Nucleic Acids Research, 2025, 1–7
https://doi.org/10.1093/nar/gkaf420
Web Server issue

MAGNETIC: a web server to fetch gene network based on motif distribution in promoters

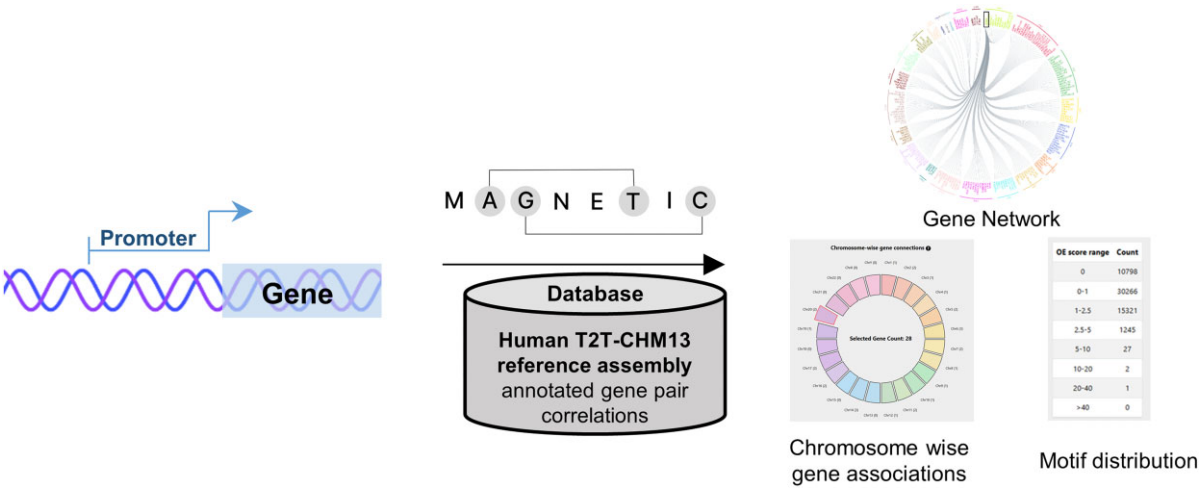
Atreyi Chakraborty[1], Saksham Srimali[1], Rohan Topno[1], Ajinkya Ranshur[1], Mallur Srivatsan Madhusudhan [1,2,*]

[1]Department of Biology, Indian Institute of Science Education and Research, Pune, Maharashtra 411008, India
[2]Department of Data Science, Indian Institute of Science Education and Research, Pune, Maharashtra 411008, India
[*]To whom correspondence should be addressed. Email: madhusudhan@iiserpune.ac.in

Abstract

Our web server MAGNETIC (Motif Associated Gene NETworks in Chromosomes) allows users to search the human genome for correlations between promoter regions of genes. The correlations take into consideration the similarity of the abundance of 5/6-mer motifs in gene promoters. The promoters could be 1, 2, or 6 kb upstream of the gene start site. Genes with similar motif abundances are linked to form a gene regulatory network. These networks could help determine or even discover gene networks of input genes. Our database could also be searched by motifs (including degenerate positions). The results identify the abundance of the motif in the promoters of all genes. The promoters where the motif is found in high abundance are likely to be target binding sites of a protein that recognizes the input sequence. MAGNETIC can also determine the similarity between gene promoters at the scale of whole chromosomes. We have showcased our server using a few examples that involve the consensus recognition motif of the transcription factor, Myc, and the genes it regulates. We have identified putative new targets of Myc and also found instances where Myc could only indirectly regulate genes. This server could help make important connections between genes and give insights into gene regulation and function.

Graphical abstract



Introduction

Genomes are read and recognized by cells using proteins such as transcription factors, repressors, etc. We had previously established that when specifically recognizing DNA, these proteins make contact with motifs of 5–6-nt base pairs [1]. We are particularly interested in the regions of the genome immediately upstream of genes [2, 3]. These are the regions that we designated as promoters, where gene regulators would bind. We had also previously established that correlating motif vectors gave us a plausible gene regulatory network [4]. Motif vectors are the distribution of all the possible motifs of sizes 5 and 6 in promoter regions. We conjecture that similar motifs in different promoters would elicit binding by the same proteins/factors that specifically recognize the motifs, leading to possible co-regulation. Similarities between motif vectors are computed by Pearson correlation.

Received: February 16, 2025. **Revised:** April 10, 2025. **Editorial Decision:** April 28, 2025. **Accepted:** May 3, 2025

In our web server, Motif Associated Gene NETworks in Chromosomes (MAGNETIC), we investigate similarities between gene promoter regions. The currently updated T2T-CHM13 reference [5] of the human genome has annotated 63 494 genes [6]. These genes include those for protein coding, rRNA, tRNA, etc. We have correlated the motif vectors of promoter regions of all genes against all other genes. MAGNETIC allows users to query a database of over 2 billion possible connections in different ways. Inputting a 5–6 letter motif identifies the genes in which these motifs are present, arranged in descending order of the number of instances of the motif. One could query the database with a gene or genes of interest to identify other genes with similar motif vectors. The implication here is that high vector correlations is plausible co-regulation. Users could also compare whole chromosomes with one another and identify those whose gene promoters are most similar. MAGNETIC offers many avenues for experimentally validating gene interactions/associations, networks, and discovering proteins that are likely to target specific DNA motifs and make evolutionary connections.

Method

1. **Network generation and correlation calculations:** The MAGNETIC server generates gene networks based on motif abundance in promoter regions. The abundance of a motif is determined by an Observed to Expected (OE) ratio. The OE ratio for all motifs of double-stranded DNA is defined as a motif vector [4]. To facilitate this, we have pre-computed a database containing annotated gene correlations derived from motif occurrences. The correlation calculations are performed using two key parameters:

a. Promoter size: The server reads the promoter lengths of 1, 2, and 6 kb upstream of the query gene transcription start sites.
b. Motif size: Motif OE ratios and correlations are computed using motif vector resolutions of 5 and 6 bp. For motifs of size 5 and 6, there are 512 and 2080 unique motifs of double-stranded DNA, respectively. A motif and its reverse complement from the other DNA are not double counted and are accounted as one motif. Correlations between genes with promoters whose motif vectors are correlated above a (user specified) threshold are considered to be networked.

2. **Database pre-computation:** For the human T2T reference assembly, motif vectors and correlations between all annotated gene pairs are pre-computed and stored in the PostgreSQL 12.22 [7] database, allowing for rapid retrieval and network construction.
3. **Visualization and web framework:** Heatmaps, graphs, and networks are generated using the D3.js module (https://d3js.org/what-is-d3) of the JavaScript library. The web framework is developed using Django 5 (https://docs.djangoproject.com/en/5.2/), ensuring efficient data handling and user interface.
4. **Implementation details:** The correlation calculations and database processing are optimized for scalability. The visualization tools provide users with dynamic and customizable gene network representations. The server interface is designed to be intuitive and accessible to researchers analysing gene regulatory networks.

Server description:

Users could query the possible gene associations in three different ways (delineated by tabs on the landing page).

(1) **Input a motif of choice:**

Users can input either a 5- or a 6-mer motif of interest and also specify the size of the promoter region, i.e. the region upstream of the gene they want to search the motif in. The user can also search for motifs with variable/degenerate characters, such as GNTRYC, where N could be any nucleotide, R could be any of the purines, and Y could be any of the pyrimidines. This generalization could help in clustering motifs. The output is presented in two steps. First, a Table of the number of genes with abundances (in terms of OE values) of the input motif in different ranges is displayed. On choosing upper and lower limits of the OE values, the user is presented with up to 20 genes whose promoters contain the input motif with the highest values, in decreasing order (Fig. 1). The users can download the whole list for a comprehensive list of all genes and the OE values of the input motifs in their promoters.

(2) **Input a gene or genes of interest**

If the users are interested in predicting what other genes could be co-regulated along with particular genes of interest, the second tab allows for such searches. The input could be a gene or multiple gene names separated by commas. All gene names have to adhere to the HUGO nomenclature [8]. In the input, users must also specify what the cut-off value of promoter vector motif correlations should be. Only results of gene pairs having correlations above this value are displayed. In addition to these inputs, the gene promoter size and motif size (5- or 6-mer) must also be specified. The output displays the total number of genes correlated to the query genes above the given correlation coefficient threshold, with chromosome-wise break up. A more detailed network of correlations between gene promoters is shown below the chromosome-wise diagram (Fig. 2). Mousing over the connection displays the pairs that are connected along with the value of their motif vector correlation and the top five motifs (in terms of OE ratios) that have contributed to the correlation. The motifs shown in red and are common to the two gene promoters. Single click on the gene will direct the user to UCSC genome browser. A double click will direct the user to the PubMed Central (https://pmc.ncbi.nlm.nih.gov/) site displaying all the hits with the keyword of selected gene 'AND' query gene. The network is only displayed for a maximum of 300 genes. Should the number exceed this number for a particular query, the only option is to download the results. In general, in addition to the graphic representation, all results could be downloaded.

An alternative search strategy using gene names would be from the output of the motif search. The top 20 genes obtained from a motif search have links that will direct the user to the gene network tab.

(3) **Investigate the connections between whole chromosomes**

The similarity of chromosomes based on the correlations between their constituent genes could be seen for different cut-off values of correlation coefficient, gene promoter size, and

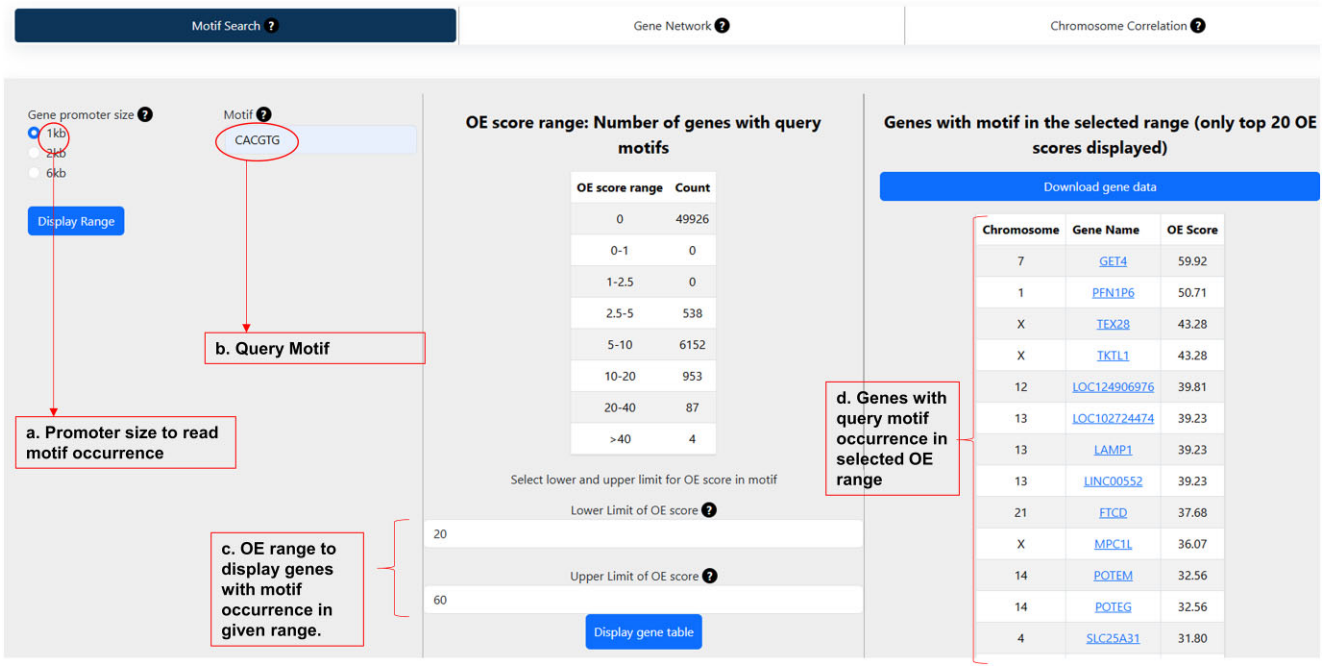


Figure 1. The description of 'Motif search' option on MAGNETIC. The text boxes labeled a–d explain the different parameters of the query search and the output results.

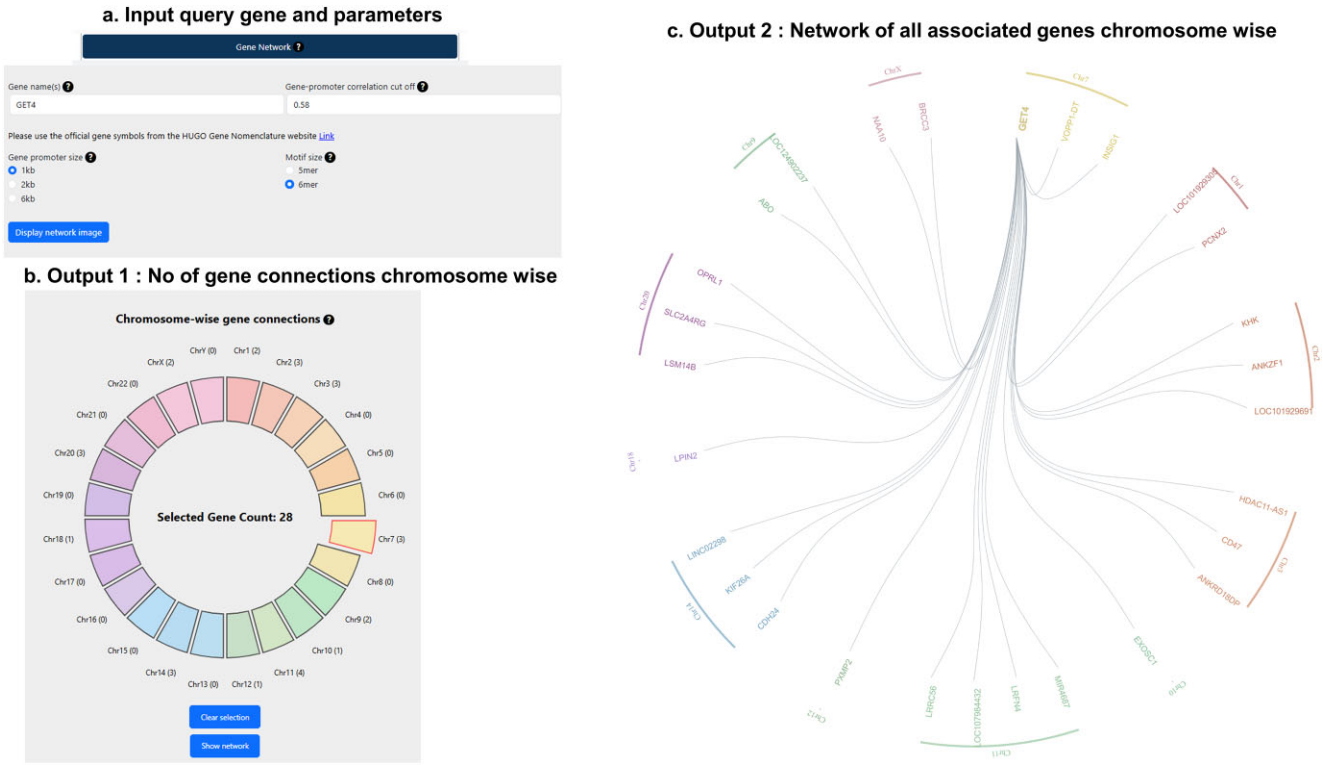


Figure 2. Using the 'Gene network' option of MAGNETIC for a single query gene. (**A**) The input query gene, '**GET4**', and parameters of motif size, promoter size, and correlation cut-off. (**B**) The first output displays the gene associations per chromosome above the input correlation cut-off threshold. The chromosome housing the query gene is highlighted with a red border. (**C**) The second part of the output shows the gene regulatory network of associations with the query gene. The different gene colours represent the different chromosomal locations of the associated genes.

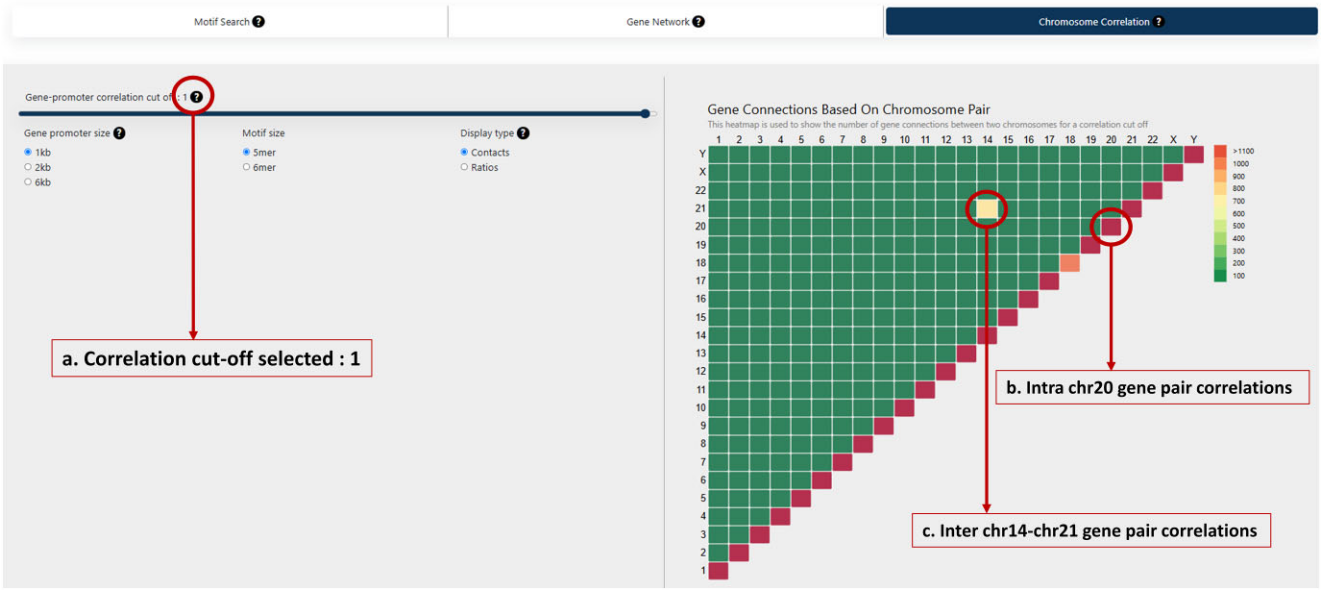


Figure 3. The 'Chromosome Correlation' option of MAGNETIC. A heat map of chromosome pair wise gene correlations. Highlighted text boxes a–c explain the different parameters of the query search and the description of the results.

motif size. Users could view heat maps of such chromosome similarities depending on whether they choose to see the correlations in terms of the total number of contacts (number of gene pairs with correlation coefficients above the threshold) or the ratios (the number of gene pairs normalized in the range 0–1) (Fig. 3).

Our server cross-references PubMed Central (https://pmc.ncbi.nlm.nih.gov/) and the UCSC genome browser [9, 10].

Results

1. **Finding motif abundances across genes:** We queried MAGNETIC for the abundance of the motif **CACGTG** (E box), a reported consensus motif recognized by the transcription factor Myc [11, 12]. The search was done 1 kb upstream of gene start site. In the Table of OE value ranges and genes, 4 genes had OE values in the range of 40–60 and 87 genes in the range of 20–40 (Fig. 1). The output displayed the top 20 genes with the highest OE scores. Of these 20 genes, we further considered only the 12 protein coding genes and excluded the non-characterized genes or the long non-coding RNA genes. Of these 12 genes, GOLIM4 had the least OE score of 29.00. The ChEA Transcription Factor Targets database [13, 14] lists 7 of the 12 protein coding genes as known targets of Myc transcription activation (Table 1a). The other five genes whose OE values for Myc consensus sites are in the range of ∼32–51 are not recognized by ChEA as targets of Myc (Table 1b). This could be because they are false positives of our method or they could indeed be activated by Myc in a cell/tissue type for which the ChIPSeq data is not curated in ChEA.
2. **Genes associated with GET4 also share a common transcription factor:** As an example of using MAGNETIC to find associations of a query gene of interest, we input the gene GET4. This gene was identified in the previous example as the one with the highest abundance of Myc Consensus Motif (Result section 1). At correlation coefficient cut-off values of 0.50, 0.55, and 0.58, we get 1626, 247, and 27 gene associations, respectively, with a promoter size of 1 kb and motif size of 6-mer (Fig. 2a).

Table 1a. List of protein coding genes with high OE score for Myc consensus motif and are Myc targets

Chromosome	Gene name	OE score
7	GET4	59.92
X	TKTL1	43.28
13	LAMP1	39.23
21	FTCD	37.68
4	SLC25A31	31.80
Y	GTPBP6	30.03
3	GOLIM4	29.00

Table 1b. List of protein coding genes with Myc consensus motif OE above 30 and not a Myc target

Chromosome	Gene name	OE score
1	PFN1P6	50.71
X	TEX28	43.28
X	MPC1L	36.07
14	POTEM	32.56
14	POTEG	32.56

Of the 27 genes associated with GET4, 9 are uncharacterized genes or non-coding RNA (Supplementary Table S1a). Of the 18 protein coding genes, 11 were found in citations when queried along with GET4 on PubMed Central (Supplementary Table S1b). Seven of the 18 protein coding genes were identified as Myc targets in the 2010 version of ChEA [13]. Subsequently, five others were also identified as Myc targets in the 2022 version, taking the number of Myc targets to 12 (Table 2a and Supplementary Table S2). Of these 12 known targets of Myc, only 8 have non-zero OE values for the Myc consensus site (ranging from OE val-

Table 2a. List of protein coding genes associated with GET4 that are identified as Myc targets and the OE score of the Myc consensus motif in their promoters

S. no.	Chromosome	Gene	OE score
1.	2	ANKZF1	0
2.	10	EXOSC1	6.32
3.	11	LRFN4	12.63
4.	12	PXMP2 (PMP22)	0
5.	20	LSM14B	0
6.	7	INSIG1	6.66
7.	X	NAA10	7.21
8.	20	SLC2A4RG	5.56
9.	11	LRRC56	25.26
10.	2	KHK	6.87
11.	3	CD47	0
12.	20	OPRL1	5.56

Table 2b. List of protein coding genes associated with GET4 which are not identified as Myc targets and the OE score of the Myc consensus motif

S. no.	Chromosome	Gene	OE score
1.	1	PCNX2	0
2.	14	CDH24	0
3.	18	LPIN2	14.15
4.	9	ABO	0
5.	X	BRCC3	0
6.	14	KIF26A	6.51

Table 3. Chromosomes wise list of genes predicted to be co-regulated along with GET4 and LAMP1 genes

S. no.	Chromosome	Common interactors
1.	1	PHC2AS1, LOC101929536
2.	2	ANKZEF1
3.	4	CTBP1-DT
4.	5	MAPK9
5.	7	INSIG1, VOPP1-DT
6.	8	MIR4470
7.	9	ABO, LOC124902237
8.	10	SEC61A2
9.	11	LRFN4, LRRC56
10.	12	RASSF3-DT
11.	14	KIF264, ZFYVE21
12.	16	CHTF18, CASKIN1, LOC100134368
13.	18	LPIN2, PPP4R1-AS1, KDSR, LOC124904317
14.	21	SUMO3
15.	X	TMEM187, HCFC1, BRCC3

ues of ∼5–25). Four genes have no occurrences of the Myc consensus site in their 1 kb promoters. Three of the four ANKZF1, PXMP2, and LSM14B have OE (0, 1.14), (3.31, 1.10), and (2.77, 3.69) at 2 and 6 kb promoter size, respectively (Supplementary Table S3). It is possible that in these cases, Myc recognizes variants of the consensus or the Myc binding sites for these genes could lie beyond 1 kb or that Myc activates another transcription factor that in turn activates these genes. Among the 6 genes that are associated with GET4 but not identified as Myc targets are two genes, LPIN2 and KIF26A that have high OE values (14.15 and 6.51 respectively) for the Myc consensus site (Table 2b). It is likely that these genes could be activated by Myc.

We compared the GET4 network to other predicted networks: GeneMANIA [15], Network Analyst [16] (using signalling network database SIGNOR 2.0 [17]), and BioGrid [18]. GeneMANIA generates gene network based on different datasets of functional association data such as protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity [15]. Network Analyst generates gene networks based on transcription factor, miRNA or signalling associations (SIGNOR2.0) [16, 17]. BioGrid generates network based on a curated database of protein and genetic interactions sourced from literature [18]. GET4 had 247, 162, 20, and 0 associations obtained from MAGNETIC (at 0.5 correlation cut-off), BioGrid, GeneMANIA, and Network Analyst, respectively (Supplementary Fig. S1). Of all the MAGNETIC associations, two genes (UBL4A and FASN) were in common with those found by BioGrid. UBL4A was also found in common with GeneMANIA. As Network Analyst had no associations, there were no common genes with MAGNETIC. The overlap between predicted network connections by the different methods is poor (Supplementary Fig. S1).

There is a fundamental difference in the networks made by MAGNETIC and the other methods. From the correlation of the motif vectors of gene promoters, we are establishing the possibility of genes to be co-regulated, i.e. recognized by the same protein partner. The resulting gene products may not necessarily belong to the same pathway or even interact. In contrast, the different gene network predictors rely on pathways, physical interactions, co-expression, etc. Our method is agnostic to these considerations and makes prediction based on plausible co-regulation. For users to verify or validate some of the predictions made, MAGNETIC also retrieves PubMed Central searches based on keyword-based literature search of genes (Supplementary Table S1b).

1. **Multiple gene network to find common associations:** The genes GET4 and LAMP1 were the multiple genes input to MAGNETIC which identified 116 genes that are likely to be co-regulated along with the input genes. Of this list, 27 genes are correlated to both input genes with correlation coefficients greater than the selected threshold of 0.57 (Table 3). These 27 genes identified across 15 different chromosomes could then be further investigated to determine a common pathway.
2. **Genes with shared promoters from inter- and intra-chromosome gene correlations:** Using the chromosome correlation tab on MAGNETIC with the correlation coefficient threshold set to 1.0 (identical motif distributions) for 1 kb promoters and 5-mer motifs, we obtained an inter- and intra-chromosome correlation heatmap. The highest inter-chromosome correlations were between 14 and 21 (Fig. 3). For this chromosome pair, 758 gene pairs involving 16 genes from chromosome 14 and 55 from chromosome 21 had a correlation coefficient of 1. On alignment, the promoters were, on average, 96.8% identical to one another. These two chromosomes are known to undergo Robertsonian translocations [19]. It is likely that the near-identical promoter regions of the genes in this part of these chromosomes are an important factor influencing the translocations.

On examining gene pairs within chromosome 20, we found nine gene pairs with a correlation coefficient of 1. Among

these, 9 are a pair CPNE1 and RBM12 [20] that are located on the same strands and have overlapping promoter sites. MAGNETIC can help examine such cases. For instance, MAGNETIC could also be used to analyse how multiple copies of a gene are regulated and what other genes are co-regulated along with each of the gene copies. Thus, in turn may help identify the genes that are expressed/regulated in particular cell/tissue types.

Conclusion

MAGNETIC attempts to link genes based on the similarity of the abundance of 5–6 letter motifs in their promoter regions. The conjecture here is that promoters with similar motif abundances would be acted upon by the same proteins and hence these genes could be part of a regulatory network. The output of MAGNETIC is not tissue type specific or affected by expression abundances, both of which would affect the results of the other methods. Our networks only make plausible connections based on an estimate of abundance of the likely protein binding target sequences. In an earlier study, we also showed that networked genes are also likely to be co-localized in the nucleus. So, in addition to mining our database for genes with similar promoters, one could also predict co-localized genes. Future additions to the web server would be along the following directions

(1) The inclusion of non-genic regions. In the human genome, the genic regions constitute less than 1/20th of the whole genome. There are likely to be correlations among regions of non-gene promoter regions that could be useful in determining regions that are co-localized. These co-localized regions could harbour distant promoter sites because of DNA packing and/or give insights into DNA packing in the nucleus. Correlated regions could also give us predictions of euchromatin and hetero chromatin (compartments A and B) stretches of the genome.
(2) We already have preliminary data on several other genomes other than the human genome. Soon, we will populate MAGNETIC with data from many model organisms such as *Drosophila melanogaster*, *Mus musculus*, *Caenorhabditis elegans*, etc. Searches similar to those described above would now be possible with these other genomes.
(3) MAGNETIC would hence serve as a platform to study conservation of motif abundances across genomes and give insights into the evolution of gene regulation.

Acknowledgements

We would like to acknowledge COSPI lab members for constructive criticism. We also acknowledge the support and computational resources provided by the PARAM Brahma facility under the National Supercomputing Mission, Government of India at the Indian Institute of Science Education and Research (IISER) Pune.

Author contributions: Atreyi Chakraborty (Conceptualization [equal], Data curation [equal], Project administration [equal], Writing—original draft [equal]), Saksham Srimali (Software [equal]), Rohan Topno (Software [equal]), Ajinkya Ranshur (Software [equal]), Mallur Srivatsan Madhusudhan (Conceptualization [equal], Funding acquisition [equal], Project administration [equal], Writing—review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Department of Biotechnology, Government of India grant to the Indian Institute of Science Education and Research Pune under BIC grant (BT/PR40262/BTIS/137/38/2022 and BT/PR40323/BTIS/137/78/2023). Funding to pay the Open Access publication charges for this article was provided by the Department of Biotechnology, Government of India.

Data availability

Users can access the MAGNETIC web server at https://cospi.iiserpune.ac.in/magnetic/ without any login or registration requirement. The site contains no cookie settings. Our contact information is provided on the site. All the data (~5 TB) that went into the server is available on request.

References

1. Nair S, Madhusudhan MS. JEDII: juxtaposition enabled DNA-binding interface identifier. bioRxiv, https://doi.org/10.1101/2022.05.19.492702, 20 May 2022, preprint: not peer reviewed.
2. Sanyal A, Lajoie BR, Jain G *et al.* The long-range interaction landscape of gene promoters. *Nature* 2012;**489**:109–13.
3. Hurst LD, Sachenkova O, Daub C *et al.* A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol* 2014;**15**:413. https://doi.org/10.1186/s13059-014-0413-3
4. Chakraborty A, Chopde S, Madhusudhan MS. Motif distribution in genomes gives insights into gene clustering and co-regulation. *Nucleic Acids Res* 2025;**53**:1178. https://doi.org/10.1093/nar/gkae1178
5. Aganezov S, Yan SM, Soto DC *et al.* A complete reference genome improves analysis of human genetic variation. *Science* 2022;**376**:54–69. https://doi.org/10.1126/science.abl3533
6. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53. https://doi.org/10.1126/science.abj6987
7. Bostock M, Ogievetsky V, Heer J. D3 data-driven documents. *IEEE Trans Visual Comput Graphics* 2011;**17**:2301–9. https://doi.org/10.1109/TVCG.2011.185
8. Seal RL, Braschi B, Gray K *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res* 2023;**51**:D1003–9. https://doi.org/10.1093/nar/gkac888
9. Navarro Gonzalez J, Zweig AS, Speir ML *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* 2021;**49**:D1046–57. https://doi.org/10.1093/nar/gkaa1070
10. Nassar LR, Barber GP, Benet-Pagès A *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* 2023;**51**:D1188–95. https://doi.org/10.1093/nar/gkac1072

11. Adhikary S, Eilers M. Transcriptional regulation and transformation by Myc proteins. *Nat Rev Mol Cell Biol* 2005;**6**:635–45. https://doi.org/10.1038/nrm1703
12. Chen H, Liu H, Qing G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct Target Ther* 2018;**3**:5.
13. Lachmann A, Xu H, Krishnan J *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 2010;**26**:2438–44. https://doi.org/10.1093/bioinformatics/btq466
14. Keenan AB, Torre D, Lachmann A *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res* 2019;**47**:W212–24. https://doi.org/10.1093/nar/gkz446
15. Mostafavi S, Ray D, Warde-Farley D *et al.* GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* 2008;**9**(**Suppl. 1**):S4. https://doi.org/10.1186/gb-2008-9-s1-s4
16. Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat Protoc* 2015;**10**:823–44. https://doi.org/10.1038/nprot.2015.052
17. Licata L, Lo Surdo P, Iannuccelli M *et al.* SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res* 2020;**48**:D504–10.
18. Oughtred R, Rust J, Chang C *et al.* The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 2021;**30**:187–200. https://doi.org/10.1002/pro.3978
19. Saeed S, Hassan J, Javed SM *et al.* A familial case of Robertsonian translocation 13;14: case report. *Cureus* 2022;**14**:e29430. https://doi.org/10.7759/cureus.29430
20. Yang W, Ng P, Zhao M *et al.* Promoter-sharing by different genes in human genome—CPNE1 and RBM12 gene pair as an example. *BMC Genomics* 2008;**9**:456. https://doi.org/10.1186/1471-2164-9-456

Received: February 16, 2025. **Revised:** April 10, 2025. **Editorial Decision:** April 28, 2025. **Accepted:** May 3, 2025