*Article*

# Gene expression landscape of Chronic Myeloid Leukemia K562 cells overexpressing the tumour suppressor gene PTPRG

**Giulia Lombardi[1,†,‡] 0000-0002-8287-518X, Roberta Valeria Latorre[2] 0000-0001-6722-5683, Alessandro Mosca[1] 0000-0003-2323-3344, Diego Calvanese[1] 0000-0001-5174-9693, Luisa Tomasello[2] 0000-0001-8700-1759, Christian Boni[3], Manuela Ferracini[4] 0000-0002-1595-6887, Massimo Negrini[5] 0000-0002-0007-1920, Nader Al Dewik [6,7,8,9] 0000-0001-5739-1135, Mohamed Yassin [10] 0000-0002-1144-8076, Mohamed. A. Ismail [6−11] 0000-0001-7647-856X, Bruno Carpentieri[1] 0000-0003-1960-9986, Claudio Sorio[3,‡,*] 0000-0003-2739-4014, and Paola Lecca [1,‡,*] 0000-0002-7224-136X**

[1] Faculty of Computer Science, Free Univesity of Bozen-Bolzano
Piazza Domenicani 3, 39100 Bolzano, Italy

[2] Department of Cancer Biology and Genetics, The Ohio State University,
Columbus, OH 43210, USA

[3] Department of Medicine, Division of General Pathology, University of Verona,
Strada Le Grazie 8, 37134 Verona, Italy

[4] Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna,
Via S.Giacomo 14, 40126 Bologna, Italy

[5] Dipartimento di Medicina Traslazionale e per la Romagna, University of Ferrara
Via Fossato di Mortara 70, 44121 Ferrara, Italy

[6] School of Life Science, Pharmacy and Chemistry, Faculty of Science, Engineering & Computing, Kingston University London, United Kingdom.

[7] Qatar Medical Genetic Center (QMGC), Hamad General Hospital (HGH), and Interim Translational Research Institute (iTRI), Hamad Medical Corporation (HMC), Doha, Qatar

[8] College of Health and Life Science (CHLS), Hamad Bin Khalifa University (HBKU), Doha, Qatar

[10] Department of Pediatrics, Hamad General Hospital (HGH), HMC, Doha, Qatar

[11] Interim Translational Research Institute (iTRI), Hamad Medical Corporation (HMC), Doha, Qatar.

* Correspondence: paola.lecca@unibz.it; claudio.sorio@univr.it;giulia.lombardi.unitn.it

† Current address: Department of Mathematics, University of Trento, via Sommarive 14, 38123 Trento, Italy

‡ These authors contributed equally to this work.

**Citation:** Lombardi G., Latorre R. V., Mosca A., Calvanese C., Tomasello L., Boni C., Ferracini M., Negrini M., Dewik N. A. , Yassin M., Ismail M, A. Carpentieri B., Sorio C, and Lecca P. Gene expression landscape of Chronic Myeloid Leukemia K562 cells overexpressing the tumour suppressor gene PTPRG. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Abstract:** This study concerns the analysis of the modulation of Chronic Myeloid Leukemia (CML) cell model K562 transcriptome following transfection with the tumor suppressor gene encoding for Protein Tyrosine Phosphatase Receptor Type G (PTPRG) and the treatment with the tyrosine kinase inhibitor (TKI) Imatinib. Specifically, we aimed at identifying genes whose level of expression is altered by PTPRG modulation and Imatinib concentration. Statistical tests as differential expression analysis (DEA) supported by gene set enrichment analysis (GSEA) and modern methods of ontological term analysis are presented along with some results of current interest for forthcoming experimental research in the field of the transcriptomic landscape of CML. In particular, we present two methods that differ in the order of the analysis steps. After a gene selection based on fold-change value thresholding, we applied statistical tests to select differentially expressed genes. Therefore, we applied two different methods on the set of differentially expressed genes. With the first method (Method 1) we implemented GSEA, followed by the identification of transcription factors. With the second method (Method 2), we first selected the transcription factors from the set of differentially expressed genes and implemented GSEA on this set. Method 1 is a standard method commonly used in this type of analysis, while method 2 is unconventional and is motivated by the intention to identify transcription factors more specifically involved to biological processes relevant to the CML condition. Both methods have been equipped by ontological knowledge mining and word cloud analysis, as elements of novelty of our analytical procedure. Data analysis identified RARG and CD36 as a potential PTPRG upregulated genes suggesting a possible induction of cell differentiation toward a erithromyeloid phenotype. The prediction was confirmed at the mRNA and protein level, further validating the approach and identifying a new molecular mechanism of tumour suppression governed by PTPRG in a CML context.

**Keywords:** Differential expression gene analysis; enrichment gene analysis; bio-ontologies; Chronic Myeloid Leukemia.

## 1. Introduction

Chronic Myeloid Leukemia (CML) is a myeloproliferative disease affecting approximately 1 per 200,000 persons per year in industrialized countries. Many treatment improvements have been achieved recently, especially in the development of new drugs, but a mortality rate of 2-3% per year remains [1,2]. A distinctive feature of CML is the reciprocal translocation, originating in hematopoietic stem cells (HSCs), between the long arms of chromosomes 9 and 22, i.e. t(9;22)(q34;q11.2), which results in the BCR-ABL1 chimeric gene. This genomic aberration generates a new fusion gene, BCR-ABL1, which encodes for a tyrosine kinase held accountable for the neoplastic transformation of these cells by affecting normal cellular pathways essential for tissue homeostasis, and thus causing the alteration of crucial cellular processes, such as apoptosis, cell cycle and autophagy [3,4]. In this context, one primary goal of the research is to identify the regulatory mechanisms antagonizing the kinase activity of BCR-ABL1 and, possibly, of other vital effectors intersecting this pathway as players other than BCR-ABL1 have been involved in the pathogenesis of the disease [5,6]. The natural history of CML, prior to the advent of small molecule protein kinase antagonists, feature a progression from a stable or chronic phase to an accelerated phase, or to a rapidly fatal blast crisis within 3–5 years. Typically blood cells differentiate normally in the stable phase, but not in the blast phase [1]. Protein Tyrosine Phosphatase receptor type G (PTPRG) is a member of the protein tyrosine phosphatase (PTP) family featuring an extracellular and a single transmembrane region and two tandem intracytoplasmic catalytic domains [7]. PTPRG is widely expressed in human tissues [8] and is involved in the regulation of cell growth, differentiation, mitotic cycle, and oncogenic transformation [9,10]. The gene encoding for this phosphatase is located in a chromosomal region (3p21-p14.2) frequently deleted in renal cell and lung carcinoma, where PTPRG acts as a tumor suppressor in many cancers [11–14]. Specifically, PTPRG was recognized as having an oncosuppressor function gene and was found downregulated in CML patients. This relevance of this gene to CML has been recently supported by several studies performed in patients and strategies aimed at restoring its expression are expected to benefit the course of the disease by improving drug efficacy or contrasting the emergence of BCR/ABL1 mutants [15–17].

Epigenetic events, such as the hyper-methylation of its promoter region as well as intron 1, negatively regulates the transcription of PTPRG, as demonstrated in CML and childhood acute lymphoblastic leukemia [16,18–20]. Re-expression of this protein occurs in leukocytes (especially neutropils) of CML patients following targeted therapy [18]. Once activated, PTPRG can reduce the phosphorylation level of BCR-ABL1 and some of its key targets, such as CRK-L and STAT512. We found that, in CML cells, PTPRG expression inversely correlates with BCR/ABL1 expression and activation, both in cell lines and primary cells models following pathways that include beta catenin [21] and possibly others are currently under investigation [17,20,21].

Our study focuses on the detection of genes and gene pathways in protein-protein interaction networks (commonly considered a proxy of gene network) that are most likely affected by the state of the gene coding for PTPRG and by the treatment a prototype tyrosine kinases inhibitor (TKI), Imatinib, in K562 cell line overexpressing the enzymatic active and enzymatic dead PTPRG. Tyrosine kinases phosphorylate proteins on tyrosine residues, producing a biologic signal that also influences many aspects of cellular functions including cell growth, proliferation, differentiation, and death. PTPs act as natural modulators of TKI signaling and it is well known how inhibition of TKI represents a strategy to disrupt signalling pathways that promote neoplastic growth and survival in haematologic malignancies and likely in other neoplasia as well. In order to identify responsive genes we implemented two analytical pipelines hereafter referred to as Method 1 and Method 2. On the set of differentially expressed genes we applied two methods of analysis. With the first method (Method 1) we implemented Gene Set Enrichment Analysis (GSEA), followed by the identification of transcription factors. With the second method (Method 2), we first selected the transcription factors (TFs) from the set of differentially expressed genes and implemented GSEA on this set. Method 1 is a standard method commonly used in this type of analysis, while Method 2 is unconventional and is motivated by the intention to identify transcription factors more specifically involved to biological processes relevant to the CML condition. In Method 1, due to a larger gene universe, we expect the set of transcription factors selected upstream of the GSEA to be either larger or related to the known role of PTPRG as modulator of hematopoietic cell differentiation [10].

## 2. Materials and methods

In this section, we report on the methods and materials relevant to the experimental activity of data collection, while we devote the next section to the description of methods pertinent to the computational analysis of data

### *2.1. Cell lines*

The human K562 chronic myeloid leukemia clones expressing full-length PTPRG, empty vector and inactive mutant holding a mutation on the catalytic domain D1028A were previously described [18] and were cultured in RPMI medium supplemented with 1% L-glutamine 100X (Biowest), 10% fetal bovine serum (FBS, Euroclone) and the selective agent G418 0,=.5 mg/mL (Sigma) at 37° C in an humidified incubator with 5% $CO_2$.

### *2.2. Quantitative Real-Time Polymerase Chain Reaction*

Total RNA was extracted from the K562 cell lines using Qiagen RNeasy Kit according to the manufacturer's protocol. Complementary DNA was synthesized using the PrimeScript RT reagent Kit (TAKARA BIO INC.) and the quantitative real-time polymerase chain reaction (qRT-PCR) was performed using TB Green Premix Ex taq (TAKARA BIO INC.). Each sample was run in triplicate and 3 ng complementary DNA was used for each reaction. The sequences of gene-specific primers used are listed in Table 1. The fold changes in mRNA levels of transcription factors (TF-DEGs) between K562 cell line expressing PTPRG and control group were determined using the $2^{-}\Delta\Delta CT$ method with GAPDH used as the internal control for normalization. Prism (GraphPad Software) was used for statistical analyses and the Student's t-test was used to determine statistically significant differences between groups.

| PRIMER | Forward | Reverse |
|---|---|---|
| MECP2 | CGTGAAGGAGTCTTCTATCCGA | GCTTCACCACTTCCTTGACC |
| TFAP2C | ATTCGCAAAGGTCCCATTTCC | GGCATTTAAGCATTCAGGTGG |
| RARG | GCAAGTATACCACGAACTCCAG | ACGCAGCATCAGGATATCTAGG |
| TRPS1 | CAAACAAGAAGCAAATCACCTG | GTGTGCTCTCCTGTAGTGTC |
| SMAD1 | TCCTTCCAACAATAAGAACCGT | CTACTGTCACTAAGGCATTCG |
| CD36 | TTTGGCTTAATGAGACTGGGAC | ACAAACATCACCACACCAACAC |

**Table 1.** Sequences of gene-specific primers used in this study.

### *2.3. Flow Cytometry Analysis*

The K562 cell lines (5×105 cells) were harvested, washed in FACS buffer (PBS supplemented with 2% FBS and 2 mM EDTA) and centrifuged at 1200 rpm for 5 min at room temperature. The cell suspensions (100$\mu$L) were plated in 96 well plate and 2uL anti-CD36 (V450 mouse 2-Human CD36; cat.no. 561535; BD Biosciences) were added. The samples were incubated in the dark for 1h at 4°C, washed again with FACS buffer and centrifugated

(1200 rpm for 5 min). FACS buffer (150 $\mu$l) was added to the cell pellet, and the samples were analyzed using MACSQuant® Analyzer 10 Flow Cytometer (Miltenyi Biotec). The data was analyzed with FlowJo$^{TM}$ v10.8.1 software and the fraction of positively stained cells (CD36+) was determined as the percentage of live population stained with Propidium Iodide (PI).

### *2.4. Data collection*

The RNAs from the samples were hybridized on Agilent whole human genome oligo microarray (#G4851A, Agilent Technologies, Palo Alto, CA). This microarray consists of 60-mer DNA probes synthesized with SurePrint technology, covering 60,000 unique human transcripts. One-color gene expression was performed according to the manufacturer's procedure. Briefly, total RNA fraction was obtained from samples by using the Trizol Reagent (Invitrogen). RNA quality was assessed by the use of Agilent 2100 Bioanalyzer (Agilent Technologies). Low quality RNAs (RNA integrity number below 7) were excluded from microarray analyses. Labeled cRNA was synthesized from 100 ng of total RNA using the Low Imput Quick-Amp Labeling Kit, one color (Agilent Technologies) in the presence of cyanine 3-CTP. Hybridizations were performed at 65°C for 17 hours in a rotating oven. Images at 3um resolution were generated by Agilent scanner and the Feature Extraction 10.7.3.1 software (Agilent Technologies) was used to obtain the microarray raw-data.

Microarray results were then analyzed by using the GeneSpring GX 11 software (Agilent Technologies). Data transformation was applied to set all the negative raw values at 1.0, followed by a normalization on the 75th percentile. A filter on low gene expression was used to keep only the probes expressed in at least one sample (flagged as Marginal or Present).

The data used in this study derive from the above-mentioned analysis carried out by microarray hybridization of the CML cell transcriptome (K562) in different conditions. The cells were transfected with full-length PTPRG and compared to several controls: cells transfected with the empty vector, cells transfected with PTPRG inactive mutant holding a mutation on the catalytic domain (D1028A), and cells treated with Imatinib targeting the oncogene BCR/ABL1. We integrated the data relating to gene expression with then gene ontology and protein-protein network data. We investigated

- the effect of the PTPRG expression and its activation status,
- the impact of PTPRG expression (both active or inactive) in the presence of TKI, hereafter called with its clinical name Imatinib.
- and the effect of Imatinib in combination with functional or mutant PTPRG expression.

For this purposes, we developed ad-hoc methods to identify differentially expressed genes, with a particular focus on gene coding for transcription factors. This class of genes was selected as they are known to act as master genes activating cell programs that include key features such as cell differentiation and proliferation other than controlling genes essential for the ontogenesis and maintenance of the normal hematopoietic system, other than being involved in the pathogenesis of leukemia [22].

## 3. Computational analysis

We implemented first differential gene expression analysis, and then gene ontology enrichment analysis of the identified differentially expressed genes (DEGs). Differential expression analysis (DEA) is a single-gene technique performed to identify differentially expressed genes (DEGs), namely genes whose expression levels vary significantly under different experimental conditions. Gene set enrichment analysis (GSEA) is a computational method applied to get biological insights from gene expression data. It is typically used to examine a given subset of interesting genes stemming from previous analyses versus an extensive reference set referred as gene universe. Unlike single-gene techniques, GSEA aims at identifying statistically significant groups of functionally-related genes by relying on current knowledge for data classification. For theoretical in-depth analyses about GSEA refer to [23]. Depending on the specific biological question that is designed to tackle, several databases can be employed to investigate *a priori* gene functional groupings.

### *3.1. Differential gene expression analysis*

In this study, differential gene expression analysis was performed to detect DEGs between two groups: control and the phosphatase inactive mutant D1028A [18] samples considered as the untreated group (4 replicates) and the treatment group referred to PTPRG expressing samples (2 replicates). Differential expression analysis was conducted on log2-trasformed data using the Bioconductor/R package `limma` (Version 3.12) [24]. Both the empirical Bayes correction on the variances and the multi-testing Bonferroni-Hochberg correction on p-values were selected.

Among the set of differentially expressed genes we focused on transcription factors (TF-DEGs). Indeed, the identification of the TF-DEGs responsive to the treatments would allow to identify the active drivers turning specific genes (possibly involved in the onset and progression of CML) "on" or "off", or boosting/repressing the gene's transcriptions.

### *3.2. Gene Ontology enrichment analysis of DEGs*

In this study, gene ontology enrichment analysis of the DEGs relied on the Gene Ontology (GO) system of classification [25] and in particular on the GO domain referred to biological processes. Therefore, over-representation of GO terms pertinent to the DEGs previously identified has been tested to reveal associations with disease phenotypes. Instead of investigating the results of GSEA applied to the whole set of DEGs, we focused on the transcription factors (TFs) detected as DEGs (TF-DEGs).

The analysis on transcription factors was carried out by applying two different GSEA methods implemented by the Bioconductor/R package `topGO` (version 3.12) [26]. Both methods combine a classical enrichment analysis with the Kolmogorov-Smirnov statistic test (`runTest` function with input parameters "algorithm = classic" and "statistic = ks"). This particular setting was selected for two reasons.

1. The methods compute the significance of a node independently from its neighbouring nodes [27]. This means that if a GO term contains the same genes as one of its children, then the traditional method give the children the same score. While this setting could cause data redundancy, on the other hand, by not discarding any GO term based on parent-child relationships, it allows keeping valuable information that can be exploited later on to investigate associations and dependencies between GO terms.
2. The Kolmogorov-Smirnov statistic computes enrichment based on gene scores [25]. Hence, it is possible to take full advantage of the information obtained by DEA by ranking genes according to their adjusted p-values.

These two considerations lead to two methods, hereafter referred to as Method 1 (corresponding to consideration 1) and Method 2 (corresponding to consideration 2) which are outlined in more detail in Figure 1. Both methods were developed on two separate streams to discriminate between GO terms associated with up-regulated and down-regulated genes, respectively. In this regard, the procedure returned a total of four lists of significant GO terms split in two methods and subsequently in two modalities (up-regulated and down-regulated). The output lists of significant GO terms obtained were analysed on a textual content level and compared modality-wise with the aim of extracting text-based insights that could guide the reader searching GO terms relevant to the case study and discriminating between the two methods at a glance. In these regards, a graphical technique was developed based on word clouds for visual representation and GSEA rankings for computing single-word significance. Specifically, the technique was based on the R packages `wordcloud` [28] and tm [29]. Its main steps are outlined in Figure 7.

The word clouds and the correlation plots were used to inspect the lists of significant GO terms returned by the two methods, compare the different results and carry out in-depth analyses on a specific subset of labels. In this regard, significant GO terms - and hence biological processes - plausibly correlated with CML have been selected and further examined at a single-gene level to implement the following objectives.

1. To compare the informative content of the labels to optimize the identification of genes relevant to CML. More generic and high-level labels were discarded in favour of more CML-specific ones.
2. To extend the analysis from TF-DEGs to their partners in the gene networks to gain biological insights on gene-gene interactions and better understand the impact of the treatment on the network topology.

## 4. Validation

Genes with adjusted p-value $< 0.05$ and $|\text{log2FoldChange}| > 0.1$ were considered to be differentially expressed. Based on these criteria, 384 genes have been selected as DEGs: 115 were up-regulated and 269 were down-regulated (see Figure 4).

In the set of genes scoring positively to the fold change test, we have identified 43 differentially expressed transcription factors (TF-DEGs), of which 24 are down-regulated and 19 are up-regulated (see Figure 1). The top 5 down-regulated TF-DEGs are ARNTL2, ZNF563, KLF7, TRPS1 and LHX2 while the top 5 up-regulated are ZNF90, ZNF492, HOXD9, MECP2 and. We then proceeded to validation of gene expression by quantitative RT-PCR on an independent set of samples. We selected a group of up and down-regulated genes based on microarray data and performed QPCR validation. Figure 2 shows the results of the analysis providing confirmation of the microarray analysis.

DEA and GSEA performed on the TF-DEGs bring to our attention a set of up and down-regulated genes that become part of a complex network reflecting on cell phenotype. Transcription factors recognize and bind to consensus sequence elements that are specific for each transcription factor, and the transcription factors then regulate downstream gene expression. We then proceeded to evaluate the phenotypic consequences of this regulation and focused on the upregulation of RARG, a gene belonging to the nuclear receptor superfamily, sharing 90% homology with retinoic acid receptor $\alpha$ (RAR$\alpha$) and retinoic acid receptor $\beta$ (RAR$\beta$), appears crucial for hematopoietic development [30] and erythroid differentiation program. Indeed Walkley et al. [30], showed that RAR$\gamma$ null mice exhibit a considerable increase in granulocytes in the peripheral blood (PB), in the bone marrow (BM) and spleen, developing a myeloproliferative-like syndrome and displaying reduction in the megakaryocyte– erythroid progenitor fraction thus altering homeostatic bone marrow erythropoiesis. Although the effect in mice appear to be the result of an erythroid cell extrinsic role (i.e. alteration of bone marrow microenvironment), a role in stress erythropoiesis or non-homeostatic erythroid demand was not excluded [30,31]. Therefore, upregulation of RARG might imply an increased propensity to erythroid differentiation in hemopoietic cells, a prediction that we verified and confirmed in the same cell model (Figure 5). As we noticed that, starting from 0.125$\mu$M IMA, the cells seem to have reached the maximum capability to produce Haemoglobin, we decided to pool these data and perform statistical analyses, such as an estimation plot (Figure 4). The statistical analyses confirmed the change in the differentiation program toward erythroid lineage . Furthermore CD36, expressed by committed erythroid progenitors that is expressed continuously on normal immature erythroblasts [32] appears one of the genes more strongly upregulated, further suggesting erythroid differentiation is modulated by PTPRG expression. We confirmaed CD36 upregulation both at the mRNA (Figure 2) and protein level in both resting condition and upon overnight treatment with IMA 5$\mu$M (Figure 3).

This relevance of this gene to CML has been recently supported by several studies performed in patients [15,16,20,21,33] and strategies aimed at restoring its expression are expected to benefit the course of the disease by improving drug efficacy or contrasting the emergence of BCR/ABL1 mutants.

## 5. Identification of molecular pathways

As a result of GSEA, the TF-DEGs identified by Method 1 and 2 are represented in Figure 6. The TF-DEGs associated to GO terms stemming from Method 2 are contained in the set returned by Method 1. Figure 8 shows the bar plots of the distributions of p-values returned by Method 1 and 2. The plots show significant differences between the two methods: Method 1 returns larger sets of GO terms with low p-values frequencies (less than 6% for both up-regulated and down-regulated genes). On the other hand, Method 2 returns smaller sets of GO terms characterized by frequencies that reach up to 20%. In this regard, Unlike Method 1, Method 2 shows fewer significant GO terms distributed in more densely populated and separated clusters. In order to better inspect the differences between the two methods, we built and analysed the weighted word clouds from such lists of GO terms with a view to improving understanding about the differences between Method 1 and Method 2. Word clouds for up-regulated genes are shown in Figure 9. We note a clear distinction between the two plots based on both data quantity and content. For example, the chunk myeloid is contained in both word clouds in two different sizes:

- in the first case, it appears as a small-sized chunk among other terms that are in all likelihood connected to CML;
- in the second case, it is represented as a middle-sized chunk among terms that seem quite distant from the target.

Correlation analysis was then performed further to investigate the informative content of the word clouds. Figure 10 shows the results achieved on the top 10 most significant words. The chunk myeloid appears in Method 2's top 10 associated with the chunks regulation, differentiation and cell which in turn show other interesting associations. On the other hand, Method 1 shows interesting correlations for all the top 10 chunks even if the word myeloid is not among them. In this regard, further investigation was conducted by going through the list of GO terms and picking attractive labels based on the insights extracted from the word clouds.

Figure 11 shows the word clouds for down-regulated genes. In this case, the two plots show mainly content-based differences. In fact, both clouds are thick and almost equally distributed in terms of word sizes. Moreover, the most powerful words are mostly in common. Both clouds show words of potential interest for experimental analyses - even if with different sizes - as immune, transcription, myeloid, leukocyte, and growth. On the other hand, Method 2 appears to be more detailed than Method 1 since it shows additional specific chunks as apoptotic, hemopoiesis, hematopoietic, p53, chondrocyte, cytokine, stem, and differentiation.

Correlation analysis was then performed to better discriminate between the two word clouds. Figure 12 shows the results performed on the top 10 most significant words. We see that the majority of the top 10 words is shared between the two methods. The chunks regulation, process, negative, metabolic, compound, and biosynthetic are represented in both plots. Moreover, since we are analyzing biological processes related to down-regulated genes, it is interesting that both methods share the association negative - regulation. However, the main difference between the two methods relies on the associated words rather than on the most significant ones themselves. In fact, Method 1 shows interesting but high-level associations that bring attention to generic biological processes. On the contrary, Method 2 shows more detailed associations as differentiation - chondrocyte, leukocyte, myeloid, compound - phosphate-containing, and negative - transcription.

After examining the GO lists on a single-word level with the aim of highlighting key words and biological insights, we thereby selected specific GO terms which showed particular relevance to biological processes involved in CML onset and development. In this regard, the plots presented hereafter split the analysis on two levels. The level defined by the GO terms set that provide labels along with the enrichment score (transformed p-value) returned by GSEA, and (ii) the level of the TF-DEGs set, report for each label the associated TF-DEGs.

Figures 13 and 14 show the selected labels for up-regulated genes. The first plot of Figure 13 shows the selected labels returned by Method 1. GO terms result to be clustered as in Table 2: The second plot of Figure 13 shows the TF-DEGs associated with the above-mentioned GO terms. We note that are only three TF-DEGs: MECP2, involved in almost all the selected biological processes, NR2E1, associated with regulation of cell migration involved in sprouting angiogenesis, and RARG2, implicated in both cell growth and regulation of myeloid cell differentiation.

| Process | References | GO ID |
|---|---|---|
| chromatin | [34] | GO:0016569, GO:0034401, GO:0097549, GO:1905269, GO:0006342 |
| acetylation | [35] | GO:0006473, GO:0006475, GO:0018393, GO:0016573, GO:0018394 |
| acylation | [36] | GO:0043543 |
| amino acid | [37] | GO:0018193 |
| cell growth | | GO:0016049 |
| myeloid cell differentiation | [38] | G0:0045637 |
| angiogenesis | [39] | GO:0090049 |

**Table 2.** GO terms of the biological processes selected by up-regulated TF-DEGs returned by Method 1. We note that the majority of GO labels is associated to the terms chromatin and acetylation while there is only one GO label (G0:0045637) directly correlated to CML.

The first plot of Figure 14 shows the selected labels returned by Method 2. There are only two GO labels that show a connection with CML (see Table 3).

| Process | References | GO ID |
|---|---|---|
| regulation of myeloid cell differentiation | [38] | GO:0045637 |
| regulation of blood vessel endothelial cell migration | [40] | GO:0043535 |

**Table 3.** GO terms of the biological processes selected by up-regulated TF-DEGs returned by Method 2. We note that unlike Method 1 the selection comprises just a few terms. Furthermore, only one term (GO:0045637) appears to be strictly correlated to CML.

Even if the selection is very different from the one carried out in Method 1, we obtained that the TF-DEGs identified are the same. Furthermore, Method 2 detected both NR2E1 and MECP2 as members of GO:0043535.

Figures 15 and 16 show the selected labels for down-regulated genes. The first plot of Figure 15 shows the selected labels returned by Method 1. GO terms can be clustered as follows based on the key words used to carry out the selection reported in Table 4. The second plot shows the TF-DEGs associated with the GO terms listed in Table 4. In the plot we see that the two most significant GO terms are both referred to the genes SOX5 and SMAD1. Moreover, GO:0071560 is a direct child of GO:0071559 and hence it is more specific than the other. Secondly, we note that several genes are identified in the high-level label GO:0002376 immune system process. Among them only ZBTB16, IFI16, and BATF3 are associated with more specific terms.

The first plot of Figure 16 shows the selected labels returned by Method 2. The selection of GO terms reported in Table 5 is more relevant to CML than the one returned by Method 1 both regarding the number of labels and their specificity. On the other hand, the set of related TF-DEGs overlaps with the one from Method 1 except for the genes LHX2 (only in Method 2), TFAP2C and TRPS1 (only in Method 1). Moreover, it is possible to notice that the TF-DEGs are always associated with more than two labels. Hence, Method 2 proves to be more detailed not only in the variety of GO terms but also in the number of

| Process | References | GO ID |
|---|---|---|
| growth | | GO:00071559, GO:0071560 |
| differentiation | [38] | GO:0046637, GO:0006475, GO:0018393, GO:0016573, GO:0046632 |
| immune | | GO:0002376, GO:0002253 |
| kinase | [38] | GO:0007178 |
| epigenetic | [41] | GO:0040029 |
| endopeptidase | [42] | GO:2000117 |
| cell population | | G0:0045637 |

**Table 4.** GO terms of the biological processes selected by Method 1 for down-regulated genes. We note that most of the GO labels are associated to the term differentiation. Moreover, also the other terms comprised in the selection appear to be clearly correlated with CML.

associated TF-DEGs. Regarding the genes left out by Method 1 we note that both TFAP2C and LHX2 were associated only to GO:00040029 regulation of gene expression, epigenetic, which results to be a high-level label. On the contrary, the new entry TRPS1 is associated to both GO:0002062 and GO:0032330, i.e. to chondrocyte differentiation. This means that Method 2 opted again in favour of a more CML-specific connotation.

| Process | References | GO ID |
|---|---|---|
| leukocyte | [43] | GO:0002521, GO:1902105, GO:0045321, GO:0002573 |
| immune | | GO:0002376, GO:0006955, GO:0002520, GO:0002550 |
| chondrocyte | [44] | GO:0002062, GO:0032330 |
| p53 | [45] | GO:0072331 |
| myeloid | [38] | GO:0030099, GO:0002573 |
| growth | | GO:0071560, GO:0071559 |
| differentiation | [38] | GO:0002521, GO:1902105, GO:0030099, GO:0002573, GO:0032330 |
| hemopoiesis | [46] | GO:0030097 |
| hematopoietic | [46] | GO:0048534 |
| cytokine | [47] | GO:0001816, GO:0001817 |
| phosphorylation | [47] | GO:0006468 |
| stem | [48] | GO:0098722, GO:0008356 |

**Table 5.** GO terms of the biological processes selected by Method 2. We note that the terms differentiation, immune and leukocyte are the most significant as far as number of associated GO labels. Moreover, lalso the other terms comprised in the selection appear to be strictly correlated to CML. In this case the correlation is clearly higher than the one returned by Method 1. This is due to the fact that the key terms comprised in the selection are referred to more specific biological processes involved in CML.

## 6. Discussion

We analyzed the modulation of CML cell model K562 transcriptome following transfection with the tumor suppressor gene PTPRG and the treatment with the tyrosine kinase inhibitor (TKI) Imatinib with the aim at identifying genes responding to the PTPRG modulation and/or treatments with Imatinib.

We developed two GSEA-based computational methods, Method 1 and Method 2, aimed at detecting all the CML-related differentially expressed transcription factors (TF-DEGs) and the biological processes involved. To summarize, the genes responsive to the treatments found by our methods are:

- Method 1:

  - Up-regulated TF-DEGs: MECP2, NR2E1, RARG;
  - Down-regulated TF-DEGs: ZBTB16, TFAP2C, SOX5, SMAD1, LHX2, IKZF3, IFI16, EPAS1, BATF3, BACH2;
- Method 2:
  - Up-regulated TF-DEGs: MECP2, NR2E1, RARG;
  - Down-regulated TF-DEGs: ZBTB16, TRPS1, SOX5, SMAD1, IKZF3, IFI16, EPAS1, BATF3, BACH2.

Method 1 was designed to take as input the whole list of DEGs stemming from DEA and afterwards select only the TF-DEGs. On the other hand, Method 2 was set to filter out only TF-DEGs identifying a smaller gene universe than Method 1. Moreover, the two methods were split in two modalities to discern between up-regulated and down-regulated genes. We observed that Method 1 returned more GO labels than Method 2 in both modalities. However, this entailed different outcomes for up-regulated and down-regulated TF-DEGs respectively. In fact, both the word clouds and the correlation analysis showed that for up-regulated TF-DEGs Method 1 returned appropriate and specific GO labels while Method 2 provided more general results. Nevertheless, the selections of CML-related TF-DEGs stemming from key term analysis identified the same list of genes for both methods. Hence, we could say that in this case Method 1 appears to be more appropriate on the grounds that it identified more specific GO labels than Method 2. For down-regulated TF-DEGs, Method 1 provided more high-level biological insights at all stages (weighted word clouds, correlation analysis and key term selection) while Method 2 showed more specific references to CML-related biological processes. However, the final lists of CML-related TF-DEGs differ for only few genes (LHX2 only for Method 2, TFAP2D and TRPS1 only for Method 1). In this case Method 2 is to be preferred to Method 1.

In conclusion, the methods here presented offer a versatile exploratory computational approach to analyze and extract meaningful biological information. The study combines statistical tests for DEA and GSEA with human-curated contents (Gene Ontology), weighted word clouds, correlation analysis and key term selection, originally born in different application domains (as textual analysis). These methods could be potentially very useful and expressive also in the descriptive statistical analyses applied to gene biology.

Finally, we provide some future development of our analysis. The identification of genes responsive to pharmacological treatments is certainly not limited to the application of these exploratory methods focused mainly on the gene as a single entity and the quantitative characteristics of its activity (e.g., its expression level), but requires analyses relevant to the field of systems biology. Indeed, the past twenty years have seen a revolution in the volume and complexity of data generated in experiments and observations in the life sciences. With the increase in available data, the need for data management, integration, and analysis has become an increasingly important challenge. Biological knowledge is inherently complex and so cannot readily be integrated into existing databases of molecular data. Since more than twenty years ontologies provide a means of formalizing biological knowledge — for example, about genes, anatomy and phenotypes — in complex hierarchies that are composed of terms and rules [49]. An ontology is a formal way of representing knowledge in which concepts are described both by their meaning and relationship. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them. The use of ontologies began in the biological sciences around 1998 with the development of the Gene Ontology [50,51], which systematically summarizes current knowledge of gene products across a wide range of species. Since then, many other databases have been created to store biological information in ontological structures. We refer the reader to [51–53] for a comprehensive review of the existing ontologies databases.

Currently, ontology databases store the knowledge about the *static* structures of biological organisms, whereas the dynamic behaviors of biological processes have, for the past half-century, been captured in the mathematical language of physics-based simulation modelling [54]. To date, there have been only a few attempts to bridge the wealth of structural knowledge and the wealth of process knowledge, i.e., of the physico-chemical laws described by equations of dynamical models. D. Cook et al. [54] introduced the terms *bio-ontology* and *biosimulation* to indicate ontologies related to biological entities and simulation of physics-based mathematical models of biological systems dynamics.

D. Cook and co-authors showed that the semantics of biosimulation models could be expressed in a formal ontology that describes the entities, the properties, and the physical laws that are encoded in the mathematical equations of a simulation model. They introduced the Ontology of Physics for Biology (OPB) [55,56] based on systems dynamics and makes explicit the biophysical semantics of physics-based biosimulation models. OPB can be used as a reference knowledge resource for annotating variables and equations of models and for deriving computable modelling code. Therefore, the future direction of this study is the development of a methodology to bridge this gap, and link the semantics of biosimulation to the knowledge in structural bio-ontologies. A possible way to pursue this goal could be the analysis of gene networks resulting from the identification of TF-DEGs of interest. More specifically, we plan to choose TF-DEGs that seem to be involved in CML-related biological processes and expand the analysis on genes that interact with them. The types of relations between genes can be retrieved from various sources as partner or pathway databases. Here we relied on Pathway Commons, a pathway database that uses the Biological Pathway Exchange (BioPAX) [57] standard to represent data. It allows to investigate multiple biological concepts such as biochemical reactions; gene regulatory networks; genetic interactions; proteins, small molecules, DNA, RNA, complexes and their cellular locations; complex assembly and transport; post-translational protein modifications; citations; experimental evidence and links to other databases e.g. protein sequence annotation [58].

For our purposes we focused on two types of gene-gene relationships involving TF-DEGs:

- control of gene expression (one-way relationship): we analyzed all the genes in control or controlled by TF-DEGs in terms of expression levels.
- interaction between genes (two-way relationship): we analyzed all the genes that chemically interact with TF-DEGs.

Since the analysis on up-regulated genes returned the same set of relevant TF-DEGs, we focused only on down-regulated genes. Figures 17, 18, 19 and 20 show the analysis results for Method 1 and Method 2 as gene networks. In conclusion, we plan to investigate the biological and chemical relations between the genes represented in the networks to enrich the exploratory methods hereby defined with additional information about the network dynamics. The construction of the equations for the dynamics of the gene networks of interest involves calibrating the model as the next step. In possession only of static data, such as those used in this study, this phase will require the development of efficient sensitivity analysis techniques given the large number of genes potentially involved and the expected non-linear dynamics. In this regard, we plan to refine the numerical techniques fpr parameter sensitivity analysis, inference and dynamic simulation developed in [59–61].

Another future research line to be furter explored is the identification and the analysis of the DEGs responsive to both the case study under examination and known pharmacological treatments with TKI. In this direction, we preliminarly performed DEA to detect DEGs between two groups: control considered as the untreated group (2 replicates) and the treatment group referred to TKI expressing samples (2 replicates). In order to discard background noise, only genes with intragroup standard deviation $< 0.3$ and distance between the group means $> 0.5$ were considered. Differential expression analysis was conducted on log2-trasformed gene expressions using the Bioconductor/R package limma (Version 3.12). Both the empirical Bayes correction on the variances and the multi-testing Bonferroni-Hochberg correction on p-values were selected. Therefore, genes with adjusted p-value $<$ 0.05 and $|\text{log2FoldChange}| > 0.1$ were considered to be differentially expressed. Based on these criteria, 568 genes have been selected as DEGs: 310 were up-regulated and 258 were down-regulated. Among them we have identified 61 transcription factors, of which 25 are down-regulated and 36 are up-regulated (see Figure 21). The top 5 down-regulated TFs are GATA3, RUNX3, HES1, TBX4, FOSL1 while the top 5 up-regulated are NPAS4, FOXN4, HOXA2, PURG, ZNF540. Then, we selected only the DEGs that occurred in both selections stemming from DEA. The results are represented in 22 and are split between up-regulated and down-regulated genes.

## 7. Conclusions

In this study we identified molecular pathways modulated by the tumour suppressor gene PTPRG and focused on transcription factors, known to act as master genes controlling a high number of downstream effectors. Future avenues might involve the identification of specific genes regulated by them and to investigate the biological and chemical relations between the genes represented in the networks to enrich the exploratory methods hereby defined with additional information about the network dynamics. The construction of the equations for the dynamics of the gene networks of interest involves calibrating the model as the next step. In possession only of static data, such as those used in this study, this phase will require the development of efficient sensitivity analysis techniques given the large number of genes potentially involved and the expected non-linear dynamics. In this regard, we plan to refine the numerical techniques developed in [59]. In conclusion, the methods here presented offer a versatile exploratory computational approach to analyse and extract meaningful biological information. The study combines statistical tests for DEA and GSEA with human-curated contents (Gene Ontology), weighted word clouds, correlation analysis and key term selection, originally born in different application domains (as textual analysis). Of note we have validated the microarray data using a group of differentially expressed genes and identified a cell differentiation program activated by the TSG PTPRG leading to a higher propensity of the blasts to differentiate toward a more mature phenotype, a condition that is further enhanced by TKI therapy. These data further support the relevance of re-expression of PTPRG in the context of CML suggesting it as a relevant therapeutic target. These methods could be potentially very useful and expressive also in the descriptive statistical analyses applied to gene biology.

## Figures

For the sake of clarity and order we collect in this section all the figures produced to show the application of the methods to the data and the obtained results.

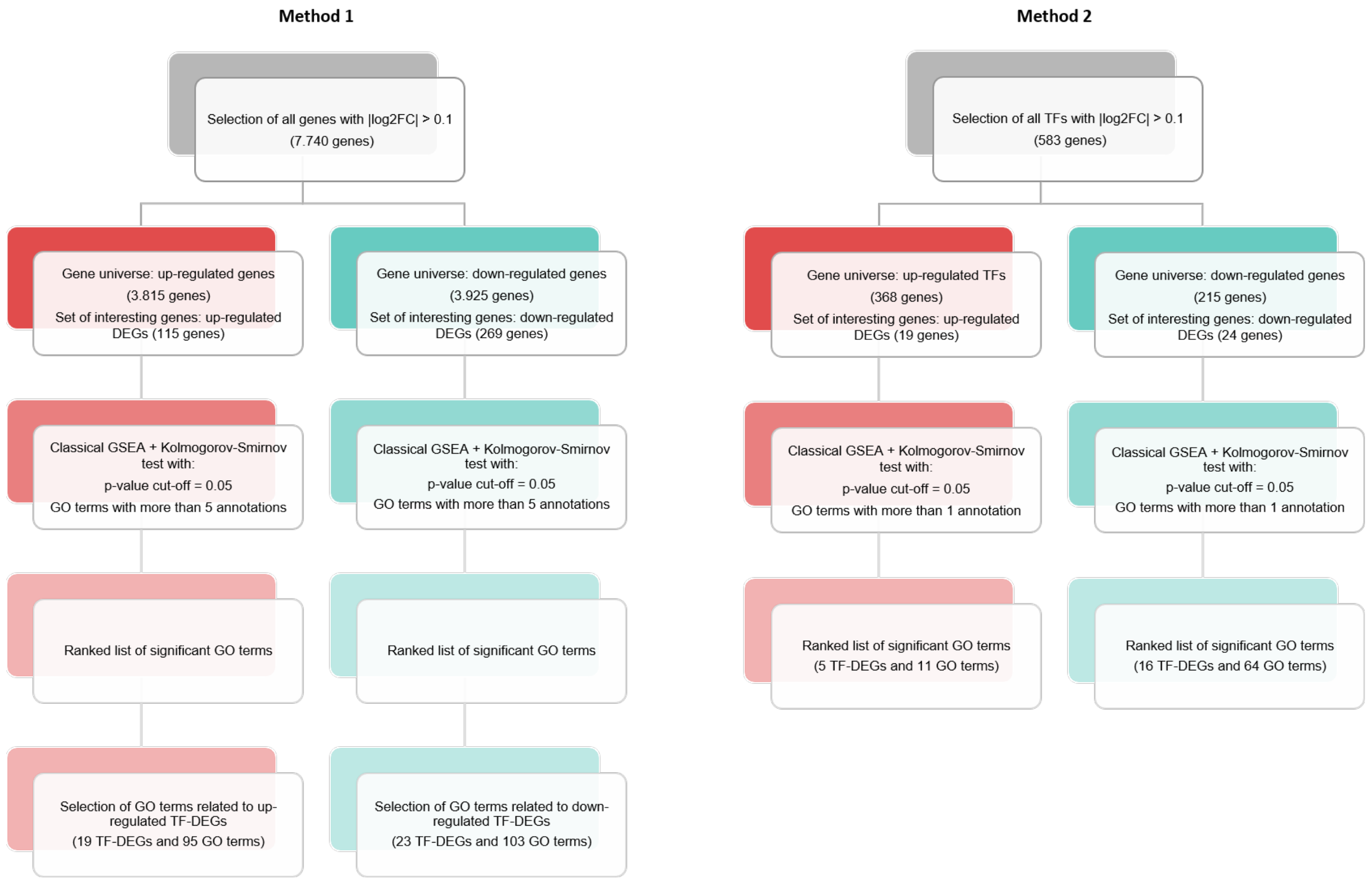

Method 1
Selection of all genes with |log2FC| > 0.1 (7.740 genes)
Gene universe: up-regulated genes (3.815 genes) Set of interesting genes: up-regulated DEGs (115 genes)
Gene universe: down-regulated genes (3.925 genes) Set of interesting genes: down-regulated DEGs (269 genes)
Classical GSEA + Kolmogorov-Smirnov test with: p-value cut-off = 0.05 GO terms with more than 5 annotations
Classical GSEA + Kolmogorov-Smirnov test with: p-value cut-off = 0.05 GO terms with more than 5 annotations
Ranked list of significant GO terms
Ranked list of significant GO terms
Selection of GO terms related to up-regulated TF-DEGs (19 TF-DEGs and 95 GO terms)
Selection of GO terms related to down-regulated TF-DEGs (23 TF-DEGs and 103 GO terms)
Method 2
Selection of all TFs with |log2FC| > 0.1 (583 genes)
Gene universe: up-regulated TFs (368 genes) Set of interesting genes: up-regulated DEGs (19 genes)
Gene universe: down-regulated genes (215 genes) Set of interesting genes: down-regulated DEGs (24 genes)
Classical GSEA + Kolmogorov-Smirnov test with: p-value cut-off = 0.05 GO terms with more than 1 annotation
Classical GSEA + Kolmogorov-Smirnov test with: p-value cut-off = 0.05 GO terms with more than 1 annotation
Ranked list of significant GO terms (5 TF-DEGs and 11 GO terms)
Ranked list of significant GO terms (16 TF-DEGs and 64 GO terms)

**Figure 1.** GSEA-based methods.

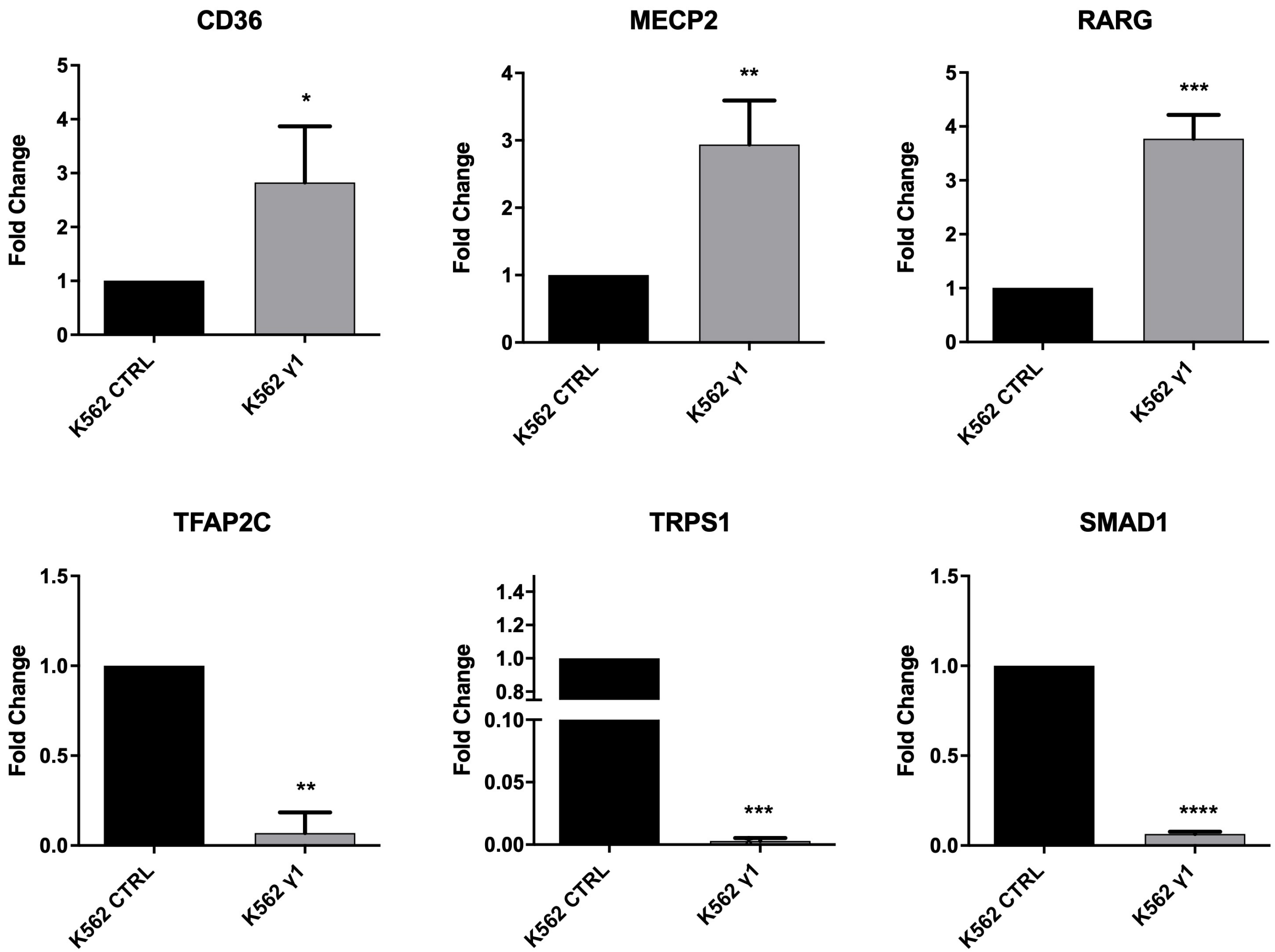


**Figure 2. mRNA level of selected up-regulated and down-regulated transcription factors (TF-DEGs) identified by Method 1 and 2.** The mRNA levels of transcription factors (TF-DEGs) were determined by qRT-PCR and the relative fold changes was calculated between K562 cell line expressing PTPRG and control group (control ∅ and D1028A). GAPDH was used as the endogenous control.

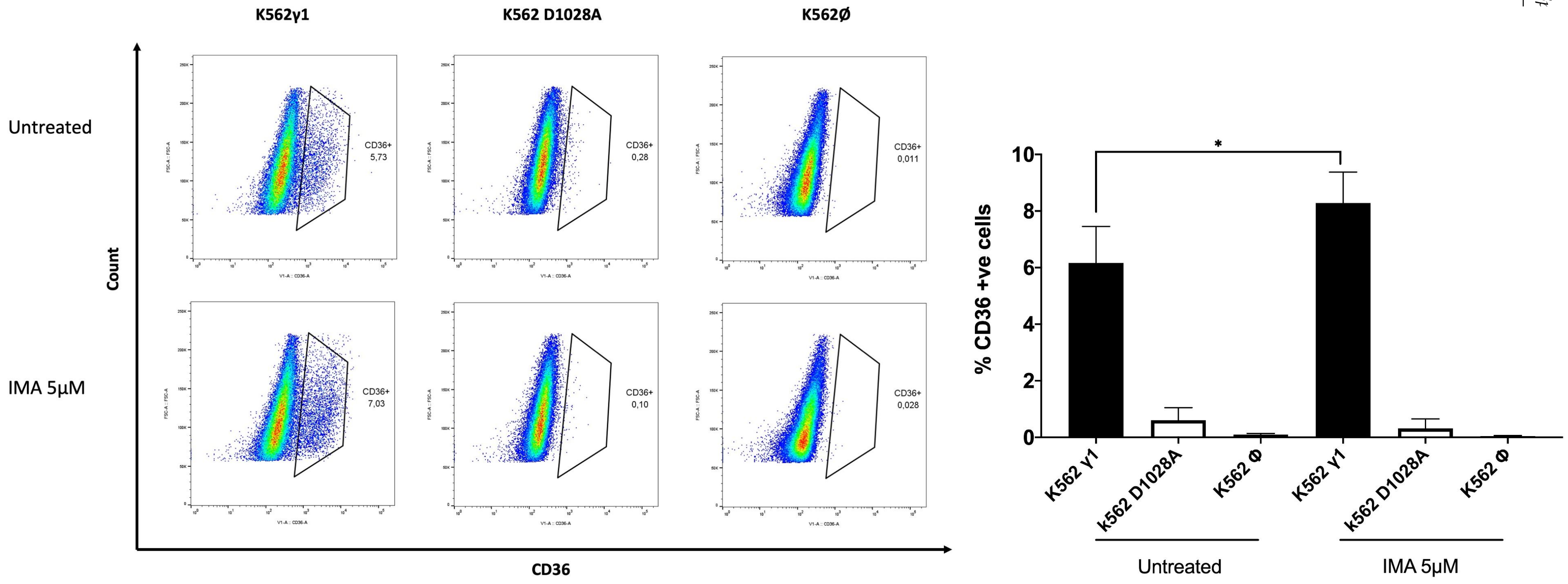


**Figure 3. Flow cytometry analysis of CD36+ surface marker on K562 subclones.** Left panel: Representative dot plot graphs show the increased expression of CD36 on the surface of the K562 cell lines expressing PTPRG $\gamma 1$ compared to the control group (control $\emptyset$ and D1028A) in pressence or absence of 5$\mu$M IMA overnight treatment. Right panel: summary of the results of a minimum of 3 experiments (p=0.03, one-tail t test).

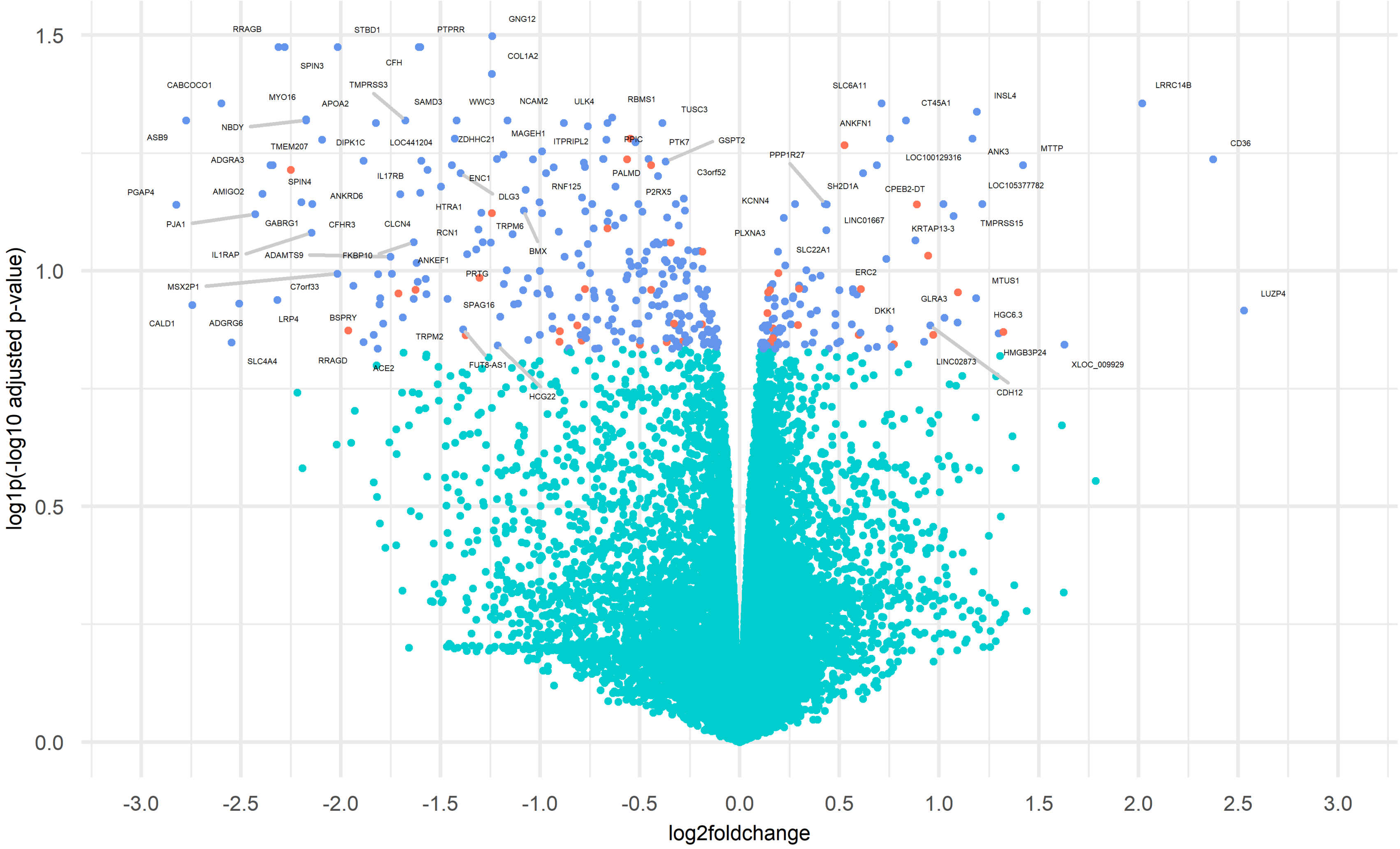


**Figure 4. Volcano plot for DEGs among controls and PTPRG overexpressing K562.** Genes are represented as scattered points: the x-axis is the log2FoldChange and the y-axis shows the log1p(-log10 adjusted p-value). Green dots represent the non-differentially expressed genes. Both red and blue dots represent genes that were identified as significantly differentially expressed (adjusted p-value $< 0.05$) with $|\text{log2FoldChange}| > 0.1$. Specifically, red dots are referred to transcription factors.

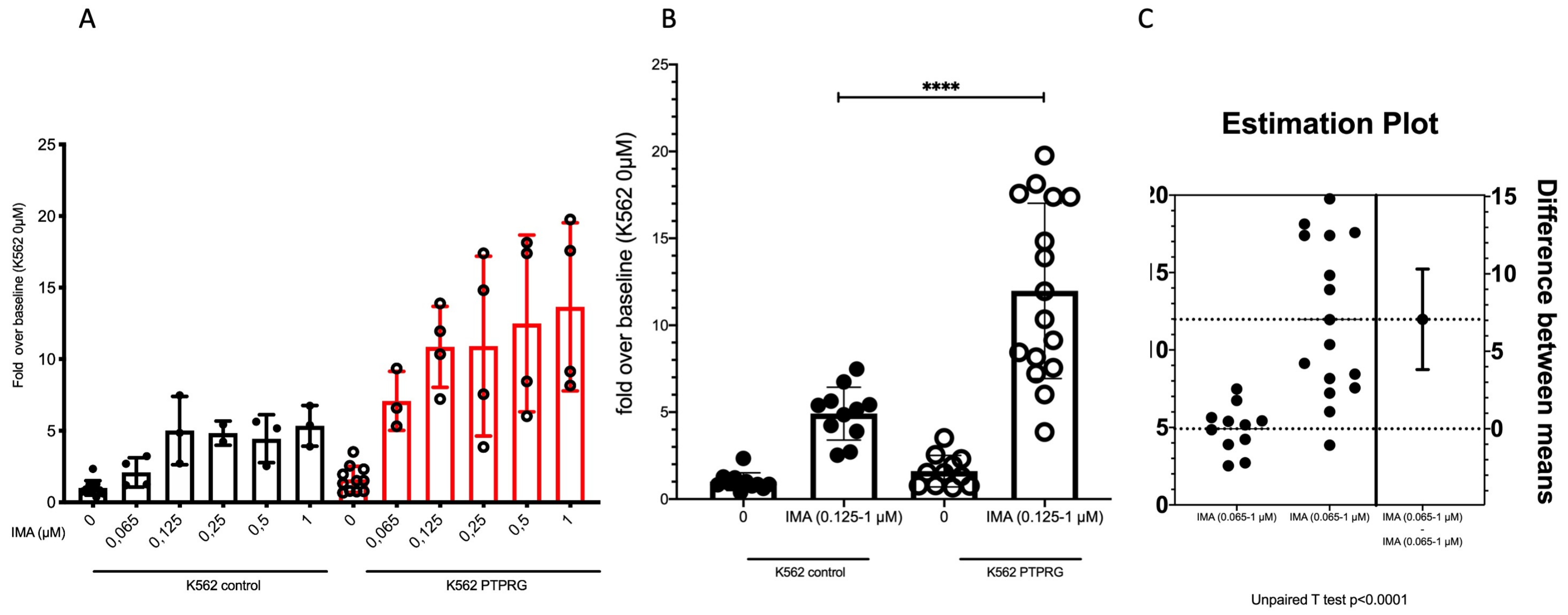


**Figure 5.** Upregulation of RARG implies an increased propensity to erythroid differentiation in hemopoietic cells. K562 were treated for 48hrs with the indicated concentrations of Imatinib (IMA). Cells were lysed and Hb content was evaluated (ng Hb$\mu$g total lysate) and expressed as fold increase over baseline (P< 0.0001 Mann-Whitney test).

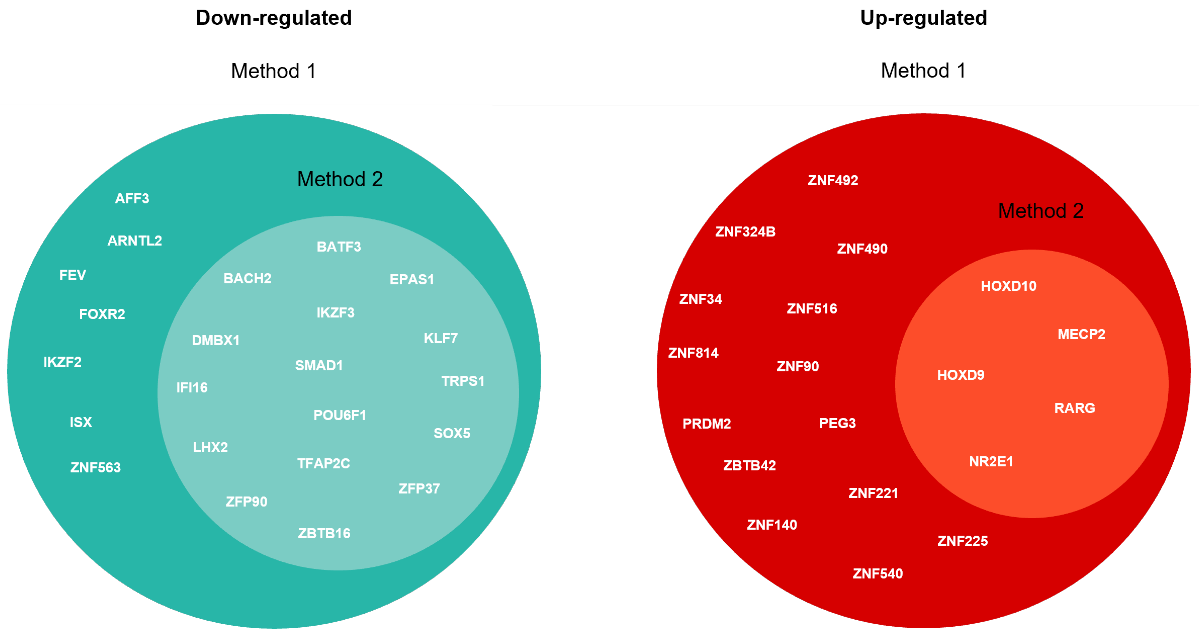
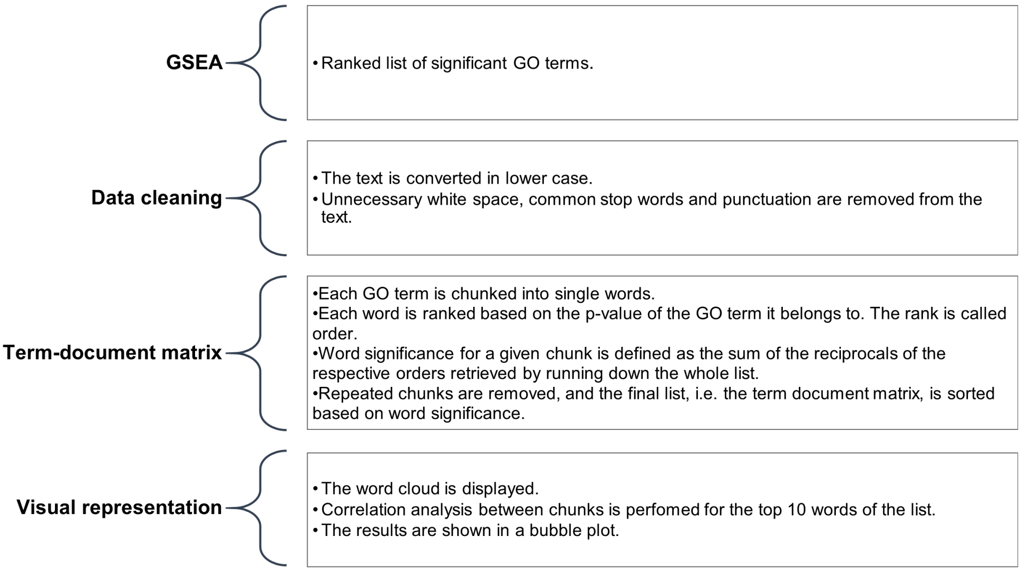


**Figure 6.** TF-DEGs detected by Method 1 and 2.



**GSEA**

- Ranked list of significant GO terms.

**Data cleaning**

- The text is converted in lower case.
- Unnecessary white space, common stop words and punctuation are removed from the text.

**Term-document matrix**

- Each GO term is chunked into single words.
- Each word is ranked based on the p-value of the GO term it belongs to. The rank is called order.
- Word significance for a given chunk is defined as the sum of the reciprocals of the respective orders retrieved by running down the whole list.
- Repeated chunks are removed, and the final list, i.e. the term document matrix, is sorted based on word significance.

**Visual representation**

- The word cloud is displayed.
- Correlation analysis between chunks is perfomed for the top 10 words of the list.
- The results are shown in a bubble plot.

**Figure 7. Main tasks of GSEA-based text mining algorithm.** The proceeding here represented was applied to all the GO lists returned by Method 1 and 2.

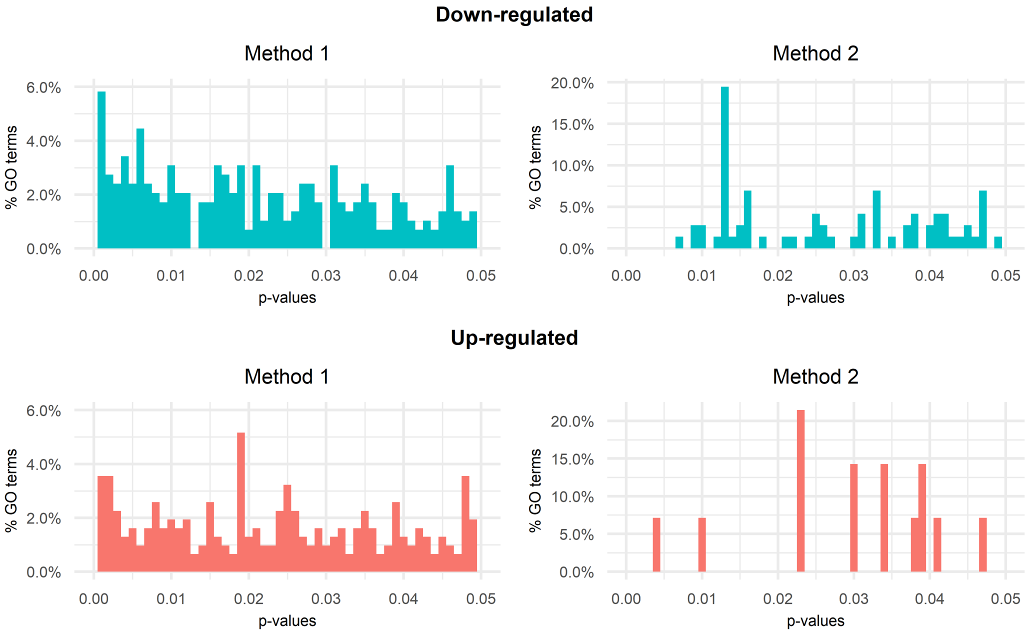


**Figure 8. Bar plots of GSEA p-values.**

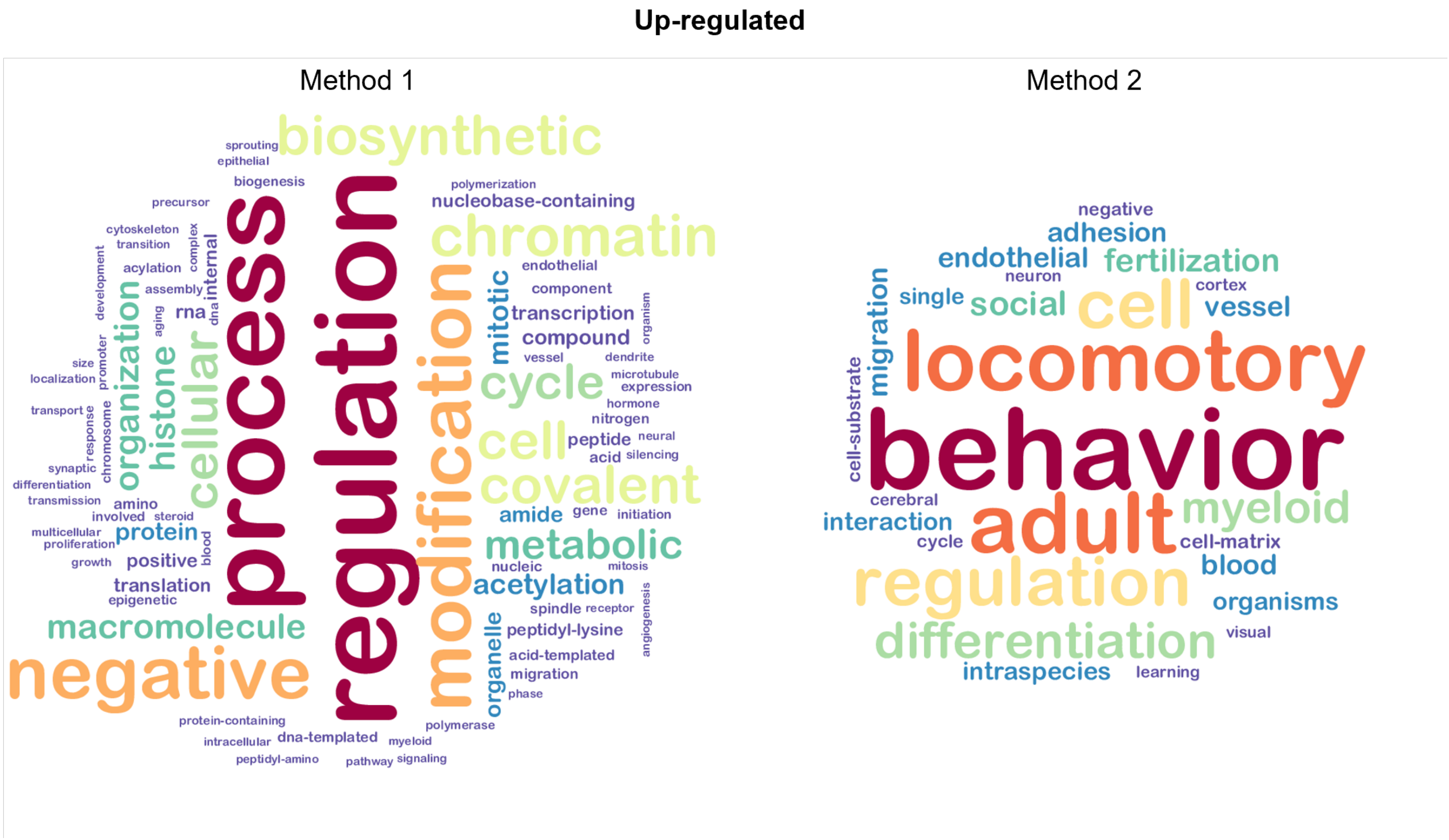


**Figure 9. Word clouds for GO terms related to up-regulated genes.** The word clouds show the top 100 words retrieved using the procedure described in Figure 7.

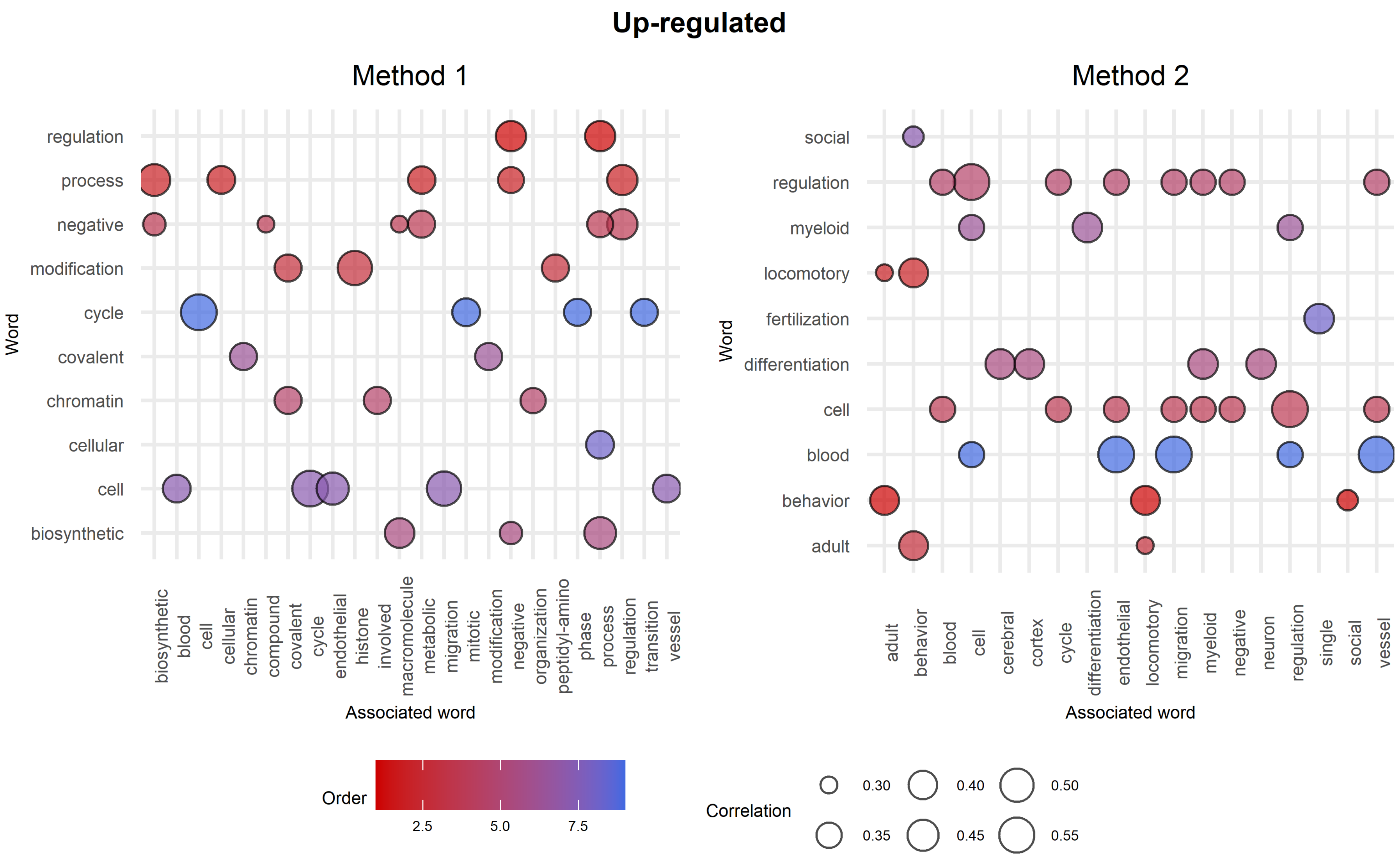


**Figure 10. Correlation analysis for the top 10 words.** The top 10 most significant words are represented on the y-axis while their respective associated words are on the x-axis. The colour of the bubbles is based on the order (significance of a given GO term as described in Figure 7) whilst the size depends on the correlation between words. Note that if two top 10 words A and B are associated to each other then the plot shows both the pairs (A,B) and (B,A).

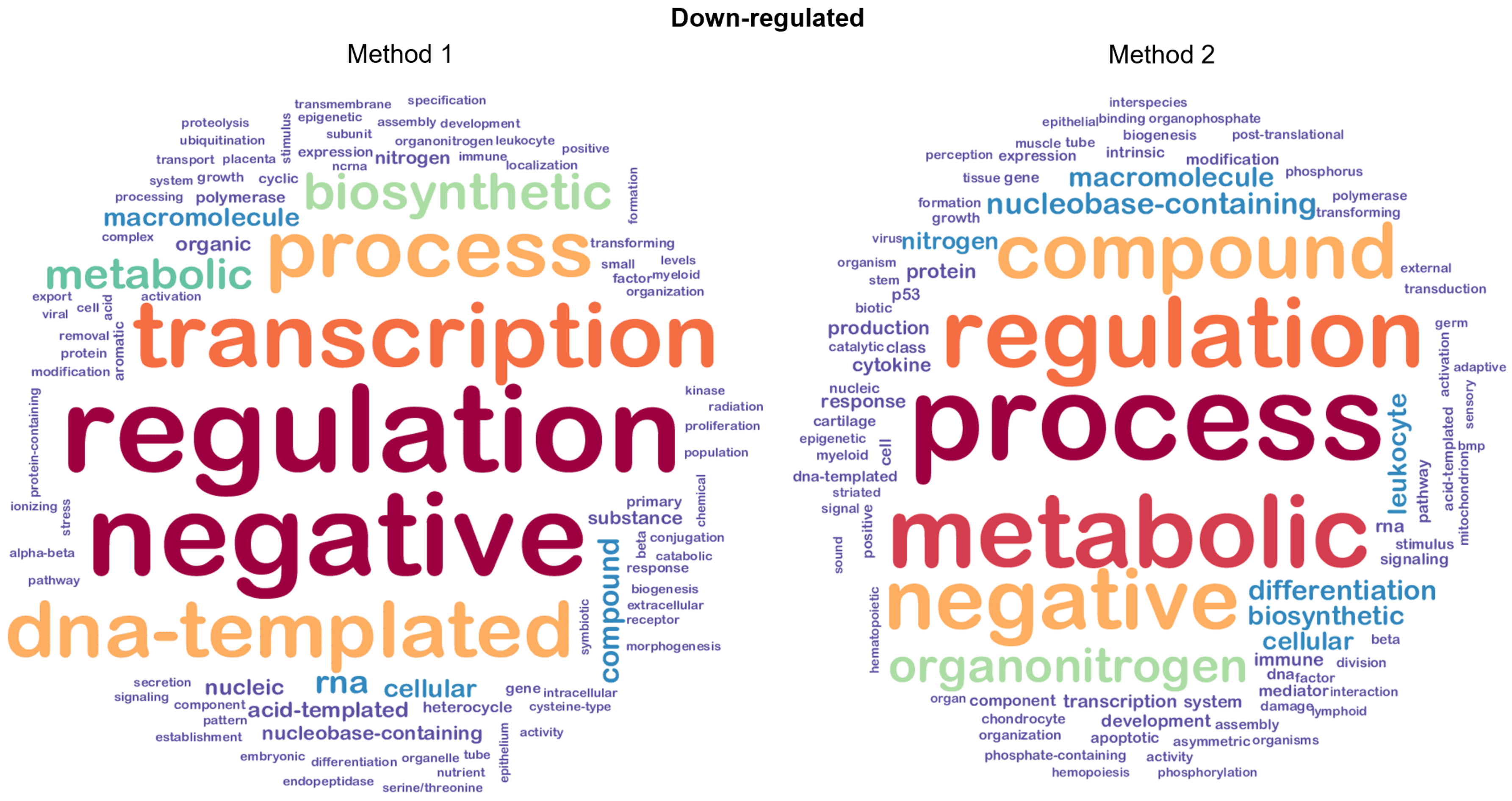


**Figure 11. Word clouds for GO terms related to down-regulated genes**. The word clouds show the top 100 words retrieved using the procedure described in Figure 7). Both single word size and color depend on word significance.

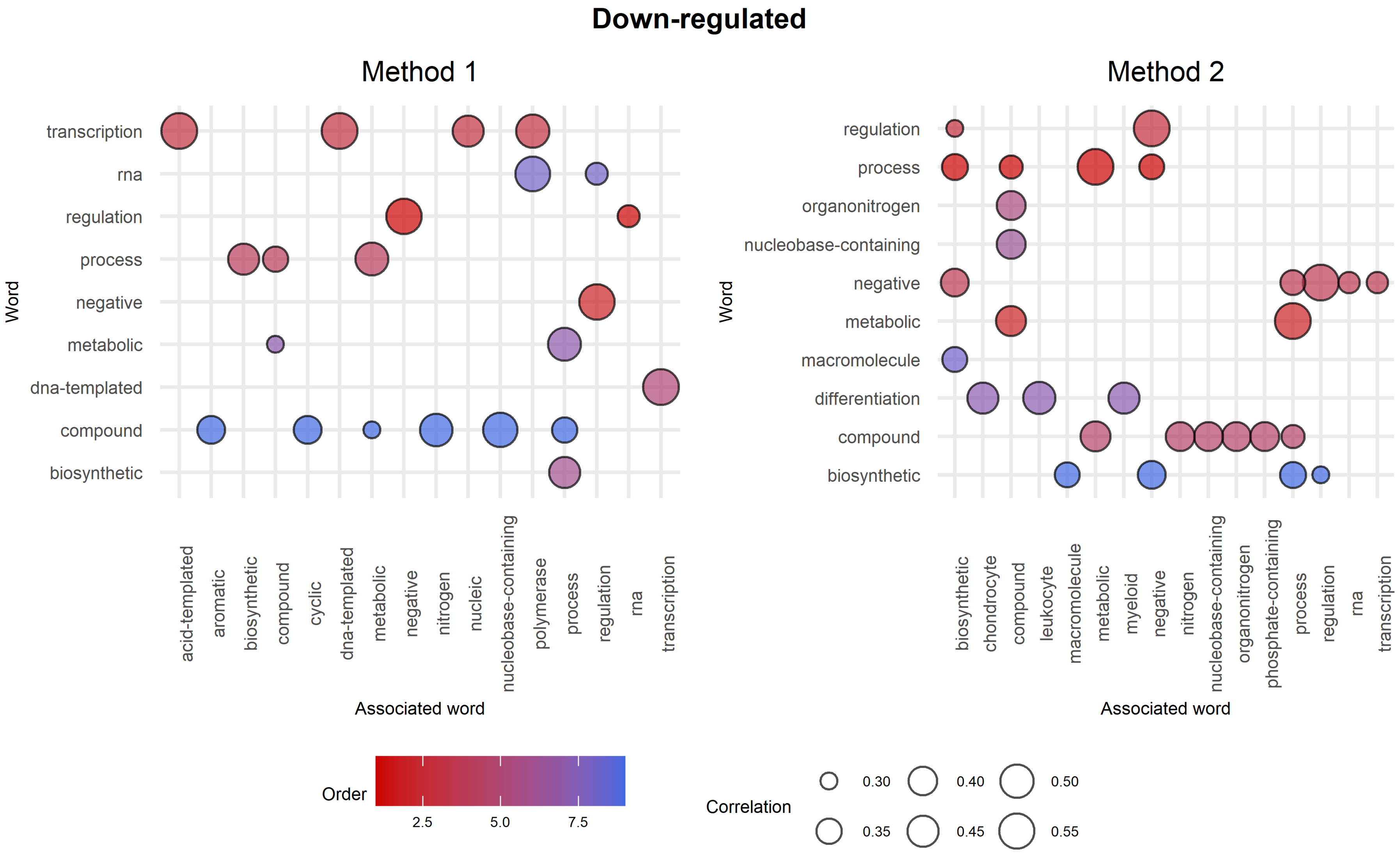


**Figure 12. Correlation analysis for the top 10 words.** The top 10 most significant words are represented on the y-axis and their respective associated words are on the x-axis. The colour of the bubbles is based on the order (significance of a given GO term as described in Figure 7 whilst the size depends on the correlation between words. Note that if two top 10 words A and B are associated to each other then the plot shows both the pairs (A,B) and (B,A).

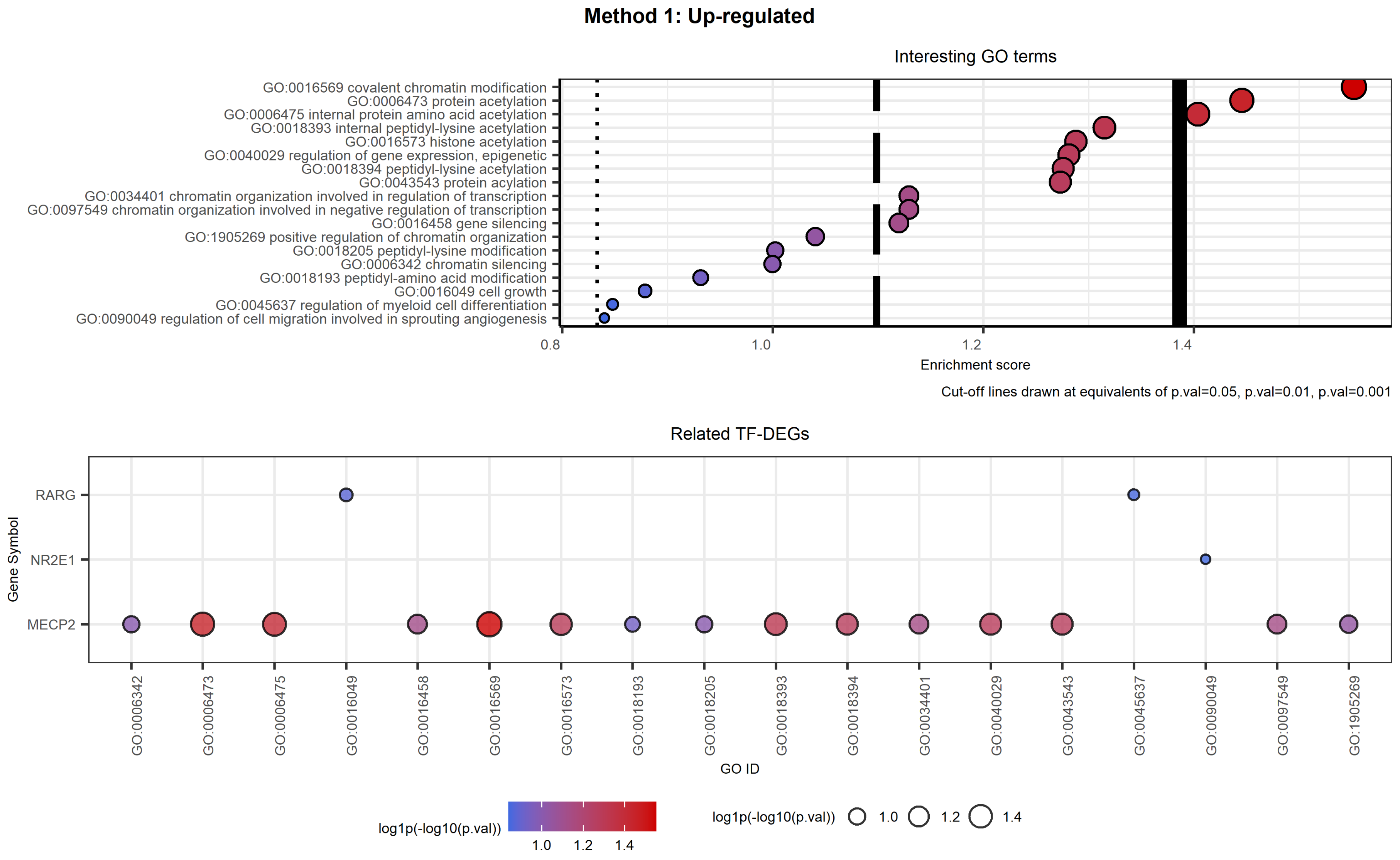


**Figure 13. Interesting GO terms and related TF-DEGs.** The first plot shows selected GO terms on the y-axis and the respective log1p(-log10(p-value)) as the Enrichment score on the x-axis. The second plot shows on the y-axis the TF-DEGs and the related GO terms on the x-axis. In both plots color and size of the bubbles depend on the Enrichment score. We note that gene MECP2 is involved in almost all the selected GO labels while genes RARG and NR2E1 results to be specific.

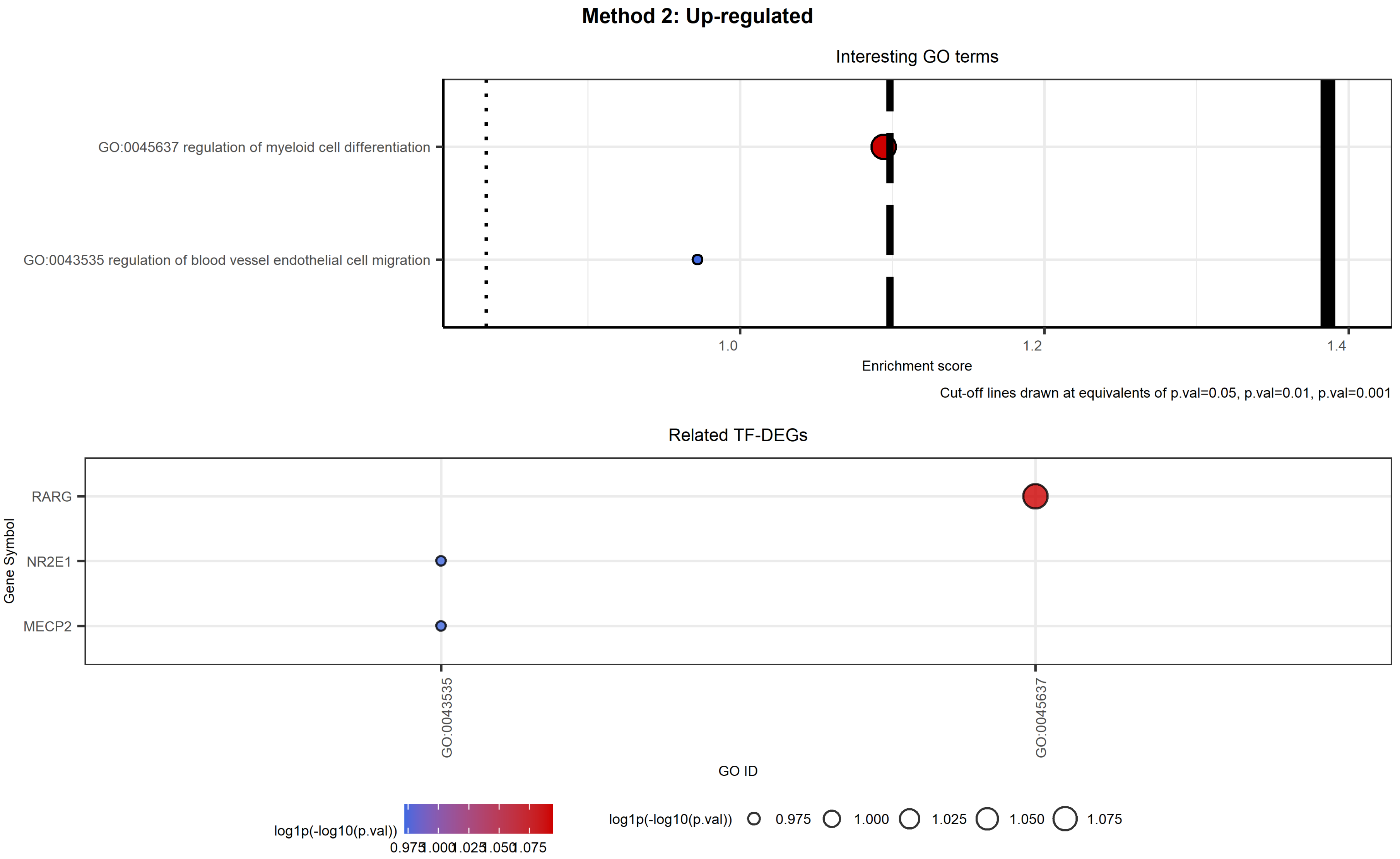


**Figure 14. Interesting GO terms and related TF-DEGs.** The first plot shows selected GO terms on the y-axis and the respective log1p(-log10(p-value)) as the Enrichment score on the x-axis. The second plot shows TF-DEGs on the y-axis and the related GO terms on the x-axis. In both plots colour and size of the bubbles depend on the Enrichment score. We note that gene RARG result to be strictly CML-related while genes MECP2 and NR2E1 are associated to a more general biological process.

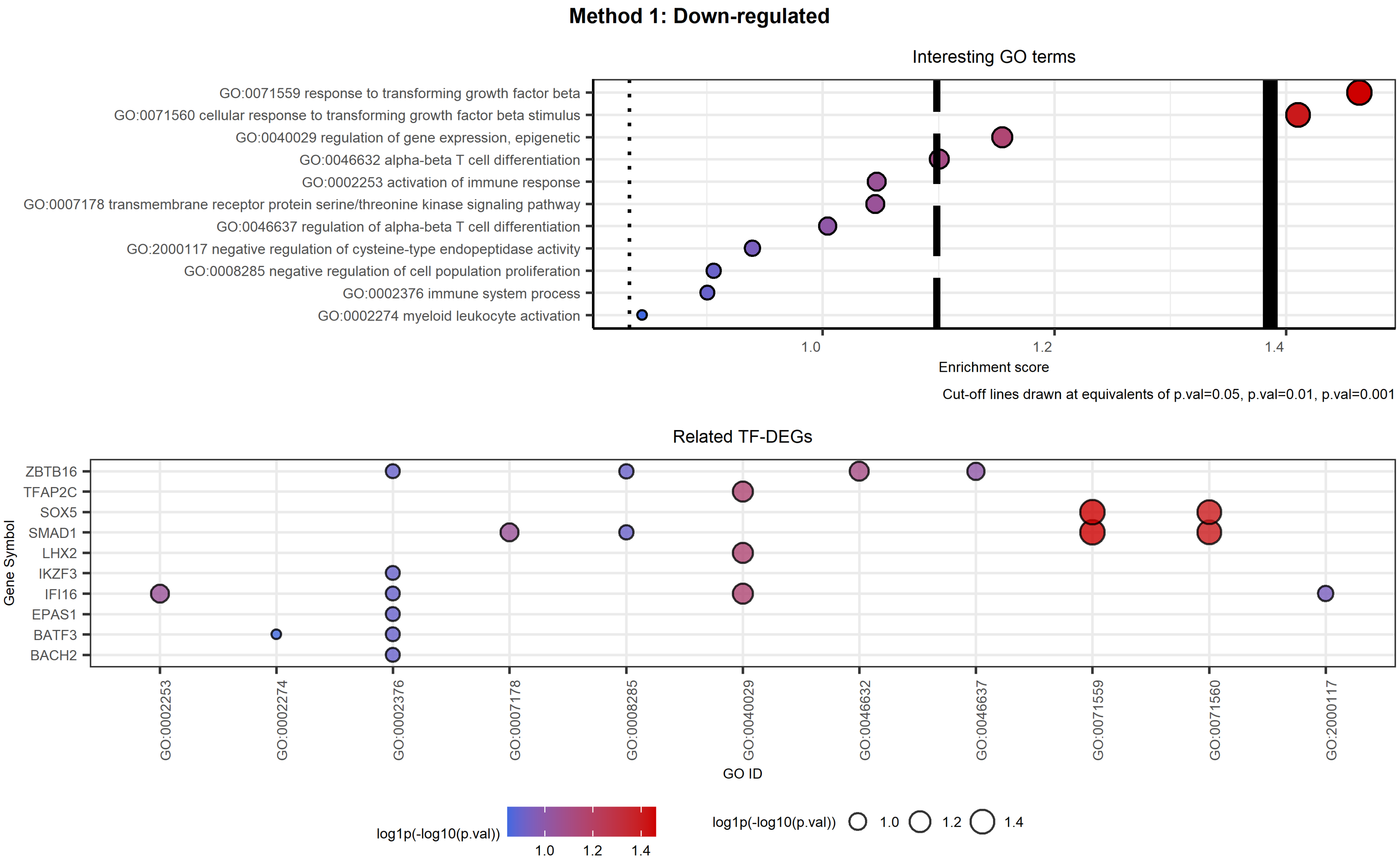


**Figure 15. Selected GO terms and related TF-DEGs.** The first plot shows selected GO terms on the y-axis and the respective log1p(-log10(p-value)) as the Enrichment score on the x-axis. The second plot shows TF-DEGs on the y-axis and the related GO terms on the x-axis. In both plots color and size of the bubbles depend on the Enrichment score. We note that genes SOX5 and SMAD1 are associated to the most significant GO labels which results to be associated with cell growth.

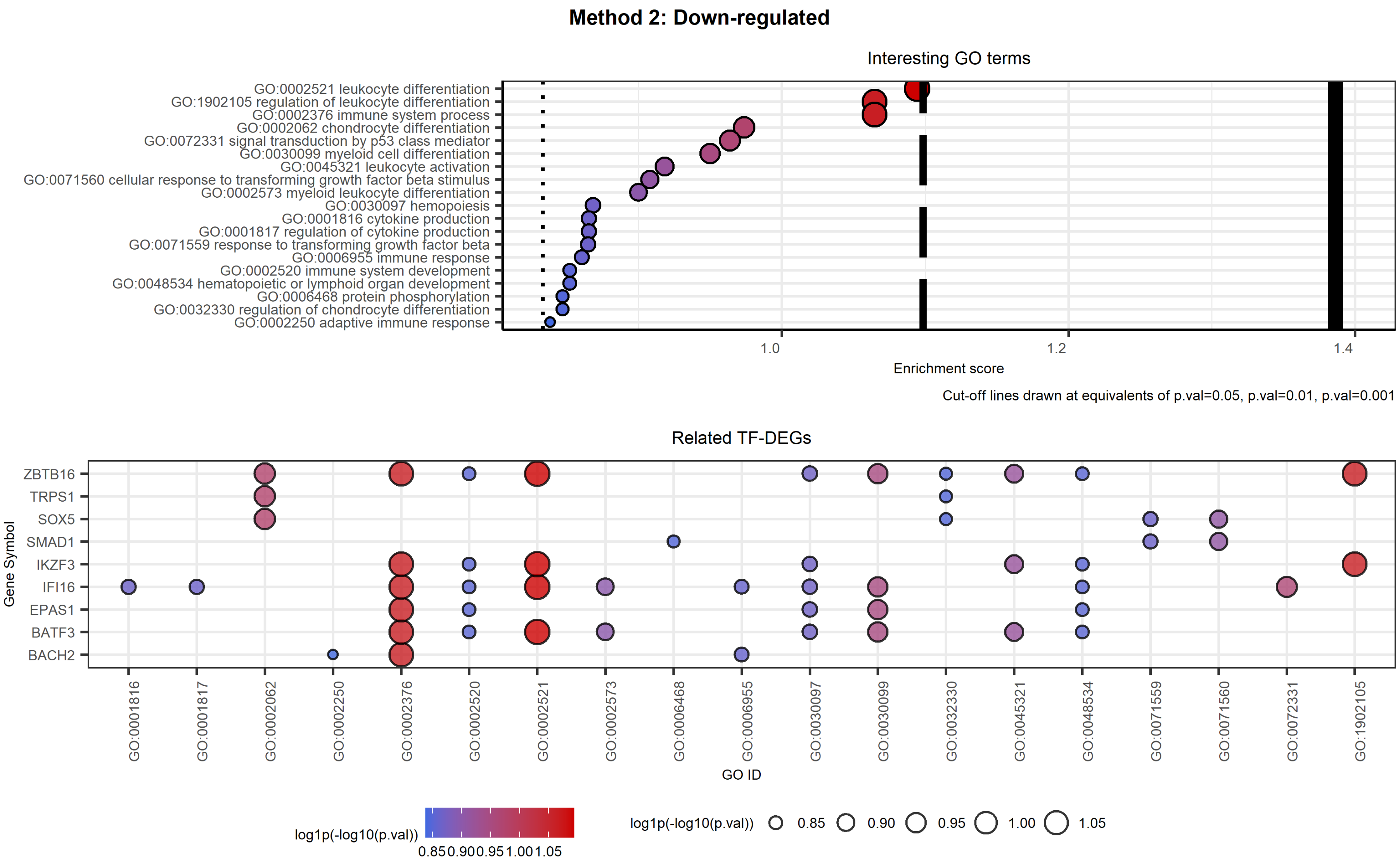


**Figure 16. Selected GO terms and related TF-DEGs.** The first plot shows selected GO terms on the y-axis and the respective log1p(-log10(p-value)) as the Enrichment score on the x-axis. The second plot shows TF-DEGs on the y-axis and the related GO terms on the x-axis. In both plots color and size of the bubbles depend on the Enrichment score. We note that both genes MECP2 and NR2E1 are involved in the same process while RARG is identified as strictly CML-related. We note that almost all genes are associated to the most significant GO label involved in leukocyte differentiation.

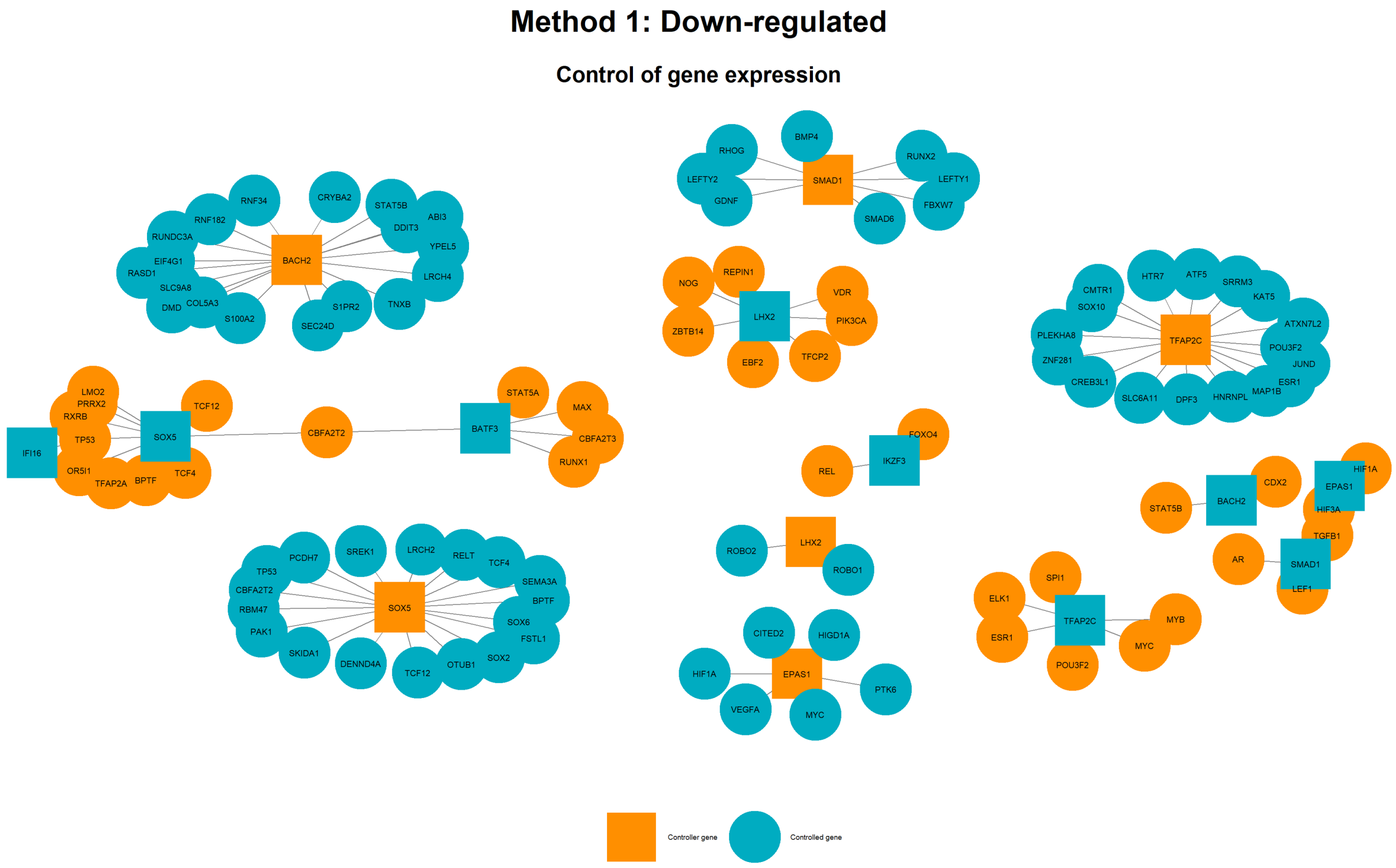


**Figure 17. Method 1: genes correlated to TF-DEGs by the function "control of gene expression".** TF-DEGs are square-shaped while gene partners are represented by circles. Orange labels are referred to genes that act as controllers while cyan is assigned to genes that are controlled by others. We note that the nets originating from IFI16, SOX5 and BATF3 are interconnected through some common genes.

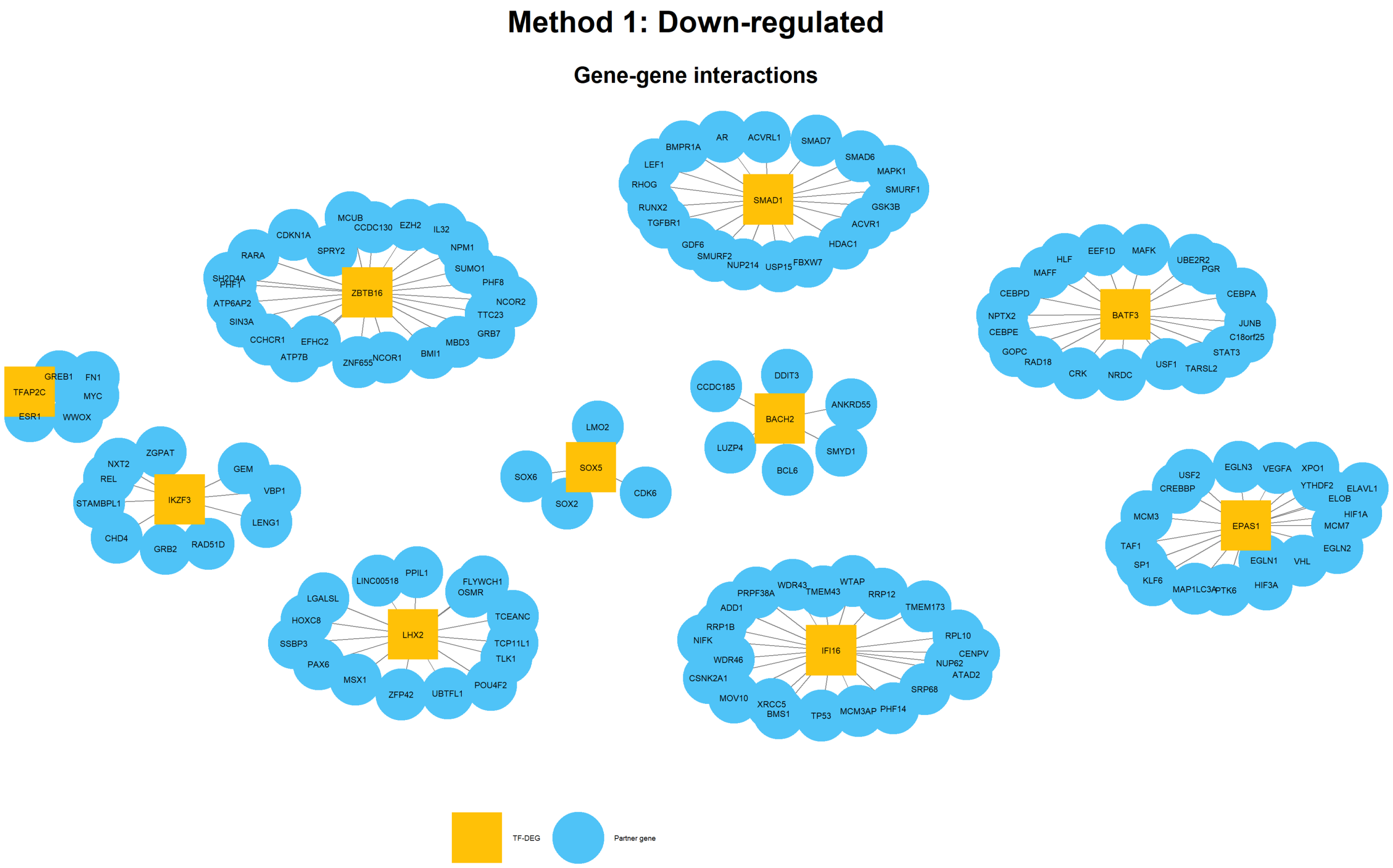


**Figure 18. Method 1: genes that interact with TF-DEGs.** TF-DEGs are square-shaped while gene partners are represented by circles. Yellow labels are referred to TF-DEGs while light blue is assigned to their partners. We note that there are no interconnected nets.

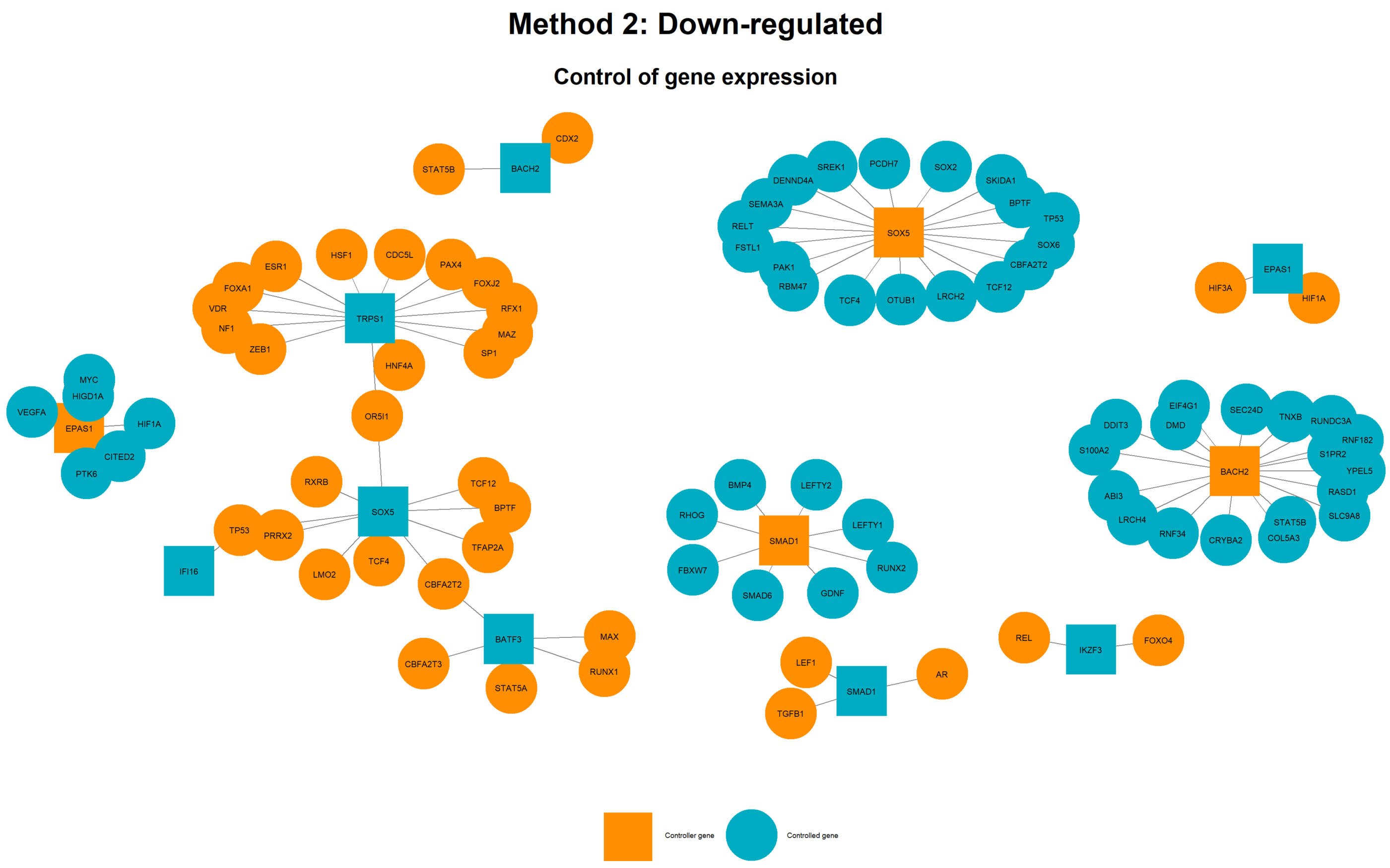


**Figure 19.** **Method 2: genes associated to TF-DEGs by a relationship of the kind "control of gene expression** The plot shows all the genes associated to TF-DEGs by a relationship of the kind "control of gene expression". TF-DEGs are square-shaped while gene partners are represented by circles. Orange labels are referred to genes that act as controllers while cyan is assigned to those genes that are controlled by others. We note that the nets originating from TRPS1, SOX5, IFI16 and BATF3 are interconnected through some common genes.

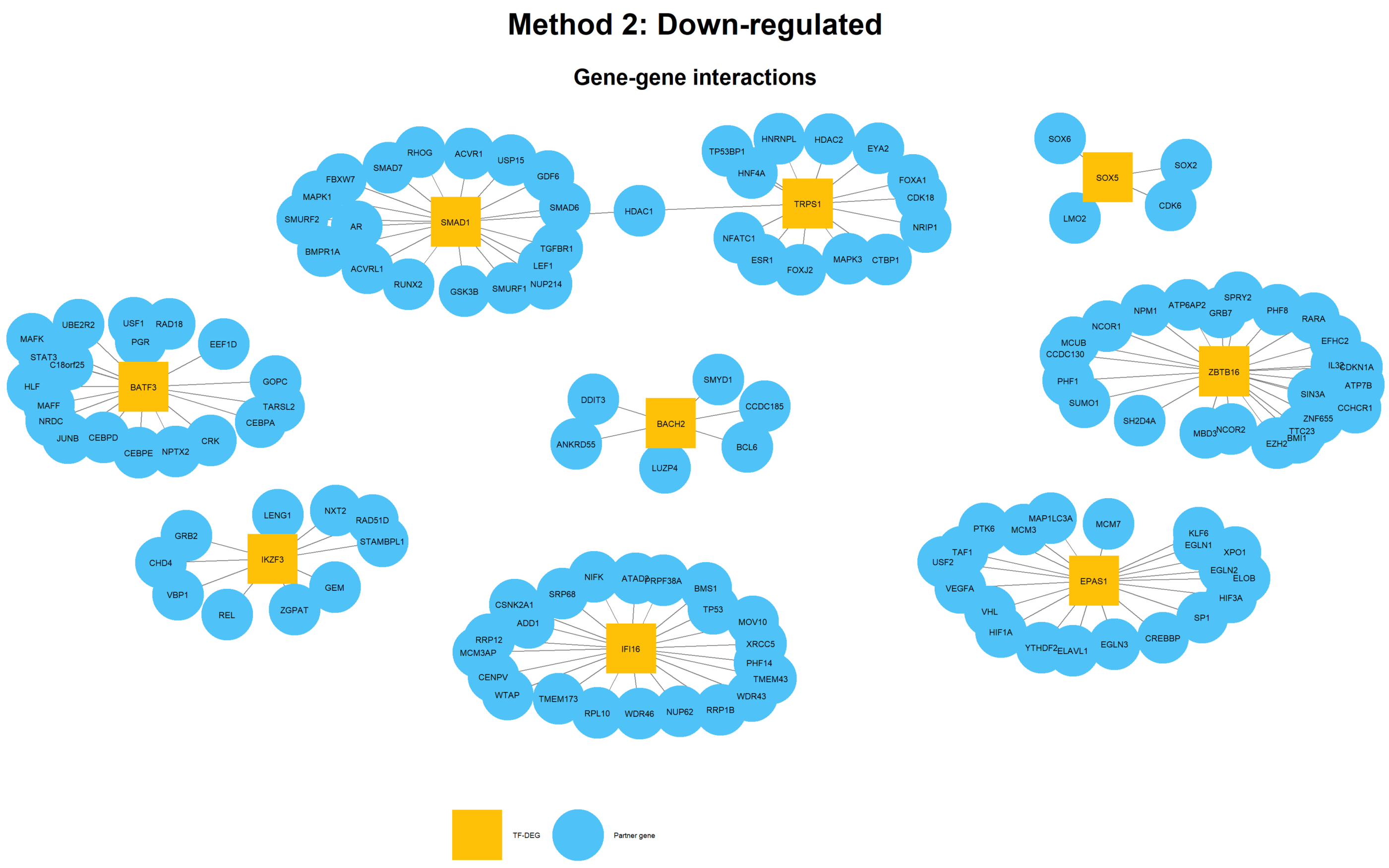


**Figure 20. Method 2: genes that interact with TF-DEGs.** TF-DEGs are square-shaped while gene partners are represented by circles. Yellow labels are referred to TF-DEGs while light blue is assigned to their partners. We note that the nets originating from SMAD1 and TRPS1 are interconnected through gene HDAC1.

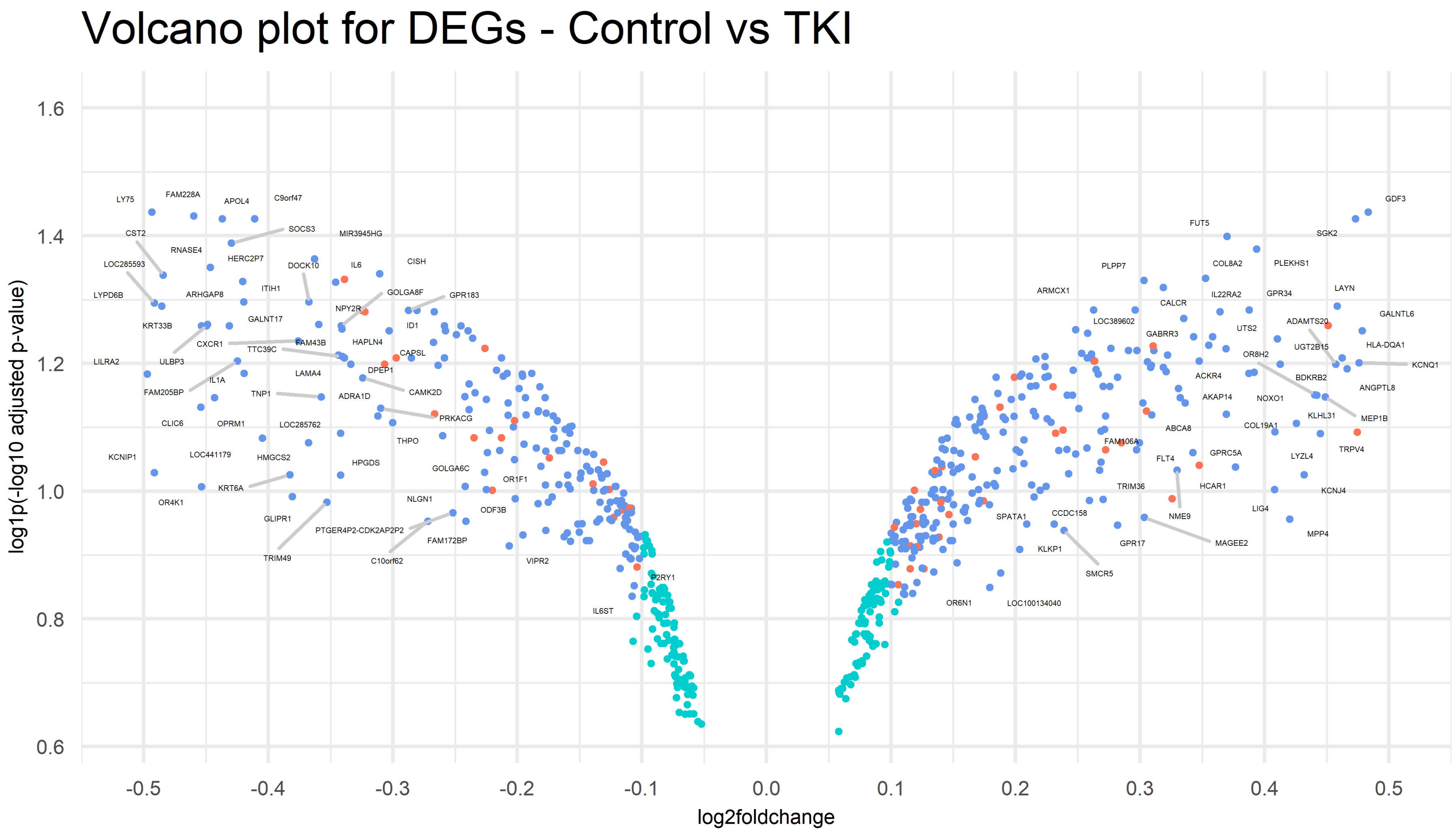


**Figure 21. Volcano plot for DEGs.** Genes are represented as scattered points: the x-axis is the log2FoldChange and the y-axis shows the log1p(-log10 adjusted p-value). Both red and blue dots represent genes that were identified as significantly differentially expressed (adjusted p-value $< 0.05$) with $|\log2FoldChange| > 0.1$. Specifically, red dots are referred to transcription factors.

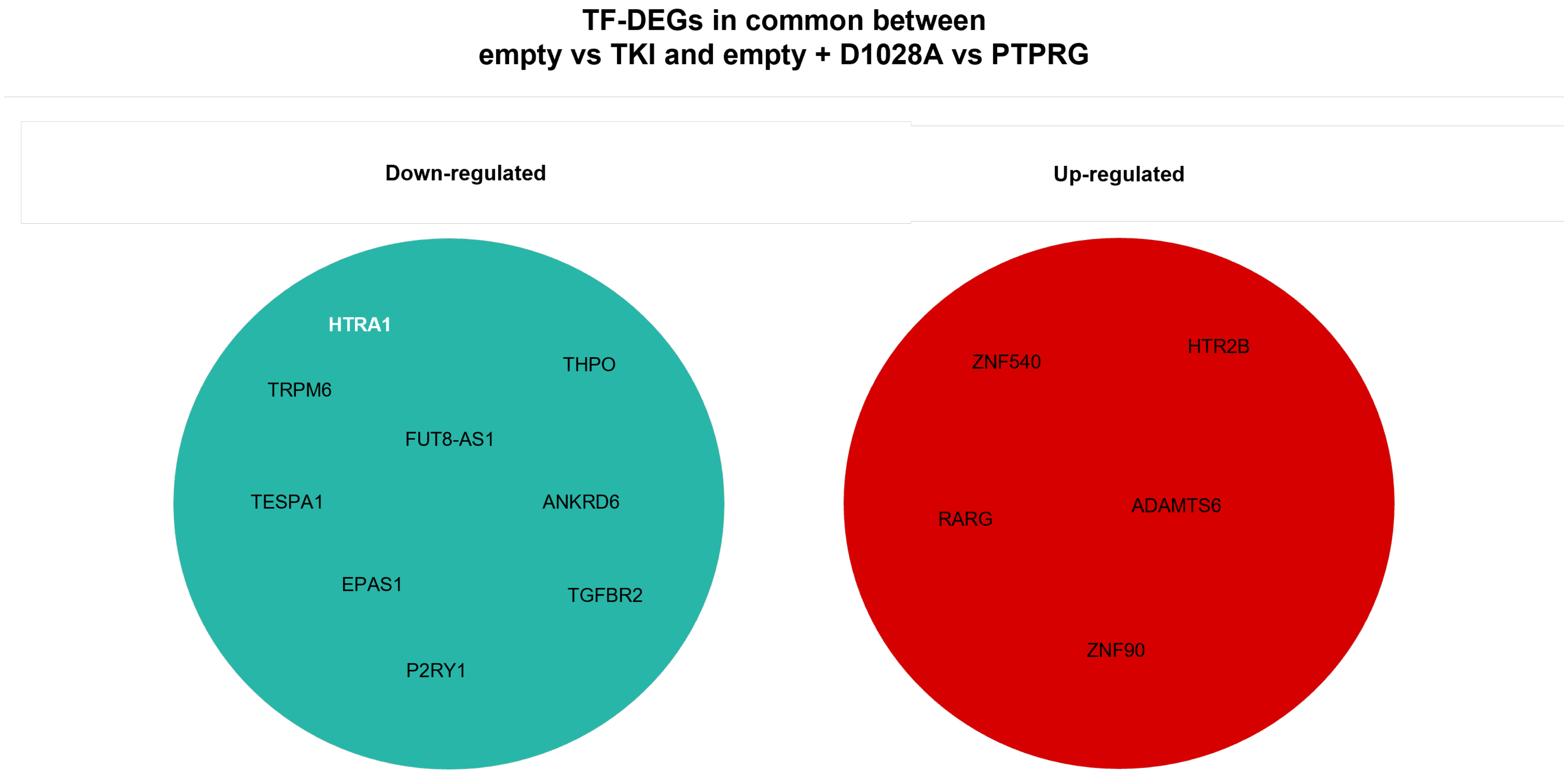


**Figure 22. TF-DEGs in common between empty vs TKI and empty + D1028A vs PTPRG.** Down-regulated genes are represented in the first set whilst up-regulated ones are shown in the second set. Moreover, TF-DEGs are written in bold white and DEGs are reported in black.

**Author Contributions:** The following statements should be used "Conceptualization, and methodology, PL, GL, and CS.; software, GL, PL.; validation, CB, RVL, , LT, formal analysis, GL, PL, MN, MF, CS; investigation, RVL, NAD, MAI.; resources, CS, MY.; data curation, PL, GL, CS, BC.; writing—original draft preparation, GL, PL, CS.; design of future directions PL, DC, AM; writing— review and editing, all authors; data visualization, GL; supervision, PL, CS; project administration, PL, CS, AM, DC.; funding acquisition, PL, CS, MY. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Interdisciplinary Grant 2020 issued by the Free University of Bozen-Bolzano, Italy. NPRP grant 4-157-3-052 from the Qatar National Research Fund (a member of the Qatar Foundation)

**Informed Consent Statement:** "Not applicable".

**Data Availability Statement:** The gene dataset and the software scripts used for its analysis of data can be found at the following GitLab link: https://gitlab.inf.unibz.it/Paola.Lecca/chronic-myeloid-leukemia-genes

**Acknowledgments:** We acknowledge the contribution of dr Elisabetta Moratti that prepared the RNA utilized for the analysis.

**Conflicts of Interest:** The authors declare no conflict of interest.'

## References

1. Cortes, J.; Pavlovsky, C.; Saußele, S. Chronic myeloid leukaemia. *The Lancet* **2021**, *398*, 1914–1926. doi:10.1016/s0140-6736(21)01204-6.
2. Hanfstein, B.; Müller, M.C.; Hochhaus, A. Response-related predictors of survival in CML. *Annals of Hematology* **2015**, *94*, 227–239. doi:10.1007/s00277-015-2327-x.
3. Julien, S.G.; Dubé, N.; Hardy, S.; Tremblay, M.L. Inside the human cancer tyrosine phosphatome. *Nature Reviews Cancer* **2010**, *11*, 35–49. doi:10.1038/nrc2980.
4. Chereda, B.; Melo, J.V. Natural course and biology of CML. *Annals of Hematology* **2015**, *94*, 107–121. doi:10.1007/s00277-015-2325-z.
5. Lecca, P.; Sorio, C. Accurate prediction of the age incidence of chronic myeloid leukemia with an improved two-mutation mathematical model. *Integrative Biology* **2016**, *8*, 1261–1275. doi:10.1039/c6ib00127k.
6. Boni, C.; Sorio, C. Current Views on the Interplay between Tyrosine Kinases and Phosphatases in Chronic Myeloid Leukemia. *Cancers* **2021**, *13*, 2311. doi:10.3390/cancers13102311.
7. Barnea, G.; Silvennoinen, O.; Shaanan, B.; Honegger, A.M.; Canoll, P.D.; D'Eustachio, P.; Morse, B.; Levy, J.B.; Laforgia, S.; Huebner, K. Identification of a carbonic anhydrase-like domain in the extracellular region of RPTP gamma defines a new subfamily of receptor tyrosine phosphatases. *Molecular and Cellular Biology* **1993**, *13*, 1497–1506. doi:10.1128/mcb.13.3.1497.
8. Vezzalini, M.; Mombello, A.; Menestrina, F.; Mafficini, A.; Peruta, M.D.; van Niekerk, C.; Barbareschi, M.; Scarpa, A.; Sorio, C. Expression of transmembrane protein tyrosine phosphatase gamma (PTP?) in normal and neoplastic human tissues. *Histopathology* **2007**, *50*, 615–628. doi:10.1111/j.1365-2559.2007.02661.x.
9. Wang, Z. Mutational Analysis of the Tyrosine Phosphatome in Colorectal Cancers. *Science* **2004**, *304*, 1164–1166. doi:10.1126/science.1096096.
10. Sorio, C.; Melotti, P.; D'Arcangelo, D.; Mendrola, J.; Calabretta, B.; Croce, C.M.; Huebner, K. Receptor protein tyrosine phosphatase gamma, Ptp gamma, regulates hematopoietic differentiation. *Blood* **1997**, *90*, 49–57.
11. Boni, C.; Sorio, C. The Role of the Tumor Suppressor Gene Protein Tyrosine Phosphatase Gamma in Cancer. *Frontiers in Cell and Developmental Biology* **2022**, *9*. doi:10.3389/fcell.2021.768969.
12. Kastury, K.; Ohta, M.; Lasota, J.; Moir, D.; Dorman, T.; LaForgia, S.; Druck, T.; Huebner, K. Structure of the Human Receptor Tyrosine Phosphatase Gamma Gene (PTPRG) and Relation to the Familial RCC t(3-8) Chromosome Translocation. *Genomics* **1996**, *32*, 225–235. doi:10.1006/geno.1996.0109.
13. van Niekerk, C.C.; Poels, L.G. Reduced expression of protein tyrosine phosphatase gamma in lung and ovarian tumors. *Cancer Letters* **1999**, *137*, 61–73. doi:10.1016/s0304-3835(98)00344-9.
14. Galvan, A.; Colombo, F.; Frullanti, E.; Dassano, A.; Noci, S.; Wang, Y.; Eisen, T.; Matakidou, A.; Tomasello, L.; Vezzalini, M.; et al. Germline polymorphisms and survival of lung adenocarcinoma patients: A genome-wide study in two European patient series. *International Journal of Cancer* **2014**, *136*, E262–E271. doi:10.1002/ijc.29195.
15. Drube, J.; Ernst, T.; Pfirrmann, M.; Albert, B.V.; Drube, S.; Reich, D.; Kresinsky, A.; Halfter, K.; Sorio, C.; Fabisch, C.; et al. PTPRG and PTPRC modulate nilotinib response in chronic myeloid leukemia cells. *Oncotarget* **2018**, *9*, 9442–9455. doi:10.18632/oncotarget.24253.
16. Ismail, M.A.; Vezzalini, M.; Morsi, H.; Abujaber, A.; Sayab, A.A.; Siveen, K.; Yassin, M.A.; Monne, M.; Samara, M.; Cook, R.; et al. Predictive value of tyrosine phosphatase receptor gamma for the response to treatment tyrosine kinase inhibitors in chronic myeloid leukemia patients. *Scientific Reports* **2021**, *11*. doi:10.1038/s41598-021-86875-y.

17. Ismail, M.A.; Nasrallah, G.K.; Monne, M.; AlSayab, A.; Yassin, M.A.; Varadharaj, G.; Younes, S.; Sorio, C.; Cook, R.; Modjtahedi, H.; et al. Description of PTPRG genetic variants identified in a cohort of Chronic Myeloid Leukemia patients and their ability to influence response to Tyrosine kinase Inhibitors. *Gene* **2022**, *813*, 146101. doi:10.1016/j.gene.2021.146101.
18. Peruta, M.D.; Martinelli, G.; Moratti, E.; Pintani, D.; Vezzalini, M.; Mafficini, A.; Grafone, T.; Iacobucci, I.; Soverini, S.; Murineddu, M.; et al. Protein Tyrosine Phosphatase Receptor Type $\gamma$ Is a Functional Tumor Suppressor Gene Specifically Downregulated in Chronic Myeloid Leukemia. *Cancer Research* **2010**, *70*, 8896–8906. doi:10.1158/0008-5472.can-10-0258.
19. Stevenson, W.S.; Best, O.G.; Przybylla, A.; Chen, Q.; Singh, N.; Koleth, M.; Pierce, S.; Kennedy, T.; Tong, W.; Kuang, S.Q.; et al. DNA methylation of membrane-bound tyrosine phosphatase genes in acute lymphoblastic leukaemia. *Leukemia* **2013**, *28*, 787–793. doi:10.1038/leu.2013.270.
20. Ismail, M.A.; Samara, M.; Sayab, A.A.; Alsharshani, M.; Yassin, M.A.; Varadharaj, G.; Vezzalini, M.; Tomasello, L.; Monne, M.; Morsi, H.; et al. Aberrant DNA methylation of PTPRG as one possible mechanism of its under-expression in CML patients in the State of Qatar. *Molecular Genetics & Genomic Medicine* **2020**, *8*. doi:10.1002/mgg3.1319.
21. Tomasello, L.; Vezzalini, M.; Boni, C.; Bonifacio, M.; Scaffidi, L.; Yassin, M.; Al-Dewik, N.; Kamga, P.T.; Krampera, M.; Sorio, C. Regulative Loop between $\beta$-catenin and Protein Tyrosine Receptor Type $\gamma$ in Chronic Myeloid Leukemia. *International Journal of Molecular Sciences* **2020**, *21*, 2298. doi:10.3390/ijms21072298.
22. Prange, K.H.; Singh, A.A.; Martens, J.H. The genome-wide molecular signature of transcription factors in leukemia. *Experimental hematology* **2014**, *42*, 637–650. doi:10.1016/j.exphem.2014.04.012.
23. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **2005**, *102*, 15545–15550. doi:10.1073/pnas.0506580102.
24. Draghici, S. *Statistics and Data Analysis for Microarrays Using R and Bioconductor*; Chapman and Hall/CRC, 2016. doi:10.1201/b11566.
25. Gene Ontology Homepage. http://geneontology.org/. Accessed: 2021-05-10.
26. Adrian Alexa, J.R. topGO, 2017. doi:10.18129/B9.BIOC.TOPGO.
27. Alexa, A.; Rahnenfuhrer, J.; Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **2006**, *22*, 1600–1607. doi:10.1093/bioinformatics/btl140.
28. R Wordcloud package. https://CRAN.R-project.org/package=wordcloud. Accessed: 2021-05-10.
29. R Text Mining package. https://cran.r-project.org/web/packages/tm. Accessed: 2021-05-10.
30. Walkley, C.R.; Olsen, G.H.; Dworkin, S.; Fabb, S.A.; Swann, J.; McArthur, G.A.; Westmoreland, S.V.; Chambon, P.; Scadden, D.T.; Purton, L.E. A Microenvironment-Induced Myeloproliferative Syndrome Caused by Retinoic Acid Receptor $\gamma$ Deficiency. *Cell* **2007**, *129*, 1097–1110. doi:10.1016/j.cell.2007.05.014.
31. Dewamitta, S.R.; Joseph, C.; Purton, L.E.; Walkley, C.R. Erythroid-extrinsic regulation of normal erythropoiesis by retinoic acid receptors. *British Journal of Haematology* **2013**, *164*, 280–285. doi:10.1111/bjh.12578.
32. Mao, B.; Huang, S.; Lu, X.; Sun, W.; Zhou, Y.; Pan, X.; Yu, J.; Lai, M.; Chen, B.; Zhou, Q.; et al. Early Development of Definitive Erythroblasts from Human Pluripotent Stem Cells Defined by Expression of Glycophorin A/CD235a, CD34, and CD36. *Stem Cell Reports* **2016**, *7*, 869–883. doi:10.1016/j.stemcr.2016.09.002.
33. Vezzalini, M.; Mafficini, A.; Tomasello, L.; Lorenzetto, E.; Moratti, E.; Fiorini, Z.; Holyoake, T.L.; Pellicano, F.; Krampera, M.; Tecchio, C.; et al. A new monoclonal antibody detects downregulation of protein tyrosine phosphatase receptor type $\gamma$ in chronic myeloid leukemia patients. *Journal of Hematology & Oncology* **2017**, *10*. doi:10.1186/s13045-017-0494-z.
34. Neff, T.; Armstrong, S.A. Chromatin maps, histone modifications and leukemia. *Leukemia* **2009**, *23*, 1243–1251. doi:10.1038/leu.2009.40.
35. Lodewick, J.; Lamsoul, I.; Polania, A.; Lebrun, S.; Burny, A.; Ratner, L.; Bex, F. Acetylation of the human T-cell leukemia virus type 1 Tax oncoprotein by p300 promotes activation of the NF-$\kappa$B pathway. *Virology* **2009**, *386*, 68–78. doi:10.1016/j.virol.2008.12.043.
36. Shi, Y.; Tomic, J.; Wen, F.; Shaha, S.; Bahlo, A.; Harrison, R.; Dennis, J.W.; Williams, R.; Gross, B.J.; Walker, S.; et al. Aberrant O-GlcNAcylation characterizes chronic lymphocytic leukemia. *Leukemia* **2010**, *24*, 1588–1598. doi:10.1038/leu.2010.152.
37. Ni, F.; Yu, W.M.; Li, Z.; Graham, D.K.; Jin, L.; Kang, S.; Rossi, M.R.; Li, S.; Broxmeyer, H.E.; Qu, C.K. Critical role of ASCT2-mediated amino acid metabolism in promoting leukaemia development and progression. *Nature Metabolism* **2019**, *1*, 390–403. doi:10.1038/s42255-019-0039-6.
38. Sell, S. Leukemia: Stem Cells, Maturation Arrest, and Differentiation Therapy. *Stem Cell Reviews* **2005**, *1*, 197–206. doi:10.1385/scr:1:3:197.
39. Mineo, M.; Garfield, S.H.; Taverna, S.; Flugy, A.; Leo, G.D.; Alessandro, R.; Kohn, E.C. Exosomes released by K562 chronic myeloid leukemia cells promote angiogenesis in a src-dependent fashion. *Angiogenesis* **2011**, *15*, 33–45. doi:10.1007/s10456-011-9241-1.
40. Lamalice, L.; Boeuf, F.L.; Huot, J. Endothelial Cell Migration During Angiogenesis. *Circulation Research* **2007**, *100*, 782–794. doi:10.1161/01.res.0000259593.07661.1e.
41. Gowda, C.; Song, C.; Ding, Y.; Iyer, S.; Dhanyamraju, P.K.; McGrath, M.; Bamme, Y.; Soliman, M.; Kane, S.; Payne, J.L.; et al. Cellular signaling and epigenetic regulation of gene expression in leukemia. *Advances in Biological Regulation* **2020**, *75*, 100665. doi:10.1016/j.jbior.2019.100665.
42. Kato, R.; Kiryu-Seo, S.; Kiyama, H. Damage-Induced Neuronal Endopeptidase (DINE/ECEL) Expression Is Regulated by Leukemia Inhibitory Factor and Deprivation of Nerve Growth Factor in Rat Sensory Ganglia after Nerve Injury. *Journal of*

*Neuroscience* **2002**, *22*, 9410–9418, [https://www.jneurosci.org/content/22/21/9410.full.pdf]. doi:10.1523/JNEUROSCI.22-21-09410.2002.

43. Gaspar, B.L.; Sharma, P.; Varma, N.; Sukhachev, D.; Bihana, I.; Naseem, S.; Malhotra, P.; Varma, S. Unique characteristics of leukocyte volume, conductivity and scatter in chronic myeloid leukemia. *Biomedical Journal* **2019**, *42*, 93–98. doi:10.1016/j.bj.2018.12.004.
44. Li, J.; Dong, S. The Signaling Pathways Involved in Chondrocyte Differentiation and Hypertrophic Differentiation. *Stem Cells International* **2016**, *2016*, 1–12. doi:10.1155/2016/2470351.
45. Koníková, E.; Kusenda, J. P53 protein expression in human leukemia and lymphoma cells. *Neoplasma* **2001**, *48*, 290–298.
46. Petzer, A.L.; Gunsilius, E. Hematopoietic stem cells in chronic myeloid leukemia. *Archives of Medical Research* **2003**, *34*, 496–506. doi:10.1016/j.arcmed.2003.09.005.
47. McCubrey, J.; May, W.S.; Duronio, V.; Mufson, A. Serine/threonine phosphorylation in cytokine signal transduction. *Leukemia* **2000**, *14*, 9–21. doi:10.1038/sj.leu.2401657.
48. Powell, A.E.; Shung, C.Y.; Saylor, K.W.; Müllendorf, K.A.; Weiss, J.B.; Wong, M.H. Lessons from development: A role for asymmetric stem cell division in cancer. *Stem Cell Research* **2010**, *4*, 3–9. doi:10.1016/j.scr.2009.09.005.
49. Bard, J.B.L.; Rhee, S.Y. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics* **2004**, *5*, 213–222. doi:10.1038/nrg1295.
50. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: tool for the unification of biology. *Nature Genetics* **2000**, *25*, 25–29. doi:10.1038/75556.
51. Robinson, P. *Introduction to bio-ontologies*; CRC Press: S.l, 2020.
52. Schuurman, N.; Leszczynski, A. Ontologies for bioinformatics. *Bioinform Biol Insights* **2008**, *2*, 187–200.
53. Bodenreider, O.; Stevens, R. Bio-ontologies: current trends and future directions. *Brief Bioinform* **2006**, *7*, 256–274.
54. Cook, D.L.; Mejino, J.L.; Neal, M.L.; Gennari, J.H. Bridging biological ontologies and biosimulation: the ontology of physics for biology. *AMIA Annu Symp Proc* **2008**, pp. 136–140.
55. Ontology of Physics for Biology. https://bioportal.bioontology.org/ontologies/OPB/?p=summary. Accessed: 2021-03-20.
56. Cook, D.L.; Neal, M.L.; Bookstein, F.L.; Gennari, J.H. Ontology of physics for biology: representing physical dependencies as a basis for biological processes. *Journal of Biomedical Semantics* **2013**, *4*, 41. doi:10.1186/2041-1480-4-41.
57. BioPAX Homepage. http://www.biopax.org/. Accessed: 2021-06-14.
58. Pathway Commons Homepage. http://www.pathwaycommons.org/. Accessed: 2021-06-14.
59. Damiani, C.; Filisetti, A.; Graudenzi, A.; Lecca, P. Parameter sensitivity analysis of stochastic models: Application to catalytic reaction networks. *Computational Biology and Chemistry* **2013**, *42*, 5–17. doi:10.1016/j.compbiolchem.2012.10.007.
60. Lecca, P.; Palmisano, A.; Priami, C.; Sanguinetti, G. A new probabilistic generative model of parameter inference in biochemical networks. In Proceedings of the Proceedings of the 2009 ACM symposium on Applied Computing - SAC 09. ACM Press, 2009. doi:10.1145/1529282.1529442.
61. Lecca, P. A time-dependent extension of gillespie algorithm for biochemical stochastic $\pi$-calculus. In Proceedings of the Proceedings of the 2006 ACM symposium on Applied computing - SAC '06. ACM Press, 2006. doi:10.1145/1141277.1141310.