Published online 5 December 2008

Nucleic Acids Research, 2009, Vol. 37, No. 2 ***533–549***
doi:10.1093/nar/gkn948

Motif discovery in promoters of genes co-localized and co-expressed during myeloid cells differentiation

Alessandro Coppe[1], Francesco Ferrari[1], Andrea Bisognin[1], Gian Antonio Danieli[1], Sergio Ferrari[2], Silvio Bicciato[2] and Stefania Bortoluzzi[1,*]

[1]University of Padova, Department of Biology, Via G. Colombo 3, 35121, Padova and [2]University of Modena and Reggio Emilia, Department of Biomedical Sciences, via Campi 287, 41100, Modena, Italy

Received October 15, 2008; Revised and Accepted November 7, 2008

ABSTRACT

Genes co-expressed may be under similar promoter-based and/or position-based regulation. Although data on expression, position and function of human genes are available, their true integration still represents a challenge for computational biology, hampering the identification of regulatory mechanisms. We carried out an integrative analysis of genomic position, functional annotation and promoters of genes expressed in myeloid cells. Promoter analysis was conducted by a novel multi-step method for discovering putative regulatory elements, i.e. over-represented motifs, in a selected set of promoters, as compared with a background model. The combination of transcriptional, structural and functional data allowed the identification of sets of promoters pertaining to groups of genes co-expressed and co-localized in regions of the human genome. The application of motif discovery to 26 groups of genes co-expressed in myeloid cells differentiation and co-localized in the genome showed that there are more over-represented motifs in promoters of co-expressed and co-localized genes than in promoters of simply co-expressed genes (CEG). Motifs, which are similar to the binding sequences of known transcription factors, non-uniformly distributed along promoter sequences and/or occurring in highly co-expressed subset of genes were identified. Co-expressed and co-localized gene sets were grouped in two co-expressed genomic meta-regions, putatively representing functional domains of a high-level expression regulation.

INTRODUCTION

Co-expression is essential to sustain normal function of cells and tissues. Genes can be co-expressed because they are co-regulated, have similar promoters, share combinations of functional regulatory sequence motifs binding transcription factors (TF), and/or they are co-localized. Genes could be co-localized because they are close to each other on a linear chromosome, thus being under the influence of the same regulators (e.g. enhancers acting locally on a limited chromosomal region) and/or under the effect of local control, possibly based on specific chromatin modifications. Genes could be considered co-localized also because they are preferentially located in a given functional district of the three-dimensional interphase nucleus, i.e. chromosome territories, thus being exposed to a particularly concentrated mixture of regulatory proteins (1). On the other hand, part of co-expressed genes (CEG) are functionally related and tend to be under the control of similar gene circuits (2–5). Thus, the integrated study of co-expression, co-regulation, co-localization and functional similarity may help in understanding basic and general rules governing genomic expression and may allow identifying mechanisms and specific switches of expression regulation in considered biological processes (6,7).

Haematopoiesis is an ideal biological model for studying regulation of gene expression in cellular differentiation since it represents a plastic process where multipotent stem cells gradually limit their differentiation potential, generating different precursor cells that finally evolve in

*To whom correspondence should be addressed. Tel: +39 49 827 6502; Fax: +39 49 827 6209; Email: stefibo@bio.unipd.it

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

eight distinct types of terminally differentiated cells. Myelopoiesis is the part of haematopoiesis leading to differentiation of myeloid cell lineages (erythrocytes, megakaryocytes, granulocytes and mono/macrophages). In a recent study on myelopoiesis (8), the analysis of the correlations between expression patterns of genes, their biological roles and their physical position in the human genome led to the identification of (i) chromatin domains containing clusters of genes relevant for specific myeloid lineages and (ii) chromosomal regions with low transcriptional activity that partially overlap genomic clusters related to non-haematopoietic functions.

In this article, we address the study on gene expression regulation by integrating analyses performed at multiple levels. Specifically, the analysis of gene co-expression and co-localization is integrated with the analysis of promoter sequences.

The search for DNA motifs in gene promoters is a challenging problem that has been of longstanding interest in computational biology (9–11). Numerous pattern discovery programs, which are based on different algorithms and methodologies, have been proposed and coded in stand-alone or web applications [e.g. MEME (12), GibbsSampler (13), Weeder (14), WordSpy (15), COOP (16), MOST (17) and RSAT (18); for a critical review see Tompa *et al*. (19)].

We introduce a novel methodology for the identification of putatively relevant motifs in gene promoters. The method is composed of a cascade of non-standard analytical procedures, and it is conducted on top of a MySQL database organizing different levels of interconnected data. In particular, this approach allows:

(i) Finding, in a selected set of promoter sequences, motifs over-represented as compared with a biologically meaningful background model.
(ii) Analyzing promoter sequences by regions. Since several functional motifs have strong position-related prevalence (20,21), accounting for positional distribution in promoter sequences could refine the search for biologically meaningful motifs.
(iii) Integrating motif over-representation with additional, biologically relevant properties of motifs, such as similarity with known-functional sequences.

The application of this computational approach to the analysis of expression regulation during myeloid cell differentiation allowed identifying a greater number of significantly over-represented motifs in the promoters of co-expressed and co-localized genes when compared with those of simply co-expressed genes. All results are stored in a dedicated website where details of promoter sequences analyses can be browsed, along with gene transcriptional and functional characteristics (http://compgen.bio.unipd.it/MoDi/).

MATERIALS AND METHODS

Myelopoiesis gene expression data

The myelopoiesis data set consists of gene expression data for 24 samples generated from 8 different types of myeloid cells. Specifically, RNA from CD34 + , haematopoietic stem/progenitor cells (HSC), myeloid precursors (myeloblasts, monoblasts, erythroblasts and megakaryoblasts) and terminally differentiated cells (monocytes, neutrophils and eosinophils) was analysed using Affymetrix GeneChip HG-U133A, as described in (8). Robust multi-array average (RMA) procedure was applied to raw signals (i.e. CEL files) in order to background adjust and normalize microarray intensities and to generate gene expression values. Raw data of 24 considered samples are publicly available as a GEO series (GSE12837). Probe–gene relationships were obtained using GeneAnnot-based custom Chip Definition Files with a total of 11 446 unique custom probesets (gahgu133a_1.1.1.cdf; www.xlab.unimo.it/GA_CDF/) (22). For each of eight cell types, RMA expression values in corresponding replicate samples were averaged, and then, the data matrix was standardized per gene, obtaining a vector of eight expression values for each gene.

Reference set of human genes: genomic localization and promoter sequences

A reference set of human genes was collected by selecting all EntrezGene human entries corresponding to trustable RefSeq nuclear genes with Known, Reviewed or Validated Status and excluding Mitochondria, Plasmids, Plastids and Pseudogenes. Each EntrezGene ID was matched to the corresponding mRNA and EST sequences of UCSC Genome Browser. The genomic region including all of these sequences was selected as the reference gene locus and used to predict the exact Transcription Start Site (TSS) position. Gene loci whose genomic position could not be unambiguously determined according to this procedure were discarded. Then, the promoter sequences were retrieved, each spanning from 1000 bp upstream to 100 bp downstream of the predicted TSS (-1000, $+100$). These sequences constituted the reference set of gene promoters (REFGP).

QT clustering of myelopoiesis expression data: groups of Co-Expressed Genes (CEG)

In order to identify sets of co-expressed human genes (CEG) along myelopoiesis, we first selected genes varying in tissue/cell type-dependent manner, and then quality threshold (QT) clustering was used to group CEG. The Shannon entropy (H) was adopted as measure of expression variability (23) and used to rank and filter genes. The genes selected as variably expressed in myeloid cells were then grouped by the similarity of expression profile. Spearman correlation was adopted as a similarity measure and cluster analysis was performed by the QT clustering (24) of TMEV software (http://www.tm4.org/mev.html). QT clustering was adopted since it allows setting *a priori* thresholds for cluster quality, such as minimum values for the correlation between gene pairs within the cluster and for the number of genes per cluster.

Local Correlation Score (LCS): Co-Expressed chromosomal Regions (CER)

We searched for chromosomal regions including CEG, which could represent functional domains of higher-level

gene expression regulation. The search for co-expressed chromosomal regions (CER) was based on the Local Correlation Score (LCS), a statistic for local correlation of gene expression patterns. The significance of local enrichment in CEG was evaluated using locally adaptive procedure (LAP) (25). The correlation between expression patterns of genes localized in a region (LCS) was computed using a sliding window whose width was set equal to twice $\mu_{id,c} + 2\sigma_{id,c}$, where $\mu_{id,c}$ and $\sigma_{id,c}$ are, respectively, the mean and standard deviation of log-transformed intergenic distances, and are computed independently on each chromosome c. In details, for each gene contained in the gene expression data matrix and located at specific position j on chromosome c, Spearman correlation was computed pairwise for all of the neighbouring genes located in the window $j \pm n_c$, with n_c equal to $\mu_{id,c} + 2\sigma_{id,c}$. The LCS was defined as the median correlation among all of the pairwise correlation coefficients. If no gene was contained in the window except the central gene at position j, the LCS was set equal to zero. Then, the significance of positive or negative local peaks in LCS values was evaluated using LAP. LAP procedure consists of three main steps: (i) adoption of a statistic for each gene contained into the gene expression data matrix; (ii) adaptive bandwidth smoothing of the statistic after sorting the statistical scores according to the chromosomal position of the corresponding genes and (iii) application of a permutation test to identify chromosomal regions with significant positive or negative peaks of the selected statistic, with a q-value correction for multiple tests. The LAP procedure was applied to LCS statistic and allowed the identification of CER with significantly high (+CER) or significantly low (–CER) levels of local correlation among gene expression patterns. These genomic regions include groups of genes co-expressed and co-localized that were considered for subsequent analyses.

Inter Regional Correlation Score (IRCS): Co-expressed Chromosomal Meta-Regions (CEMR)

Distinct chromosome arms and chromatin domains may occupy discrete territories in the cell nucleus, whose topological characteristics are essential for gene regulation (26). Therefore, we searched for groups of CER showing similar expression patterns. Previously selected +CER were clustered according to the similarity of expression, quantified in terms of IRCS. IRCS between two regions CER_A and CER_B was defined as the median value of all pairwise Spearman correlations between the n genes of CER_A and the m genes of CER_B. QT clustering based on IRCS was used to group different CER into CEMR. These CEMR may be indicative of functional domains characterized by a high-level expression regulation.

Framework for motif discovery in promoters sequences

A computational framework was developed for identifying putative regulatory motifs in promoter sequences of selected groups of genes (SELGPi) as compared with REFGP. The framework comprises methodologies and software for completing a number of analysis steps, including (i) approximate patterns enumeration, (ii) significance scoring of over-representation, (iii) generation of motifs and (iv) their comparison with known regulatory sequences (Figure 1). Data obtained from different levels of analysis are recorded in a MySQL database and integrated with expression data.

Groups of exact patterns over-represented in promoter windows. Full-length promoter sequences of the REFGP are divided into overlapping sequence windows, with window width and overlap defined by the user. For each window, sequence patterns of a given length are examined. The occurrences (in both strands) of each pattern and the sequences that contain each pattern are counted and stored in MySQL database tables, thus providing an estimate of expected frequencies in the REFGP.

Then, as shown in Figure 1, for each of the selected groups of gene promoters (SELGPi), a specific sequence window is considered, and approximate patterns occurring in at least s-sequences are identified with SPEXS (27). We considered ungapped patterns with two, not lateral, variable positions, in which any of the four nucleotides is allowed, e.g. ANATGNTCGT, N={A,C,T,G}). However, the evaluation of over-representation of the approximate patterns may produce many false positive and false negative results. In fact, approximate patterns can match both over-represented and under-represented exact patterns, thus biasing the over-representation statistic. Therefore, each approximate pattern is first associated to the group of corresponding exact patterns, which are subsequently filtered to select only those occurring in at least two different sequences and being more represented in SELGPi than expected by chance, according to the estimated frequency in the REFGP. Thus, each approximate pattern is associated to the group X containing a set of h exact patterns satisfying these conditions and the total occurrences n of these exact patterns in SELGPi are compared to the total expected frequencies, according to REFGP. Finally, the probability of observing more than n occurrences, for the group X within the SELGPi, is evaluated using the binomial distribution as described in (28). The false discovery rate (FDR) is then used to control false positive results due to multiple statistical tests (29).

In this way, for each SELGPi (and for each combination of pattern length and sequence window), a given number of significantly over-represented groups of exact patterns is identified.

k-medoids clustering of significant patterns generates motifs. For each SELGPi, all the exact patterns belonging to any over-represented group X are then compared and assembled into motifs with k-medoids clustering (30), using the TAMO package (31). In order to assess the pairwise distances between patterns, we adopted the end-space free alignment algorithm, whose peculiarity is avoiding penalization for mismatched overlapping prefixes and/or suffixes of aligned sequences. Since k-medoids clustering requires an *a priori* determined number of clusters k and is a heuristic process to find the optimal value of k (i.e. the minimum number of clusters producing a sufficiently good pattern partitioning), the clustering analysis was

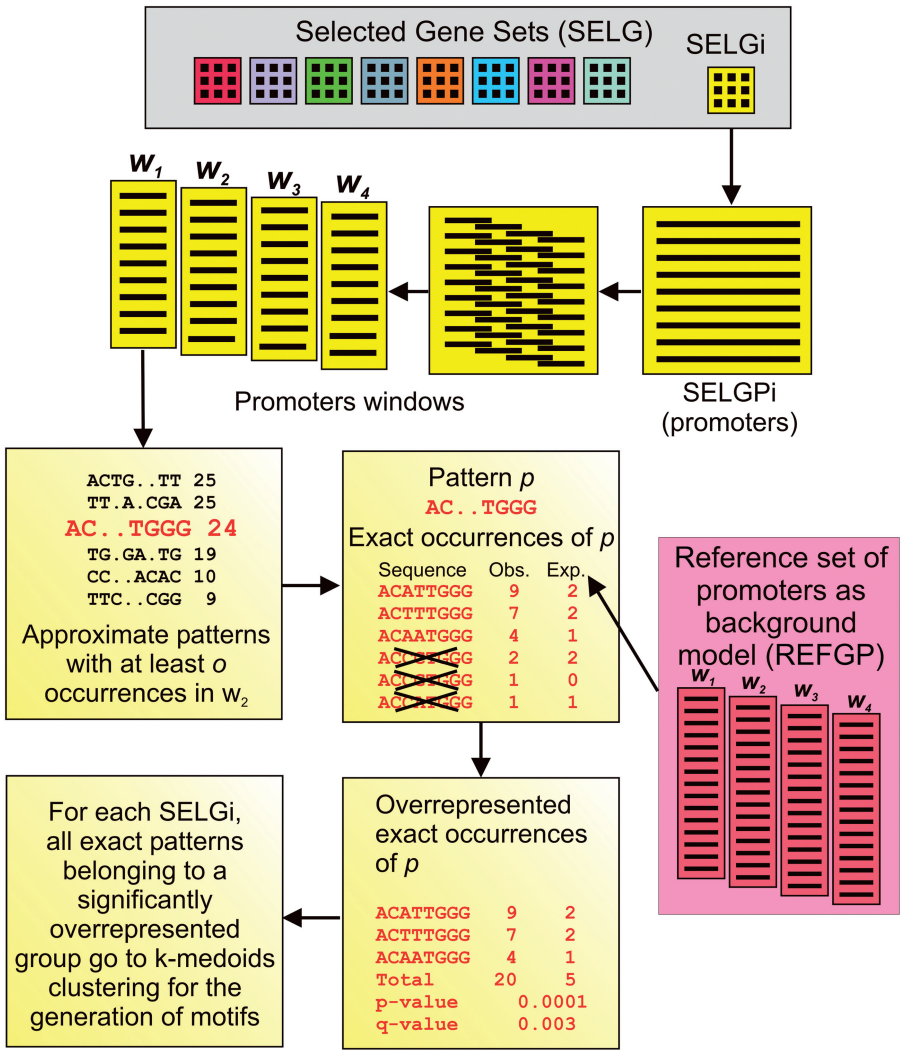


Figure 1. Scoring exact patterns over-representation in a selected set of gene promoters. For each SELGi, the corresponding SELGPi is divided in overlapping sequence windows. Approximate patterns occurring in at least 30% of the considered promoters are used to select the corresponding group of exact patterns. Then, within the groups of exact patterns, only patterns showing more occurrences than expected by chance: their total number of occurrences is used to compute over-representation *P*-value of the whole group of exact patterns. If the *P*-value is significant after Benjamini FDR correction, they are clustered together with other over-represented patterns that identify over-represented motifs corresponding to the specific SELGPi.

repeated with increasing values of k until partitioning optimality was less or equal to a user-defined threshold. Partitioning optimality (d_k) was defined as the maximum of the distances between cluster medoids and individual cluster members. To stabilize results, for each incremental value of k, the k-medoids clustering was performed 100 times. Then the partitioning having the minimum d_k was selected as the best among the 100 trials.

Finally, to obtain a single over-represented motif corresponding to a cluster of over-represented patterns, all patterns within a cluster were multialigned using ClustalW (32) with a stringent gap open penalty (100) to avoid gaps in the resulting consensus sequence. A matrix of nucleotide frequencies in motif positions was computed from patterns alignment result and visualized with a sequence logo. Over-represented motifs occurring in <30% of SELGPi were not included in final results.

Comparison with known TFBS. A set of known binding sequences for human transcription factors (TFs) was collected from JASPAR and TESS databases. In total 142 sequence motifs or variant of motifs, each known to be recognized by a TF were collected. This group also included motifs corresponding to 21 TFs for which a specific involvement in myeloid cell differentiation is well known (Supplementary Data file 1). For each TF, all available binding sequences were grouped with k-medoids clustering, as above described, to generate one motif and the corresponding matrix of nucleotide frequencies.

Significantly over-represented motifs were compared with known-transcription factor binding sequences (TFBS) using the 'scan' function of the TAMO library (31) to evaluate the similarity of the corresponding groups of sequences. Two motifs were considered similar if the match among their corresponding sets of sequences

exceeds 70% of the best possible score that could be obtained, given the number and length of sequences.

The scripts constituting our motif discovery framework are freely available at the URL http://compgen.bio.unipd.it/MoDi/ along with a copy of the promoter database used for the analyses and associated documentation.

Uniformity analysis of motifs distribution along promoter sequences

Uniformity analysis was performed to identify motifs with occurrences non-homogeneously distributed along promoter sequences. The occurrences of each motif (M_i) in each promoter sequence of the SELGPi are compared with uniform distribution: Chi-squared test is used to evaluate the significance of the differences between the observed and the expected occurrences of M_i, in a set of non-overlapping windows of promoter sequences, as described in (21).

RESULTS

Datasets

Myelopoiesis gene expression data. As detailed in Materials and methods section, the gene expression dataset was contained in a data matrix with 11 446 genes/custom probesets and 24 samples for eight different cell types of the human myeloid lineage (Supplementary Data file 2: RMA gene expression data matrix).

Reference set of human genes: promoter, position. A reference set of human genes was collected, as the complete set of EntrezGeneIDs corresponding to trustable nuclear genes. For 15 138 of these genes, the genomic position was precisely defined and promoter sequences retrieved. Thus, 15 138 promoter sequences, each encompassing 1000 bp upstream and 100 bp downstream of the predicted gene TSS (–1000, +100), constituted the REFGP (freely available at the URL http://compgen.bio.unipd.it/MoDi/).

Integrated dataset. The intersection of 11 446 genes/probesets of the expression data matrix with the reference set of 15 138 gene/promoters comprised 9716 genes for which genomic localization, promoter sequence, and expression data in myeloid cells were available. For each gene in the integrated dataset, Gene Ontology functional annotations were retrieved from the EntrezGene database.

Sets of CEG in myeloid cells

Sets of genes showing similar expression patterns constitute the first most intuitive candidates for sharing regulatory motifs. Genes with variable expression were selected and grouped according to their expression patterns into sets of CEG (Figure 2, yellow panel, and Figure 3). A set of 2796 (29%) genes with highly to moderately variable expression were selected (Shannon entropy, $H \leqslant 2.8$) and then grouped by similarity of expression using QT clustering. Setting maximum cluster diameter to 0.25 (minimum correlation of 0.75) with at least 15 genes per cluster, we obtained 44 gene clusters, including a total of 2455 genes. Each cluster represents a group of human genes co-expressed during myelopoiesis (CEG) (Supplementary Data file 3: sets of human genes co-expressed during myelopoiesis). In Figure 3, expression plots of 15 CEG sets with at least 40 genes per set are reported, whereas all plots are available in the Supplementary Data file 3. Functional GO terms enrichment was tested on each considered CEG detecting significantly enriched terms (hypergeometric test with a P-value $\leqslant 0.05$ and at least 10% of geneset genes, or five genes, in each category). Results are available online (http://compgen.bio.unipd.it/MoDi/).

CER, sets of neighbouring genes similarly expressed during myelopoiesis

A number of evidences support the existence of mechanisms for positional regulation of gene expression, influencing transcription within specific chromosomal regions (7). This high level of gene expression regulation was taken into account as well, and we looked for CER along myelopoiesis (Figure 2, green panel). The analysis of CER was carried out using the LCS statistic. LCS is computed for each gene position, considering the correlation with neighbouring genes within a specific window, as described in Materials and methods section. Window width was selected independently for each chromosome, taking into account the different gene density of chromosomes. The average window width was 3.66 Mb, with values ranging from 0.72 Mb (chr 19) to 5.69 Mb (chr 13), thus including on average 5.7 genes per window. Then, by applying LAP to LCS statistic (q-value $\leqslant 0.01$), we identified chromosomal regions, including at least five genes per region, with significantly high (positively correlated regions, +CER) or significantly low correlation among gene expression patterns (negatively correlated regions, –CER). We identified 34 +CER, including a total of 922 genes covering 211.84 Mb (7% of the human genome), and 4 –CER, with a total of 53 genes covering 16.77 Mb (0.54% of the human genome) (Figure 4).

It could be noticed that the number and the span of negatively correlated regions are considerably lower than those of positively correlated even if the correlation coefficients between genes are symmetrically distributed around zero, within the genomic windows used for computing LCS (Supplementary Data file 4, panel A). Therefore, we also investigated the relationship between the physical distance and the correlation of adjacent genes. This analysis showed an apparent inverse correlation between the distance of adjacent gene pairs and the correlation coefficient of corresponding expression patterns. Positively correlated genes tend to be closer to each other, whereas negatively correlated genes tend to be separated by larger intergenic regions (Supplementary Data file 4, panel B).

For the subsequent analyses, we focused only on positively correlated regions. Since the widths of +CER (6.23 Mb in average, min 1.31 Mb, max 16.48 Mb) exceed those of the original windows used for LCS computation, we verified the actual level of correlation between genes included into +CER. The distribution of pairwise Spearman correlations between genes of each +CER was evaluated and 26 CER with high correlation between

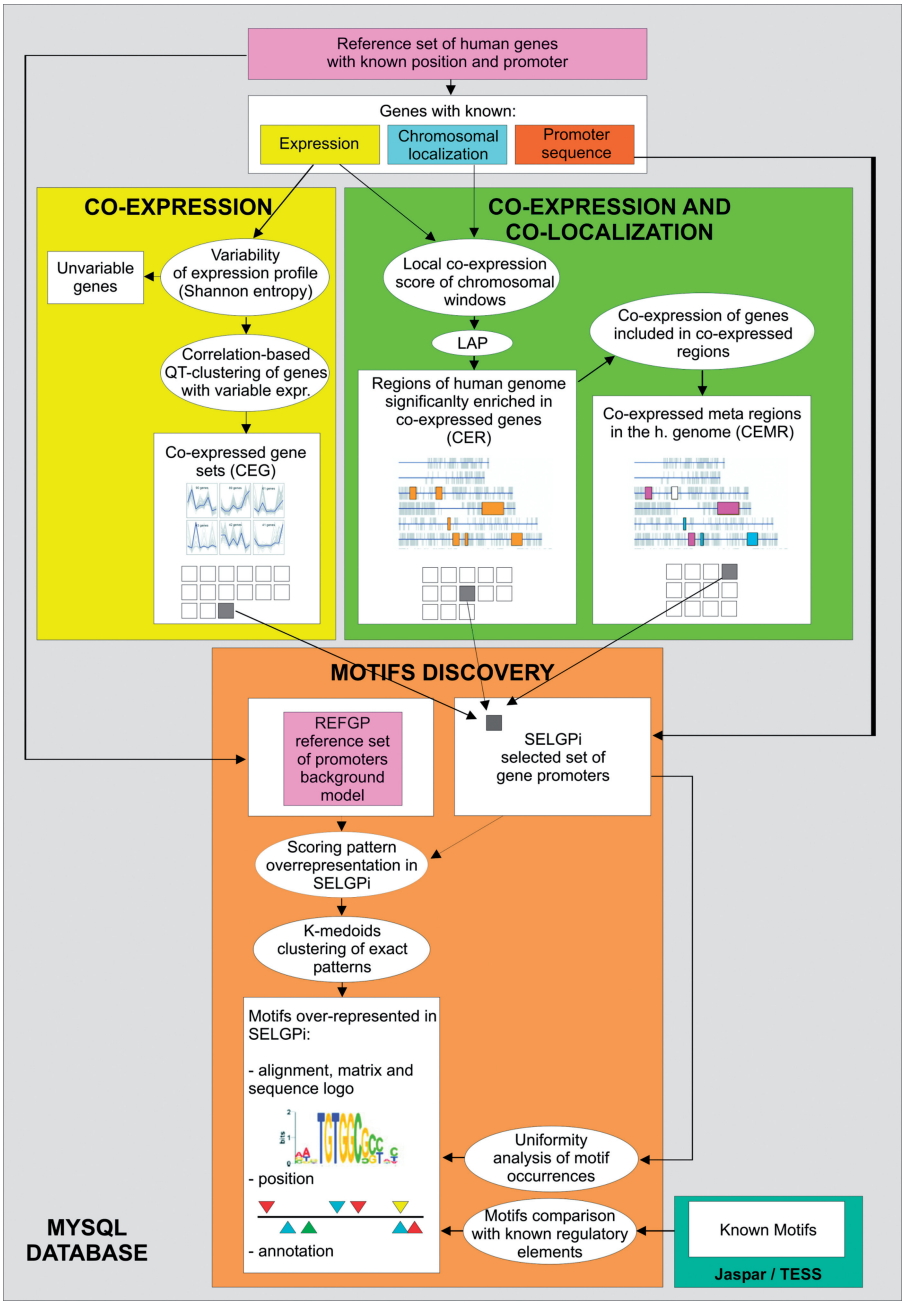


Figure 2. Experimental schema. Genomic databases with information on gene localization and sequences are used to generate the database of reference promoter sequences. Information on genes including expression data, chromosomal localization and promoter sequences are then taken into account by the integrated analytical framework. CEG sets are selected by the mere analysis of expression data. Then, expression data and genomic information are combined to select co-expressed and co-localized genes, thus identifying CER, which are subsequently grouped in CEMR, according to the similarity of the associated expression patterns. All of these set of genes are then used to select the corresponding sets of SELGPi, which are analysed with our motif discovery procedure. The motif discovery analysis combines the use of approximate patterns for grouping of underlying exact patterns, binomial distribution with FDR correction, in order to select over-represented exact patterns with an adequate statistical significance. Then, clustering of over-represented patterns leads to the identification of significantly over-represented motifs. Posterior analyses of over-represented motifs allow their further annotations with likely biologically relevant characteristics, including their non-uniform distribution along promoters, their occurrences in SELGPi, their matches with known TFBS and their occurrence in a subset of highly correlated genes.

corresponding genes (third quartile of pairwise correlations $\geqslant 0.5$) were selected for further analyses, as reliable groups of co-localized genes and co-expressed genes. Figure 5 reports the expression plot and heatmap, with genes ordered by genomic position, for a +CER including 13 genes localized in chromosome 12 (81 270 750 – 90 100 937); plots and heatmaps for the complete set of 26 +CER are in Supplementary Data file 5. The information about CER is also available using distributed annotation system (DAS) (33) (http://compgen.bio.unipd.it/Annotations/das/). Functional GO terms enrichment of CER was conducted as described for CEG.

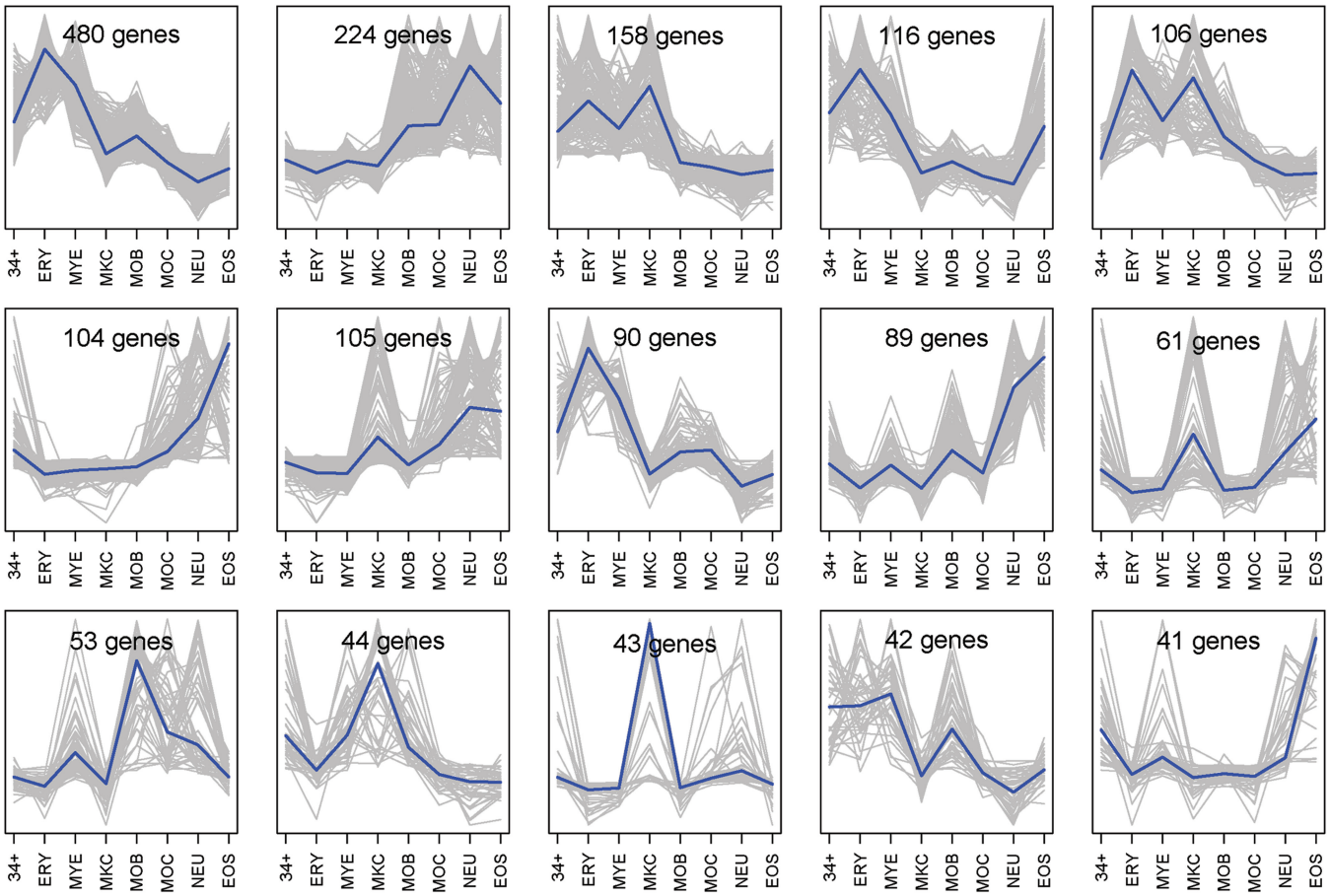


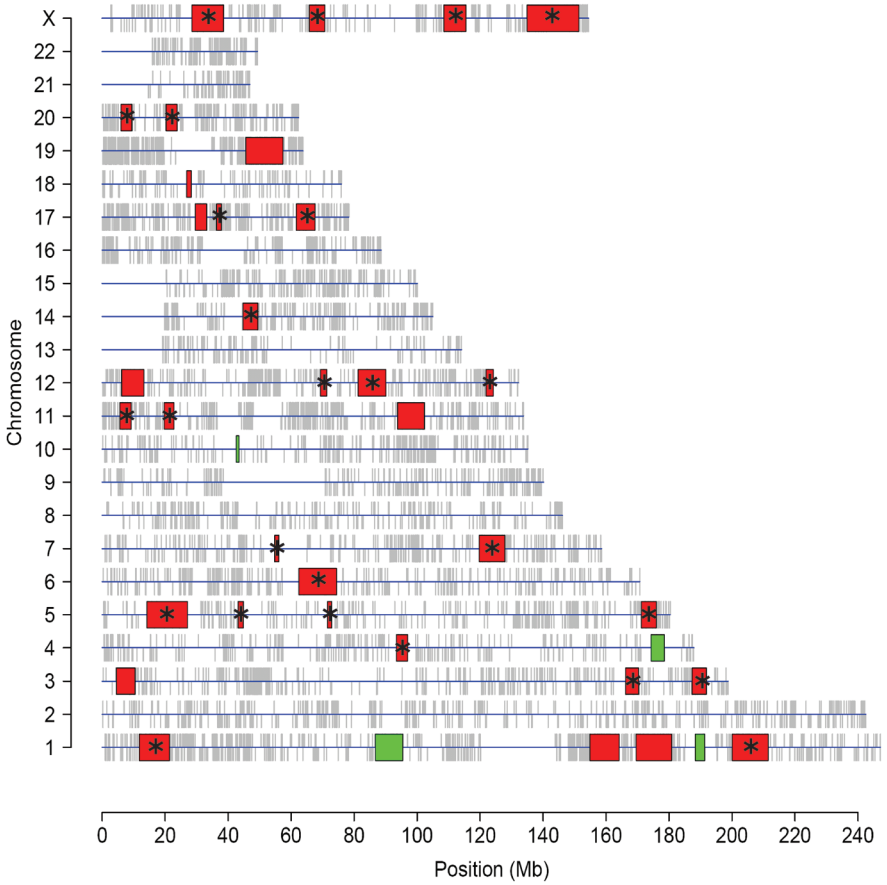
Figure 3. Expression plots of 15 co-expressed gene sets with at least 40 genes per set. QT clustering analysis allowed identifying 44 groups of highly correlated genes: CEG sets. The plot shown in grey expression patterns corresponds to the 15 CEG including more than 40 genes; the blue-bold line represents the median of the expression profiles in each gene set. (34+: CD34+ HSC; ERY: Erythroblasts; MYE: Myeloblasts; MKC: Megakaryoblasts; MOB: Monoblasts; MOC: Monocytes; NEU: Neutrophils; EOS: Eosinophils).

CEMR

We reasoned that the number of possible variants of gene expression profiles, calculated on eight different cell types, is expected to be limited and that it would be possible to find similarity among pairs or groups of profiles corresponding to different +CER. Moreover, since distinct chromosome portions may occupy discrete territories in the cell nucleus (26,34), +CER with similar expression profiles might constitute chromatin domains with a specific functional role and a peculiar localization within the nucleus (CEMR; Figure 2, green panel). Therefore, as detailed in Materials and methods section, IRCS as similarity measure and QT clustering (with maximum distance set to 0.7) were used for grouping +CER into meta-regions (CEMR). Fifteen out of 26 selected +CER were grouped into two CEMR: one CEMR includes 10 +CER, distributed in seven different chromosomes, whereas the other is composed of five +CER located in four different chromosomes. The remaining 11 +CER cannot be grouped into CEMR according to the selected thresholds. The genes belonging to each CEMR clearly show a specific expression profile and have median correlation among them $\geqslant$0.4 (Figure 6). Figure 6 reports the position of the 26 original +CER in human chromosomes. Magenta and light blue colours indicate +CER belonging to the two selected CEMR, for which the expression profiles are also given, whereas white blocks represent +CER, which cannot be grouped into CEMR. CEMR information is available as DAS annotation (33). Functional GO terms enrichment of CEMR was conducted, as described for CEG.

Identification of motifs over-represented in promoters of CEG sets, CER or CEMR, with putative regulatory role

As above described, 44 CEG sets were identified by classical analysis of gene expression data (see also Supplementary Data file 3 and Figure 3), whereas integrated analysis of gene expression and chromosomal localization allowed identifying 26 sets of genes co-expressed and co-localized (CER), and two sets of genes included in CEMR of the human genome (CEMR). For each of these gene sets, the corresponding group of gene promoters was considered and analysed to discover significantly over-represented motifs in SELGPi, as compared with a large group of 15 138 promoters (REFGP, being the



	Positively correlated regions	Span (Mb)	# genes
*	chr 1: 11768373 - 21375927	9.61	57
	chr 1: 154878363 - 164143961	9.27	62
	chr 1: 169550109 - 180840166	11.29	37
*	chr 1: 200064910 - 211513430	11.45	81
	chr 3: 4510135 - 10466546	5.96	19
*	chr 3: 166179379 - 170346787	4.17	8
*	chr 3: 187347683 - 191851995	4.5	17
*	chr 4: 93444830 - 96981647	3.54	6
*	chr 5: 14196828 - 27074446	12.88	9
*	chr 5: 43157398 - 44851371	1.69	5
*	chr 5: 71551111 - 72897254	1.35	5
*	chr 5: 171221160 - 175955375	4.73	14
*	chr 6: 62447823 - 74420411	11.97	15
*	chr 7: 54787434 - 56099176	1.31	5
*	chr 7: 119700957 - 127885706	8.18	23
*	chr 11: 5641363 - 9243447	3.6	27
*	chr 11: 19691487 - 22791122	3.1	8
	chr 11: 93790114 - 102331672	8.54	24
	chr 12: 6179815 - 13260974	7.08	79
*	chr 12: 69318128 - 71345688	2.03	8
*	chr 12: 81270750 - 90100937	8.83	13
*	chr 12: 121915834 - 124193824	2.28	13
*	chr 14: 44742143 - 49431484	4.69	8
	chr 17: 29606408 - 33179182	3.57	27
*	chr 17: 36332477 - 37828800	1.5	28
*	chr 17: 61638613 - 67634147	6	18
	chr 18: 26825028 - 28304445	1.48	10
	chr 19: 45638590 - 57421482	11.78	193
*	chr 20: 6003492 - 9459171	3.46	7
*	chr 20: 20296764 - 23808380	3.51	12
*	chr X: 28515479 - 38550640	10.04	15
*	chr X: 65732207 - 70712465	4.98	22
*	chr X: 108502790 - 115506651	7	19
*	chr X: 134895251 - 151370486	16.48	28
	Negatively correlated regions	**Span (Mb)**	**# genes**
	chr 1: 86785348 - 95485361	8.7	30
	chr 1: 188333419 - 191341867	3.01	7
	chr 4: 174326529 - 178520994	4.19	11
	chr 10: 42598264 - 43464332	0.87	5

Figure 4. Genomic positions of co-expressed regions in the human chromosomes. The analysis of CER was carried out using the LCS statistic, with a sliding window approach applied to the human genome. Regions harbouring genes with highly positive and negative correlation of expression in myelopoiesis are shown in red and green, respectively. The table on the right reports details on CER localization, span, and the number of genes from the region, which is present in the gene expression data matrix. An asterisk is used to mark positively correlated CER, which were selected for subsequent analyses, including the CEMR and motif discovery analyses.

background model), in order to identify putative regulatory elements (Figure 2).

Each selected set of genes is related to a specific expression pattern and likely to a specific biological role in a given differentiation context. Thus, significant motifs identified for each considered gene set are expected to play a functional role in a specific program of cellular differentiation.

Incidentally, since each considered promoter sequence was defined separately, we checked for overlapping promoter sequences belonging to gene pairs included in the same gene set: only 0.17% of considered sequences overlap in 13 (mostly very small) regions. Among all the considered genes, only two promoters, belonging to genes included in the same gene set, show an overlap of a number of nucleotides close to 1100, which is the length of considered promoter sequences: this is a pair of divergent genes with a unique bidirectional promoter (PDCD10, programmed cell death 10 and SERPINI1, neuroserpin precursor). This finding is in accordance with previous data on co-expression of genes with bidirectional promoters: PDCD10 and SERPINI1 were included in the same set of CEG (35).

From each considered 1100 bp promoter, five sequence windows of 300 bp in width and overlapping each other 100 bp, were extracted. For each sequence window, we selected approximate patterns of six and eight nucleotides, occurring in at least 30% of promoter sequences in the SELGPi. Since a previous systematic survey of known regulatory sites and motifs, available in TRANSFAC database, enlightened the over-representation of even-length functional motifs (16), we focused on even-length patterns as well. Among exact sequences matched by a given approximate pattern, a group of exact patterns was selected as over-represented (with FDR set to 0.05) in the specific SELGPi sequence window, as detailed in Materials and methods section. Thus, for each of the considered windows and for each pattern length, groups of over-represented exact patterns were identified, and subsequently clustered to obtain over-represented sequence motifs, each of them defined by a motif consensus sequence (represented as sequence logo) and a matrix of nucleotide frequencies in motif positions. Each detected motif is associated to the following attributes (Figure 7, and dedicated website: http://compgen.bio.unipd.it/MoDi/): (i) matrix of nucleotides frequencies in motif