but rather via neighboring sequences. Of note, upon mutation of the AGAAT site, binding of an additional high molecular weight complex became possible, which suggests that this domain represses binding of an additional protein(s). The identification of the factors that bind to the novel sequences identified in this study will be a focus of future investigation.

Analyses of all genes, pursued as shown for GPV (Figure 8), are illustrated in Supplementary Figures A-1–B-6. For the mouse PF4 promoter, clusters of conserved sequences are located within the first Kilobase, but also upstream to it (Figure A-2). Noted were conserved Ets binding sites which correspond to those published as functional, in addition to the conserved sequences, for which mammalian transcription factor binding sites have not been identified yet (6). In the human PF4 promoter all the conserved sequences (the same as in Figure 8C for hGPV) are concentrated around −2.5 kb (Figure B-2). Another example is the GPIIb promoter. In this case, functional Ets and GATA binding sites were described to reside within −60–32 bases (with zero denoting the transcriptional start) (7). In our search, the surrounding sequences of these sites were not identified as unique to megakaryocytes (Supplementary Figure A-4). The same conserved and unique to this lineage clusters, which were identified and described for GPV (the DNA signature shown in Figure 8C), were also found in GPIIb, but they were spread with gaps >1–4 kb in the case of mGPIIb and >1.7–5 kb in the case of hGPIIb.

DISCUSSION

The merit of a customized system is in its ability to fulfill the specific objectives of a particular research project. Two web-based systems for customized application of Bioinformatics have been described in the literature (23); and the *Taverna* system http://www.biomoby.org; http://taverna.sourceforge.net/. The methodology outlined in the present report was developed to respond to specific research requirements, providing for the development and application of customized-tool kits. Our choice of the customization approach may reflect on the lack of sufficient expertise and/or resources to exploit the potential of available web-based methodologies, while having access to expertise in the development of customized-computer applications.

The objective of the present study was to identify unique sequences, and clusters of sequences, common to a 9-member MegaKP group, and to explore the associated sequence-location information, which may shed light on the role played by such sequences in regulating lineage-specific expressions.

Our methodology is outlined in the context of some of the results obtained for the megakaryocytic gene promoter group. The model employed has evolved through consecutive

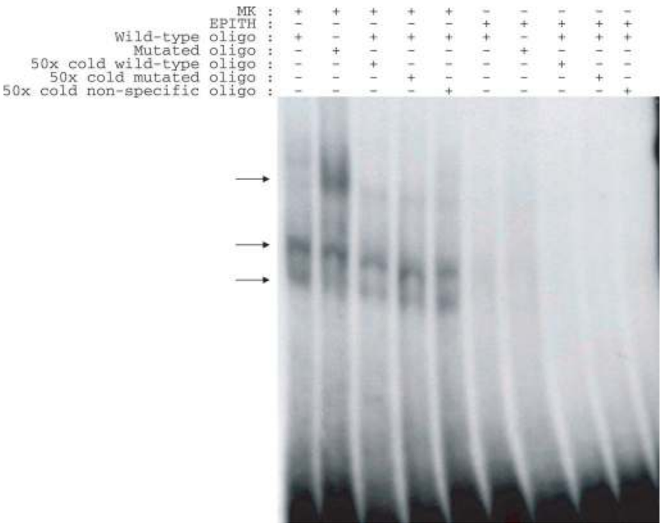

MK : + + + + + − − − − −
EPITH : − − − − − + + + + +
Wild-type oligo : + − + + + + − + + +
Mutated oligo : − + − − − − + − − −
50x cold wild-type oligo : − − + − − − − + − −
50x cold mutated oligo : − − − + − − − − + −
50x cold non-specific oligo : − − − − + − − − − +

Figure 9. Megakaryocyte nuclear protein binding to a sequence that is conserved in genes expressed in megakaryocytes. Electromobility shift assay was employed to examine protein binding (denoted by an arrow) using labeled wild-type: 5′-AGCTACTTGNGAGGCCGAGGCAG***GAGAAT***TGCTTGAA-3′ or mutated oligomers (oligo) 5′-AGCTACTTGNGAGGCCGAGGCAG***CATGTA***TGCTTGAA-3′ as probes (sequences derived from Figure 8). Nuclear extracts were prepared from megakaryocytic cells (designated as MK) and as a control also from epithelial cells, NMuMG (designated as EPITH), as described under Materials and Methods. No significant binding was noted in the epithelial cells compared with megakaryocytes, while the binding to a control Sp1 oligo (see Materials and Methods) was comparable in both protein extracts (data not shown). Data shown are representative of four experiments with two different batches of nuclear extracts.

simple conceptual steps: (i) a simple model was initially implemented, consisting of a search for contiguous segments common to the entire group of sample; (ii) an extension of the searches to include single-gap sequences; (iii) an extension to searches for the common presence of variable-single-gap sequences; (iv) identification of common multi-gap sequences; (v) further extension of segment alignment through the addition/deletion of a small number of elements and (vi) a contemplated extension to include a search for ‘closely placed active sites’, as reported by Lepage *et al*., allowing for a variable separation between the sites (32).

The results presented in this report indicate that, when applied to the full MegaKP group, including both human and murine samples, the observed common sequences correspond to FPI indicators of the order of 10%. However, splitting the study into the murine and human subgroups, results in an observed FPI-indicator Cap of 1% for the murine sub-group, while the human sub-group yielded a list characterized by FPI-indicator Cap as low as 0.01%. These groups of sequences, which are primarily present in genes that are expressed selectively in the megakaryocytic lineage, are high candidates for regulatory domains. Hence, our study suggests the existence of a DNA signature that consists of a unique combination of binding sites in the megakaryocytic, and provides a list of candidate sequences for future mutation studies to examine their functional significance individually and in combination. The list of combination of sequences to potentially mutate for promoter activity studies is large and its availability to the research community allows full examination of these regions.

It should be emphasized that the development of a PC-based customized platform does not preclude complementary, web-based studies. Of most interest, when the newly identified human sequences were subjected to a web-based search against a database of known binding sites for transcription factors, several putative Ets binding sites were identified. Additional clusters of sequences were identified, for which binding factors have not been identified yet, e.g. the mammalian homolog of the yeast HSF (Figure 8C). When the murine conserved sequences in the megakaryocyte group (Figure 7) were similarly searched against the same databases, the Ets binding core was identified in sequence # 17 and HSF putative sites were recognized in sequences # 1 and 14 (Figure 7). These findings further validate our unbiased search method for conserved regulatory elements that might be significant within a related group of genes, as Ets biding sites have been previously described to regulate genes specifically expressed in megakaryocytes (see Introduction). This search could be applied in the future to identify potential regulatory elements in other groups of genes related by virtue of their unique expression in a specific lineage.

The power of a methodology is manifested in having the flexibility to recognize and pursue new research venues based on the unique sequences identified. The mechanism and implication of such a characteristic difference between the human and mouse groups are not clear at the moment. Nevertheless, it raises an interesting possibility that there exists an intrinsic difference in the organization of upstream regulatory regions between the two mammalian genomes and that the human megakaryocyte-specific genes are commonly requested by a direct, or indirect, activation of the uniquely shared sequences identified in this study.

The usefulness of a methodology is demonstrated through the achievement and publication of meaningful scientific results, which is the main purpose of the present paper. The validity of its application is proven via the identification of clusters of common sequences in a group of genes related by their specific expression in the lineage. These sequences have been identified here for the first time.

Comment: The evolving software is available from the authors on request.

SUPPLEMENTARY DATA

Supplementary data are available at NAR online.

ACKNOWLEDGEMENTS

K.R. is an Established Investigator with the American Heart Association. This work was partially supported by an NCI core grant to the Cancer Center, Boston University School of Medicine. Funding to pay the Open Access publication charges for this article was provided by an NCL core grant to the Cancer Center, Boston University School of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

2. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
3. Ravid,K., Lu,J., Zimmet,J.M. and Jones,M.R. (2002) Roads to polyploidy: the megakaryocyte example. *J. Cell Physiol.*, **190**, 7–20.
4. Shivdasani,R.A. (2001) Molecular and transcriptional regulation of megakaryocyte differentiation. *Stem Cells*, **19**, 397–407.
5. Kaluzhny,Y., Poncz,M. and Ravid,K. (2001) Transcription Factors involved in lineage-specific gene expression during megakaryopoiesis. In Ravid,K. and Licht,J. (eds.), *Transcription Factors: Normal and Malignant Development of Blood Cells*. Wiley-Liss, New York, NY, pp. 31–50.
6. Lu,J., Pazin,M.J. and Ravid,K. (2004) Properties of ets-1 binding to chromatin and its effect on platelet factor 4 gene expression. *Mol. Cell Biol*, **24**, 428–441.
7. Wang,X., Crispino,J.D., Letting,D.L., Nakazawa,M., Poncz,M. and Blobel,G.A. (2002) Control of megakaryocyte-specific gene expression by GATA-1 and FOG-1: role of Ets transcription factors. *Embo. J*, **21**, 5225–5234.
8. Azorsa,D.O., Moog,S., Ravanat,C., Schuhler,S., Follea,G., Cazenave,J.P. and Lanza,F. (1999) Measurement of GPV released by activated platelets using a sensitive immunocapture ELISA—its use to follow platelet storage in transfusion. *Thromb. Haemost*, **81**, 131–138.
9. Ravanat,C., Morales,M., Azorsa,D.O., Moog,S., Schuhler,S., Grunert,P., Loew,D., Van Dorsselaer,A., Cazenave,J.P. and Lanza,F. (1997) Gene cloning of rat and mouse platelet glycoprotein V: identification of megakaryocyte-specific promoters and demonstration of functional thrombin cleavage. *Blood*, **89**, 3253–3262.
10. Ravanat,C., Freund,M., Mangin,P., Azorsa,D.O., Schwartz,C., Moog,S., Schuhler,S., Dambach,J., Cazenave,J.P. and Lanza,F. (2000) GPV is a marker of *in vivo* platelet activation—study in a rat thrombosis model. *Thromb. Haemost*, **83**, 327–333.
11. Holmes,M.L., Bartle,N., Eisbacher,M. and Chong,B.H. (2002) Cloning and analysis of the thrombopoietin-induced megakaryocyte-specific glycoprotein VI promoter and its regulation by GATA-1, Fli-1, and Sp1. *J. Biol. Chem.*, **277**, 48333–48341.
12. Deveaux,S., Filipe,A., Lemarchandel,V., Ghysdael,J., Romeo,P.H. and Mignotte,V. (1996) Analysis of the thrombopoietin receptor (MPL) promoter implicates GATA and Ets proteins in the coregulation of megakaryocyte-specific genes. *Blood*, **87**, 4678–4685.
13. Mignotte,V., Deveaux,S. and Filipe,A. (1996) Transcriptional regulation in megakaryocytes: the thrombopoietin receptor gene as a model. *Stem Cells*, **14** (Suppl. 1), 232–239.
14. Friedman,A.D. (2002) Transcriptional regulation of granulocyte and monocyte development. *Oncogene*, **21**, 3377–3390.
15. Lesk,A.M. (2002) *Introduction to Bioinformatics*. Oxford University Press.
16. Orengo,C.A. (2003) Sequence comparison method. In Orengo,C.A., Jones,D.T. and Thornton,J.M. (eds), *Bioinformatics: Genes, Proteins and Computers, Chapter 3*. BIOS Scientific Publishers Ltd. Oxford, UK, pp. 29–48.
17. In Orengo,C., Jones,D.T. and Thornton,J.A. (eds), *Bioinformatics: Genes, Proteins and Computers*. BIOS Scientific Publishers Ltd, Oxford, UK.
18. Eidhammer,I., Jonassen,I. and Taylor,W.R. (2004) *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*. John Wile & Sons, Ltd, Chichester, England.
19. Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1995). Approaches to the automatic discovery of patterns in biosequences. Reports in Informatics, Report No. 113. University of Bergen, Bergen, Norway, pp. 21–22.
20. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
21. Thompson,J.D., Plewniak,F. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
22. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
23. Buisine,N. and Chalmers,R. (2004) yBlast, a graphical front end for the standalone BLAST suite. *Biotechniques*, **37**, 987–989.
24. Hazony,Y. (1996) System Generator for producing manufacturing application. *IBM Syst J.*, **35**, 69–93.
25. Hazony,Y. (2002) Rule-based process-cell, monitor and control design. *Robotics Comp. Int. Manuf.*, **18**, 105–123.
26. Merika,M. and Orkin,S.H. (1993) DNA-binding specificity of GATA family transcription factors. *Mol. Cell Biol.*, **13**, 3999–4010.
27. Wasylyk,B., Hahn,S.L. and Giovane,A. (1993) The Ets family of transcription factors. *Eur. J. Biochem.*, **211**, 7–18.
28. Plaza,S., Grevin,D., MacLeod,K., Stehelin,D. and Saule,S. (1994) Pax-QNR/Pax-6, a paired- and homeobox-containing protein, recognizes Ets binding sites and can alter the transactivating properties of Ets transcription factors. *Gene. Expr.*, **4**, 43–52.
29. Lo,K., Landau,N.R. and Smale,S.T. (1991) LyF-1, a transcriptional regulator that interacts with a novel class of promoters for lymphocyte-specific genes. *Mol. Cell Biol.*, **11**, 5229–5243.
30. Hoch,M., Seifert,E. and Jackle,H. (1991) Gene expression mediated by cis-acting sequences of the *Kruppel* gene in response to the Drosophila morphogens bicoid and hunchback. *Embo J.*, **10**, 2267–2278.
31. Fernandes,M., Xiao,H. and Lis,J.T. (1994) Fine structure analyses of the Drosophila and Saccharomyces heat shock factor—heat shock element interactions. *Nucleic Acids Res.*, **22**, 167–173.
32. Lepage,A., Uzan,G., Touche,N., Morales,M., Cazenave,J.P., Lanza,F. and de La Salle,C. (1999) Functional characterization of the human platelet glycoprotein V gene promoter: a specific marker of late megakaryocytic differentiation. *Blood*, **94**, 3366–3380.
33. Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A. *et al.* (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.