W322–W329 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 26 May 2022*
https://doi.org/10.1093/nar/gkac393

FABIAN-variant: predicting the effects of DNA variants on transcription factor binding

Robin Steinhaus [1,2,*], **Peter N. Robinson** [3,4] **and Dominik Seelow** [1,2]

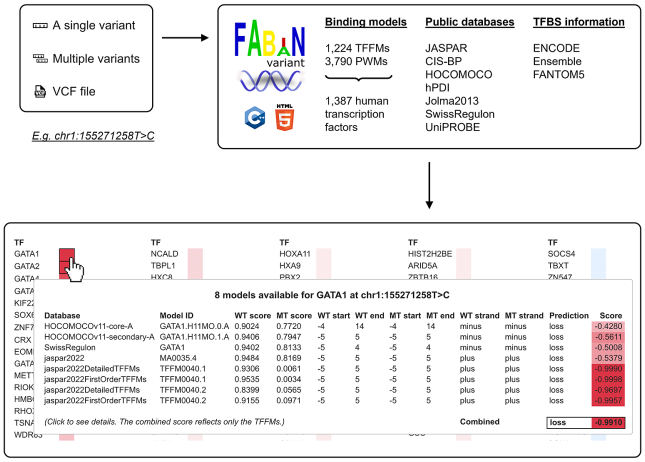
[1] Exploratory Diagnostic Sciences, Berlin Institute of Health, 10117 Berlin, Germany, [2] Institute of Medical Genetics and Human Genetics, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, 13353 Berlin, Germany, [3] The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030, USA and [4] Institute for Systems Genomics, University of Connecticut, Farmington, CT 06030, USA

Received March 15, 2022; Revised April 22, 2022; Editorial Decision May 02, 2022; Accepted May 06, 2022

ABSTRACT

While great advances in predicting the effects of coding variants have been made, the assessment of non-coding variants remains challenging. This is especially problematic for variants within promoter regions which can lead to over-expression of a gene or reduce or even abolish its expression. The binding of transcription factors to the DNA can be predicted using position weight matrices (PWMs). More recently, transcription factor flexible models (TFFMs) have been introduced and shown to be more accurate than PWMs. TFFMs are based on hidden Markov models and can account for complex positional dependencies. Our new web-based application FABIAN-variant uses 1224 TFFMs and 3790 PWMs to predict whether and to which degree DNA variants affect the binding of 1387 different human transcription factors. For each variant and transcription factor, the software combines the results of different models for a final prediction of the resulting binding-affinity change. The software is written in C++ for speed but variants can be entered through a web interface. Alternatively, a VCF file can be uploaded to assess variants identified by high-throughput sequencing. The search can be restricted to variants in the vicinity of candidate genes. FABIAN-variant is available freely at https://www.genecascade.org/fabian/.

GRAPHICAL ABSTRACT



8 models available for GATA1 at chr1:155271258T>C

Database	Model ID	WT score	MT score	WT start	WT end	MT start	MT end	WT strand	MT strand	Prediction	Score
HOCOMOCOv11-core-A	GATA1.H11MO.0.A	0.9024	0.7720	-4	14	-4	14	minus	minus	loss	-0.4280
HOCOMOCOv11-secondary-A	GATA1.H11MO.1.A	0.9406	0.7947	-5	5	-5	5	minus	minus	loss	-0.5611
SwissRegulon	GATA1	0.9402	0.8133	-5	4	-5	4	minus	minus	loss	-0.5008
jaspar2022	MA0035.4	0.9484	0.8169	-5	5	-5	5	plus	plus	loss	-0.5379
jaspar2022DetailedTFFMs	TFFM0040.1	0.9306	0.0061	-5	5	-5	5	plus	plus	loss	-0.9990
jaspar2022FirstOrderTFFMs	TFFM0040.1	0.9535	0.0034	-5	5	-5	5	plus	plus	loss	-0.9998
jaspar2022DetailedTFFMs	TFFM0040.2	0.8399	0.0565	-5	5	-5	5	plus	plus	loss	-0.9697
jaspar2022FirstOrderTFFMs	TFFM0040.2	0.9155	0.0971	-5	5	-5	5	plus	plus	loss	-0.9957

(Click to see details. The combined score reflects only the TFFMs.) Combined loss -0.9910

INTRODUCTION

Any human individual harbours about 4 million short variants (single nucleotide variants and short indels) in their genome compared to the reference genome (1). Many computer programs are available for assessing the disease-causing potential of variants in coding regions (e.g. MutationTaster (2), PolyPhen (3), and SIFT (4)), but coding regions make up only about 1.5 % of the human genome. Much less is known about the effect of variants in the remaining non-coding DNA, and tools aiming at the prediction of such variants (e.g. CADD (5), Genomiser (6), and RegulationSpotter (7)) are hampered by the lack of known disease mutations outside of protein-coding genes as training cases.

Variants in regions essential for gene expression (especially promoters and enhancers) can alter the binding affinity of transcription factors, leading to up- or down-regulation of transcriptional activity (8). This may result in non-expression or severe under-expression of a gene, the subsequent loss of the encoded protein (null mutation), and lead to disease (9–11).

[*] To whom correspondence should be addressed. Tel: +49 30 450 543685; Email: robin.steinhaus@bih-charite.de

The prediction of transcription factor binding sites (TFBSs) presents an ongoing challenge in computational biology (12,13). The standard method for assessing the binding affinity of a transcription factor to DNA in silico is to compare the DNA bases with a position weight matrix (PWM) specific for the transcription factor. A PWM model is obtained from an assumed common binding motif by counting adenine, cytosine, guanine, and thymine bases at each position in experimentally-confirmed binding sites for a transcription factor. Many thousands of PWM profiles have been published in open-access databases (e.g. JASPAR (14), HOCOMOCO (15), and SwissRegulon (16)).

PWMs are relatively simple models that ignore the positional dependencies that have been repeatedly observed in TFBSs (17–20). More advanced models have in many cases been shown to give better results in identifying experimentally verified binding sites (21–23). A number of alternative modelling approaches have been proposed, several of which attempt to integrate dependencies between adjacent and/or distant positions. These include the binding energy model (BEM) (24), dinucleotide weight matrices (DWMs) (25), and transcription factor flexible models (TFFMs) (22). Among these, TFFMs have been gaining visibility since their inclusion in the JASPAR database for transcription factors, which includes >1000 human TFFMs in its current release (14). TFFMs are based on hidden Markov models (HMMs) and can account for complex positional dependencies as well as variable length nucleotide patterns. TFFM motifs are defined in terms of HMM states, transitions between states, initials, and emissions. Two types of TFFMs are commonly used and supported in FABIAN-variant: In first-order TFFMs, each position within a TFBS is represented by a HMM state emitting a nucleotide with probabilities dependent on the nucleotide found at the previous position. In detailed TFFMs, each HMM state in the first-order HMM is decomposed into four states (one per nucleotide) and transition probabilities reflect the emission probabilities of the first-order HMM (22). Like PWMs, TFFMs are derived from experimentally verified binding sites. Unlike PWMs, TFFMs cannot be evaluated using standard mathematical operations, but require dedicated algorithms for HMMs. Although several tools for evaluating PWMs exist (e.g. FIMO (26), motifbreakeR (27), and Pscan (28)), we have only found one other web application that works with TFFMs (TFBSPred (29)). However, TFBSPred does not provide a mechanism for evaluating DNA variants.

This article introduces a new user-friendly web application for predicting the effects of DNA variants on transcription factor binding. FABIAN-variant offers 1224 TFFMs and 3790 PWMs from different databases for 1387 different human transcription factors. It has different modes for analysing single variants, lists of variants, or up to 10 000 variants from a VCF file. For each variant and transcription factor, FABIAN-variant evaluates available models in the 'wild-type' and variant sequence and returns a combined score indicating whether or not and to which degree transcription factor binding may be affected. The backend is written in C++ for speed and most types of analysis are completed in just a few seconds. For VCF-based analyses, users can choose to be notified by email once the run completes. Results are visualised in the browser and can also be downloaded. Various filters for regions, genes, variants, and transcription factors are implemented (e.g. search in promoter regions of candidate genes, search with TFFMs only). Genome builds GRCh37 (hg19) and GRCh38 (hg38) are supported. FABIAN-variant is free and open to all users without login requirement.

SOFTWARE/BACKEND

Evaluation of TFFMs and PWMs

Different models for human TFBSs (TFFMs and PWMs) were downloaded from various data sources (Table 1) and imported into FABIAN-variant.

FABIAN-variant evaluates each selected TFFM and PWM model in a sliding window from −15 to +15 nucleotides around the variant location in both the reference sequence ('wild-type', *WT*) and the variant sequence ('mutated', *MT*). Both strands are considered. Then the highest scores for both sequences ($0 \le WT, MT \le 1$) are compared for each model. A greater *WT* score indicates a weakened binding affinity, and a greater *MT* score indicates an increased binding affinity caused by the variant. For each model, FABIAN-variant generates a joint score *S* between −1 (likely TFBS loss) and 1 (likely TFBS gain),

$$S = \frac{2}{1 + 2^{-2F}} - 1, \quad F = \begin{cases} -\dfrac{1 - MT + \alpha}{1 - WT + \alpha} + 1 & WT > MT \\[2ex] \dfrac{1 - WT + \alpha}{1 - MT + \alpha} - 1 & WT \le MT \end{cases}$$

with pseudocount $\alpha = 0.1$ to avoid zero in the denominator. We use the inverse of *WT* and *MT* in the ratio (e.g. $1 - WT$) to account for the fact that PWM and TFFM scores correlate with the likelihood that binding is possible in the first place and the ratio is comparatively higher with small denominators. *WT*, *MT* and the joint score *S* per model are shown on the results page of FABIAN-variant.

For most transcription factors, several models (TFFMs and PWMs) exist. To obtain the combined prediction per variant per transcription factor, FABIAN-variant calculates the average of joint scores *S* of the individual models. If both TFFMs and PWMs are available, FABIAN-variant by default uses only the results from TFFMs for the combined prediction (this setting can be changed on the results page so that both types of models are included in the combined score).

Known TFBSs

To allow the restriction to known TFBSs, we collected data from ChIP-seq experiments from ENCODE (30) and Ensembl Regulation (31), as well as from cap analysis of gene expression (CAGE) experiments from FANTOM5 (32) (Table 1). Please note that TFBSs obtained from ChIP-seq experiments are regions of several hundred bases, whereby the precise location of the actual TFBS within the region is however unknown.

We did not use Ensembl's motif-derived predicted binding sites.

Table 1. Data sources included in FABIAN-variant

Source	Data	URL	Reference
JASPAR 2022	612 detailed TFFMs	https://jaspar.genereg.net/	(14)
JASPAR 2022	612 first-order TFFMs	https://jaspar.genereg.net/	(14)
JASPAR 2022	877 PWMs	https://jaspar.genereg.net/	(14)
MotifDb 1.36.0	*	https://doi.org/10.18129/B9.bioc.MotifDb	(41)
CIS-BP 1.02	313 PWMs	http://cisbp.ccbr.utoronto.ca/	(42)
HOCOMOCO 11	768 PWMs	https://hocomoco11.autosome.org/	(15)
hPDI	436 PWMs	http://bioinfo.wilmer.jhu.edu/PDI/	(43)
Jolma 2013	710 PWMs	https://doi.org/10.1016/j.cell.2012.12.009	(44)
SwissRegulon	684 PWMs	https://swissregulon.unibas.ch/sr/	(16)
UniPROBE	2 PWMs	http://the_brain.bwh.harvard.edu/uniprobe/	(45)
ENCODE 3	7,374,455 TFBSs†	https://www.encodeproject.org/	(30)
Ensembl Regulation 102	7,808,345 TFBSs†	https://www.ensembl.org/	(31)
FANTOM5 SSTAR	4,987 TFBSs†	https://fantom.gsc.riken.jp/5/sstar/	(32)

Data included in this table and in FABIAN-variant is for human transcription factors only.
*MotifDb 1.36.0 is an annotated collection of PWM models, and we obtained all PWMs listed in this table except for JASPAR 2022 from MotifDb.
†Data for genome build GRCh37 is shown.

FEATURES/FRONTEND

Search interface

On the search page (Figure 1), users can select transcription factors and choose models to be included in the search. Variants can either be entered into a text field or uploaded as a VCF file. In the latter case, we provide filter options for candidate gene regions, custom genomic regions, coverage, homozygosity, and restriction to rare variants using data from gnomAD (33) and the 1000 Genomes Project (34).

Users can choose to include all 5014 models and all 1387 transcription factors or limit the search to specific factors and models. The search can be restricted to transcription factors for which there are known binding sites at the genomic location of the variant. Other options are to use only TFFMs, only PWMs, or only models from a specific database.

Results for search of a single variant are available immediately after clicking on ‘Analyse’. If all models and not >100 variants are included in the search, FABIAN-variant usually returns the results in <90 s.

Results overview

A sample results page for two pathogenic promoter variants is shown in Figure 2 (chr1:155271258T>C has been reported to disable binding of the erythroid transcription factor GATA1 (35) and chr1:160001799G>C to disrupt binding of SP1 (36)). Coloured cells indicate the likelihood of a loss (red) or gain (blue) of a TFBS due to the variant based on the combined prediction of different models per variant. Deeper shades of the colour represent a greater loss or gain. Moving the mouse pointer over a coloured cell reveals the individual model scores.

The results page provides access to all results for all variants, transcription factors, and models included in the search. Because the amount of results can be overwhelming for large searches, there are several sorting and filtering options at the top that can be used to reorder or hide information on the page. Changes to the options are immediately reflected in the results table. The filter options allow the user to only show transcription factors within a specified genomic region, with a predicted loss or gain of a TFBS, with a known TFBS at the location of a variant, or those which are manually selected with the mouse pointer. FABIAN-variant automatically applies pagination beyond 100 variants. We provide a download of the complete results in TSV format and a summary based on the selected filters.

Detailed results

Clicking on a coloured cell brings up detailed results for an individual variant and a single transcription factor (Figure 3). The page includes the model scores, an option to print results, reference and variant sequences, a list of all known TFBSs at the variant location, as well as sequence logos for the different models.

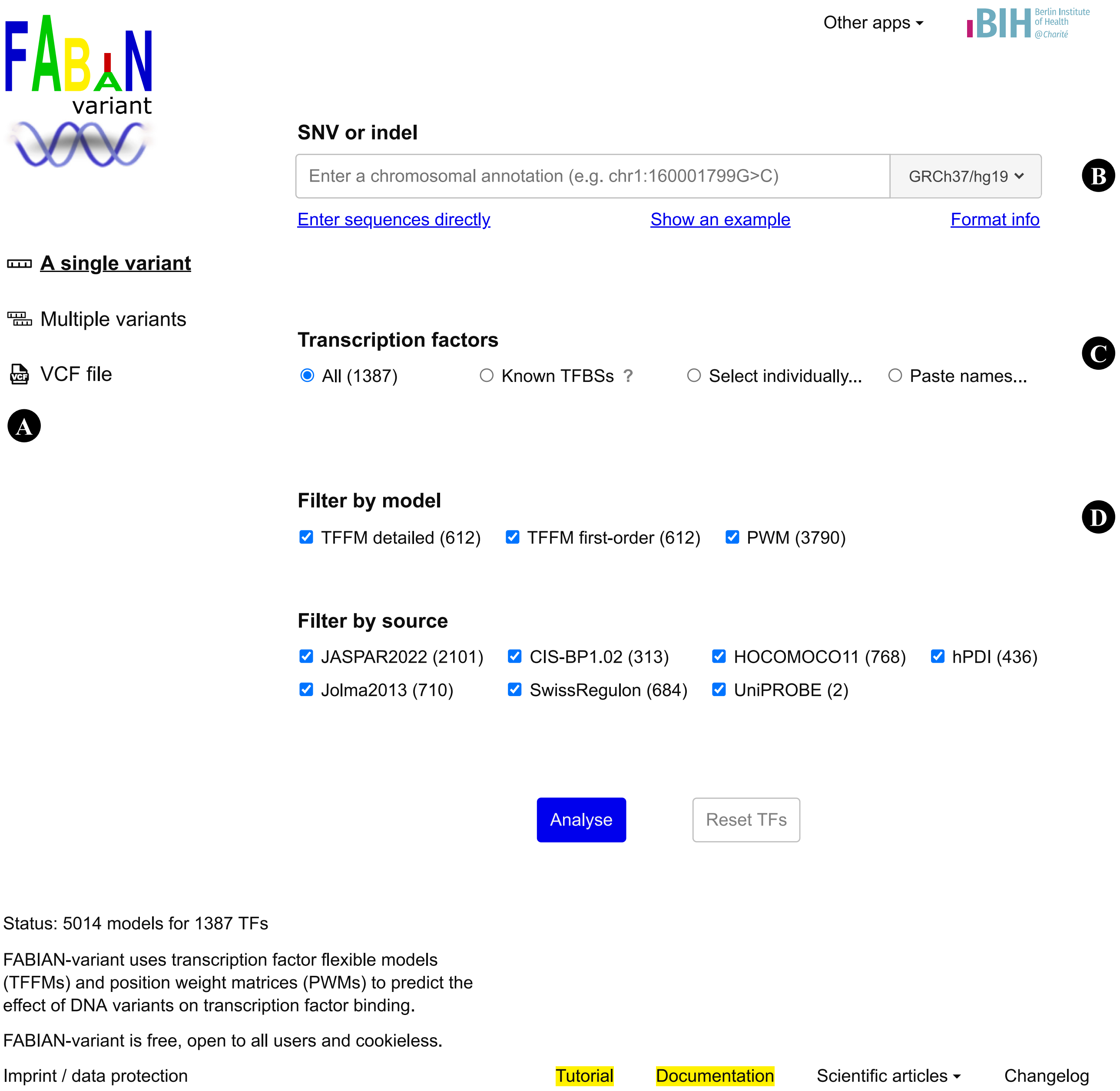
IMPLEMENTATION

FABIAN-variant is an acronym for FAst BInding-site ANalysis and has been optimised for computational efficiency. The FABIAN-variant web server uses Perl CGI to run the C++ backend and a PostgreSQL database. The frontend includes JavaScript and Ajax for interactivity. Job scheduling is provided by Slurm. The code does not use other third-party libraries.

Each TFFM score is computed with a custom C++ implementation of the forward-backward algorithm from the GHMM library (37). Position count matrices (PCMs) were converted to PWMs using the method described in (38) based on the background nucleotide distribution in the human genome.

DISCUSSION

Since their inclusion in the JASPAR database for transcription factors, TFFMs are gaining visibility.



SNV or indel

Enter a chromosomal annotation (e.g. chr1:160001799G>C) | GRCh37/hg19 ▾

Enter sequences directly | Show an example | Format info

Transcription factors

● All (1387) ○ Known TFBSs ? ○ Select individually... ○ Paste names...

Filter by model

☑ TFFM detailed (612) ☑ TFFM first-order (612) ☑ PWM (3790)

Filter by source

☑ JASPAR2022 (2101) ☑ CIS-BP1.02 (313) ☑ HOCOMOCO11 (768) ☑ hPDI (436)
☑ Jolma2013 (710) ☑ SwissRegulon (684) ☑ UniPROBE (2)

Status: 5014 models for 1387 TFs

FABIAN-variant uses transcription factor flexible models (TFFMs) and position weight matrices (PWMs) to predict the effect of DNA variants on transcription factor binding.

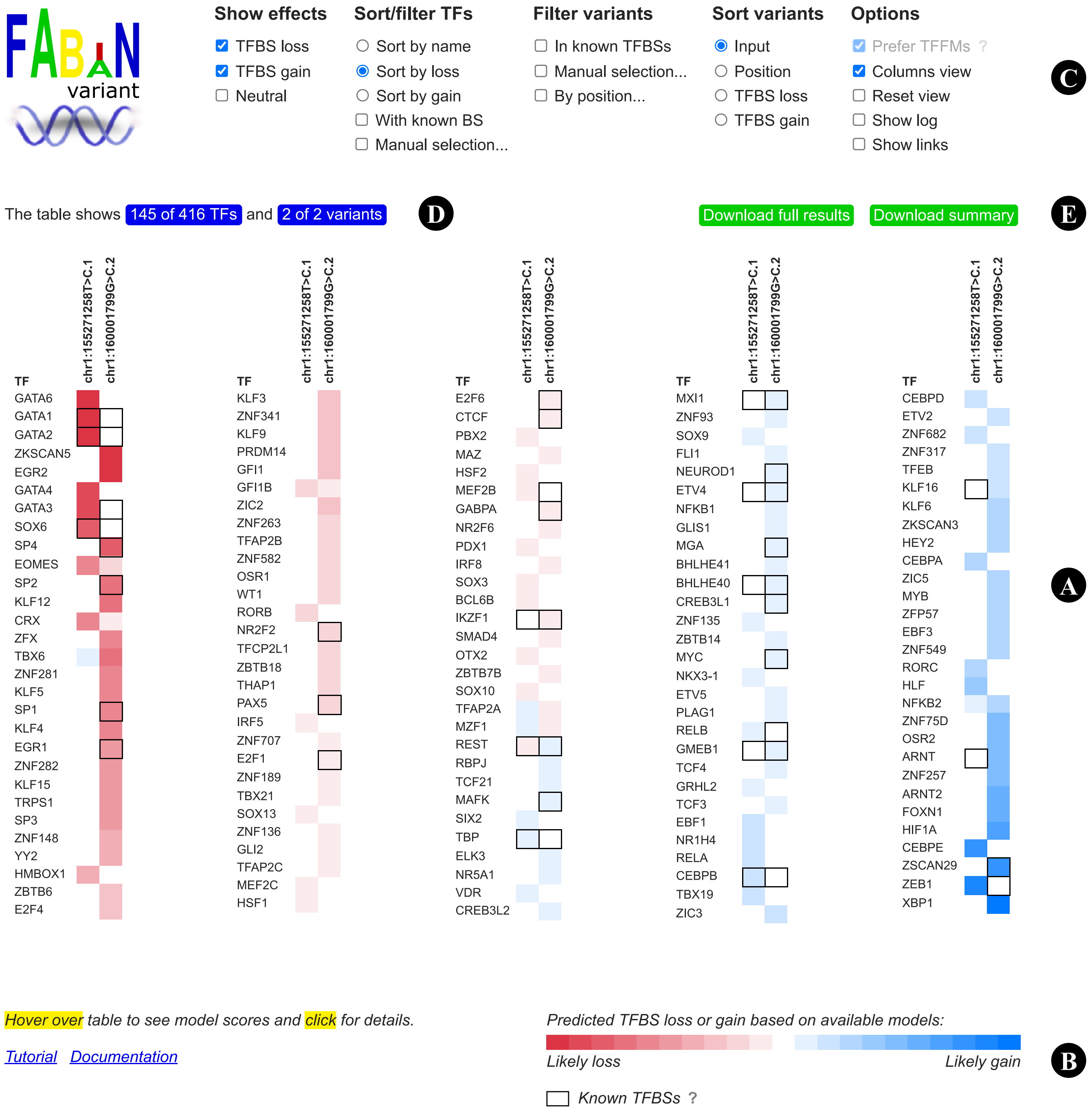
FABIAN-variant is free, open to all users and cookieless.

Figure 1. FABIAN-variant interface for a single variant. (**A**) Users can choose between a single variant, multiple variants, or a VCF file. (**B**) Input field for a chromosomal annotation of a single variant (e.g. 1:160001799G>C). Variants can also be entered as nucleotide sequences by clicking 'Enter sequences directly' (e.g. GGCCCTC...>TCACACT...). (**C**) 'Known TFBSs' searches for transcription factors known to bind at the location of the variant based on ENCODE, Ensembl, or FANTOM5 data. 'Select individually...' and 'Paste names...' open fields to submit a custom set of transcription factors. (**D**) The type of models can be restricted to TFFMs, PWMs, or data from specific sources. The numbers in parentheses update automatically and indicate the number of models included in the search based on the current input.

FABIAN-variant is the first web application that can not only analyse variant effects with PWMs but also with TFFMs.

Because there are millions of non-coding variants in any human genome, it is not helpful to search for effects on transcription factor binding for all of them – one would drown in results. However, the search for potentially regulatory variants may be very helpful if restricted to candidate genes known to be involved in the patient's disease. This might also reveal the 'second mutation' in case of recessive disorders where likely deleterious variants such as premature termination codons are only found on one allele.

Although FABIAN-variant is in principle capable of analysing all variants found in a typical whole-genome sequencing project, we have reduced the number of variants subjected to analysis to 10 000. Generating millions of results for each of the 1387 transcription factors covered by FABIAN-variant would lead to a plethora of data nobody would or could study. Instead, we provide filter options to



Show effects
☑ TFBS loss
☑ TFBS gain
☐ Neutral

Sort/filter TFs
○ Sort by name
● Sort by loss
○ Sort by gain
☐ With known BS
☐ Manual selection...

Filter variants
☐ In known TFBSs
☐ Manual selection...
☐ By position...

Sort variants
● Input
○ Position
○ TFBS loss
○ TFBS gain

Options
☑ Prefer TFFMs ?
☑ Columns view
☐ Reset view
☐ Show log
☐ Show links

The table shows 145 of 416 TFs and 2 of 2 variants

Download full results
Download summary

Hover over table to see model scores and click for details.

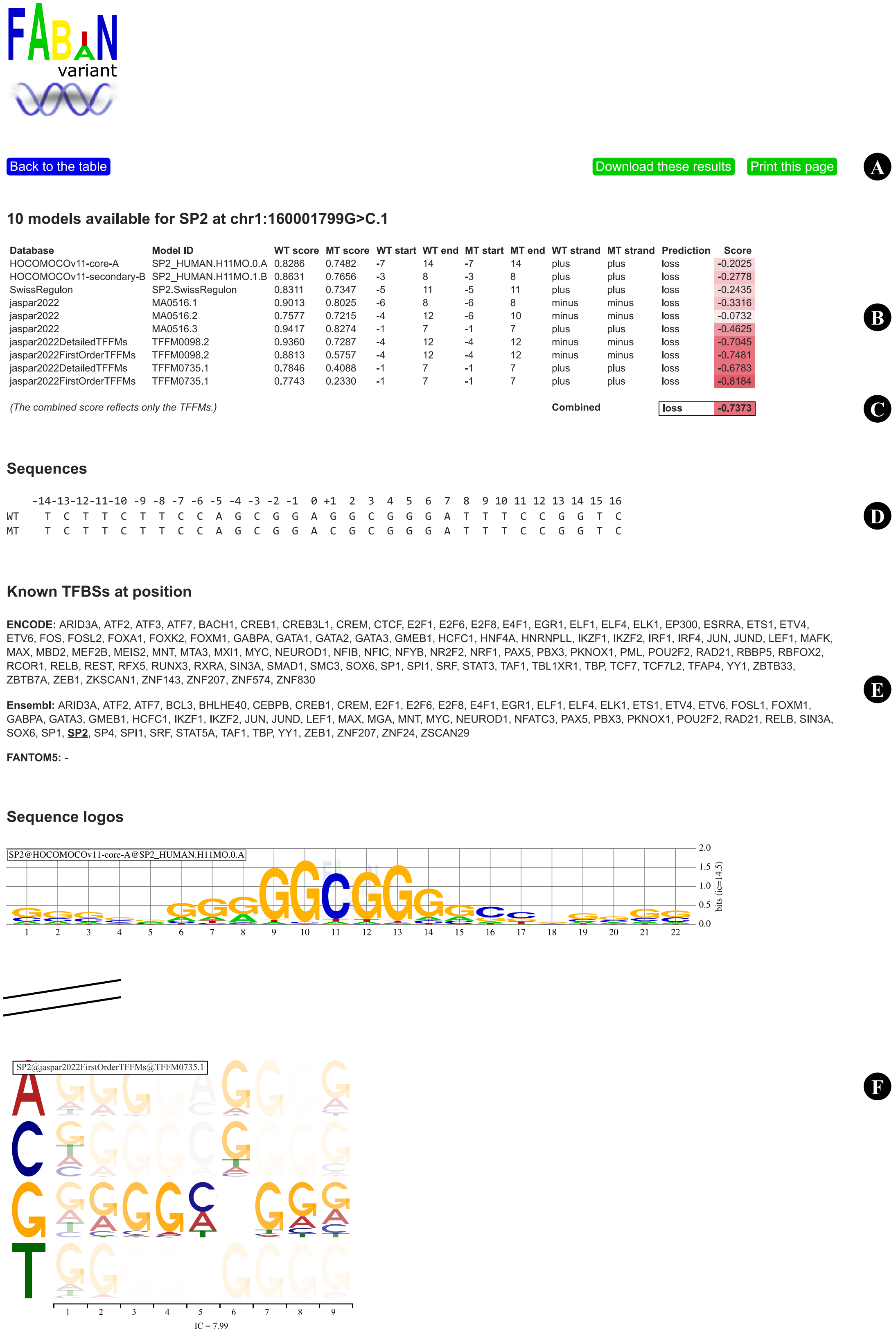
Tutorial Documentation

Predicted TFBS loss or gain based on available models:

Likely loss Likely gain

□ Known TFBSs ?

Figure 2. Results page for two promoter variants. (**A**) The results are divided into five sections for better readability. Variants are plotted in columns, transcription factors in rows. Coloured cells indicate the potential loss (red) or gain (blue) of a TFBS due to the variant. (**B**) Legend. Deeper shades of red or blue represent a greater loss or gain. Known TFBSs at the location of a variant are displayed with a border around the cell. Please note that the TFBSs obtained from ChIP-seq experiments are regions of several hundred bases and we do not know where within these stretches the real binding sites are located. (**C**) Users can define sorting and filters to limit the displayed data. (**D**) 145 transcription factors are currently shown in the table. The number is refreshed automatically based on active filters. (**E**) Results can be downloaded.



Back to the table

Download these results

Print this page

10 models available for SP2 at chr1:160001799G>C.1

Database	Model ID	WT score	MT score	WT start	WT end	MT start	MT end	WT strand	MT strand	Prediction	Score
HOCOMOCOv11-core-A	SP2_HUMAN.H11MO.0.A	0.8286	0.7482	-7	14	-7	14	plus	plus	loss	-0.2025
HOCOMOCOv11-secondary-B	SP2_HUMAN.H11MO.1.B	0.8631	0.7656	-3	8	-3	8	plus	plus	loss	-0.2778
SwissRegulon	SP2.SwissRegulon	0.8311	0.7347	-5	11	-5	11	plus	plus	loss	-0.2435
jaspar2022	MA0516.1	0.9013	0.8025	-6	8	-6	8	minus	minus	loss	-0.3316
jaspar2022	MA0516.2	0.7577	0.7215	-4	12	-6	10	minus	minus	loss	-0.0732
jaspar2022	MA0516.3	0.9417	0.8274	-1	7	-1	7	plus	plus	loss	-0.4625
jaspar2022DetailedTFFMs	TFFM0098.2	0.9360	0.7287	-4	12	-4	12	minus	minus	loss	-0.7045
jaspar2022FirstOrderTFFMs	TFFM0098.2	0.8813	0.5757	-4	12	-4	12	minus	minus	loss	-0.7481
jaspar2022DetailedTFFMs	TFFM0735.1	0.7846	0.4088	-1	7	-1	7	plus	plus	loss	-0.6783
jaspar2022FirstOrderTFFMs	TFFM0735.1	0.7743	0.2330	-1	7	-1	7	plus	plus	loss	-0.8184
(The combined score reflects only the TFFMs.)								**Combined**		loss	-0.7373

Sequences

```
    -14-13-12-11-10 -9 -8 -7 -6 -5 -4 -3 -2 -1  0 +1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
WT    T  C  T  T  C  T  T  C  C  A  G  C  G  G  A  G  G  C  G  G  G  A  T  T  T  C  C  G  G  T  C
MT    T  C  T  T  C  T  T  C  C  A  G  C  G  G  A  C  G  C  G  G  G  A  T  T  T  C  C  G  G  T  C
```

Known TFBSs at position

ENCODE: ARID3A, ATF2, ATF3, ATF7, BACH1, CREB1, CREB3L1, CREM, CTCF, E2F1, E2F6, E2F8, E4F1, EGR1, ELF1, ELF4, ELK1, EP300, ESRRA, ETS1, ETV4, ETV6, FOS, FOSL2, FOXA1, FOXK2, FOXM1, GABPA, GATA1, GATA2, GATA3, GMEB1, HCFC1, HNF4A, HNRNPLL, IKZF1, IKZF2, IRF1, IRF4, JUN, JUND, LEF1, MAFK, MAX, MBD2, MEF2B, MEIS2, MNT, MTA3, MXI1, MYC, NEUROD1, NFIB, NFIC, NFYB, NR2F2, NRF1, PAX5, PBX3, PKNOX1, PML, POU2F2, RAD21, RBBP5, RBFOX2, RCOR1, RELB, REST, RFX5, RUNX3, RXRA, SIN3A, SMAD1, SMC3, SOX6, SP1, SPI1, SRF, STAT3, TAF1, TBL1XR1, TBP, TCF7, TCF7L2, TFAP4, YY1, ZBTB33, ZBTB7A, ZEB1, ZKSCAN1, ZNF143, ZNF207, ZNF574, ZNF830

Ensembl: ARID3A, ATF2, ATF7, BCL3, BHLHE40, CEBPB, CREB1, CREM, E2F1, E2F6, E2F8, E4F1, EGR1, ELF1, ELF4, ELK1, ETS1, ETV4, ETV6, FOSL1, FOXM1, GABPA, GATA3, GMEB1, HCFC1, IKZF1, IKZF2, JUN, JUND, LEF1, MAX, MGA, MNT, MYC, NEUROD1, NFATC3, PAX5, PBX3, PKNOX1, POU2F2, RAD21, RELB, SIN3A, SOX6, SP1, **SP2**, SP4, SPI1, SRF, STAT5A, TAF1, TBP, YY1, ZEB1, ZNF207, ZNF24, ZSCAN29

FANTOM5: -

Sequence logos

Figure 3. The detailed results page is shown after clicking on the corresponding cell in the results table. (**A**) Options to download or print details on this page. (**B**) Six PWMs and four TFFMs for transcription factor SP2 were evaluated for variant chr1:160001799G>C (GRCh37). Higher scores in the reference sequence (WT) than in the variant sequence (MT) indicate a possible loss of a TFBS. (**C**) The combined prediction is shown below the individual model scores. (**D**) Reference and variant sequences (variant at position +1). (**E**) A list of known TFBSs at the variant location. (**F**) Sequence logos and information content for the ten models (abridged in the Figure).

restrict the analysis to variants found in a specific region or near functional or positional candidate genes.

A limitation of FABIAN-variant is that larger deletions that abolish the complete TFBS cannot be analysed because our application is aimed at the analysis of variants within the TFBS.

FABIAN-variant is aimed at the fast analysis of the variant effect on transcription factor binding on the sequence level and does not consider regulatory features that might affect transcription factor binding (e.g. chromatin accessibility, as implemented in SEMpl (39)).

OUTLOOK

The infrastructure of FABIAN-variant has been implemented in a way that is easily extensible when new versions of the underlying data are released. Additionally, we are considering adding deep learning-based models such as DeepBind (40) to the application.

We also plan to provide a simple API for the analysis of single variants from within other applications.

DATA AVAILABILITY

FABIAN-variant can be accessed at https://www.genecascade.org/fabian/. This website is free and open to all users without login requirement or use of cookies.

The results page for each analysis has a unique URL that can be used to access, share, or download results at a later time. Results are kept on the server for three days, after which time they are automatically deleted. Users can also choose to directly delete their data on the results page.

Links to the documentation and a tutorial are provided on the homepage. The documentation has a link to download the underlying TFFM and PWM model definitions.

FUNDING

PNR was supported by NIH NICHD [5R01HD103805-02]. DS was supported by Deutsche Forschungsgemeinschaft (DFG) [FOR2841 TP05, TP09]. Funding for open access charge: Open Access Publication Fund of the Charité – Universitätsmedizin Berlin.

Conflict of interest statement. None declared.

REFERENCES

1. Reuter,J.A., Spacek,D.V. and Snyder,M.P. (2015) High-throughput sequencing technologies. *Mol. Cell*, **58**, 586–597.
2. Steinhaus,R., Proft,S., Schuelke,M., Cooper,D.N., Schwarz,J.M. and Seelow,D. (2021) MutationTaster2021. *Nucleic Acids Res.*, **49**, W446–W451.
3. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
4. Sim,N.-L., Kumar,P., Hu,J., Henikoff,S., Schneider,G. and Ng,P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452–W457.
5. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
6. Smedley,D., Schubach,M., Jacobsen,J.O., Köhler,S., Zemojtel,T., Spielmann,M., Jäger,M., Hochheiser,H., Washington,N.L., McMurry,J.A. *et al.* (2016) A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.*, **99**, 595–606.
7. Schwarz,J.M., Hombach,D., Köhler,S., Cooper,D.N., Schuelke,M. and Seelow,D. (2019) RegulationSpotter: annotation and interpretation of extratranscriptic DNA variants. *Nucleic Acids Res.*, **47**, W106–W113.
8. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
9. Nougier,C., Roualdes,O., Fretigny,M., d'Oiron,R., Costa,C., Negrier,C. and Vinciguerra,C. (2014) Characterization of four novel molecular changes in the promoter region of the factor VIII gene. *Haemophilia*, **20**, e149–e156.
10. Xu,Y., Krishnan,A., Wan,X.S., Majima,H., Yeh,C.-C., Ludewig,G., Kasarskis,E.J. and St Clair,D.K. (1999) Mutations in the promoter reveal a cause for the reduced expression of the human manganese superoxide dismutase gene in cancer cells. *Oncogene*, **18**, 93–102.
11. Jang,Y.J., LaBella,A.L., Feeney,T.P., Braverman,N., Tuchman,M., Morizono,H., Ah Mew,N. and Caldovic,L. (2018) Disease-causing mutations in the promoter and enhancer of the ornithine transcarbamylase gene. *Hum. Mutat.*, **39**, 527–536.
12. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
13. Hombach,D., Schwarz,J.M., Robinson,P.N., Schuelke,M. and Seelow,D. (2016) A systematic, large-scale comparison of transcription factor binding site models. *BMC Genomics*, **17**, 1–10.
14. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A., Manosalva Pérez,N. *et al.* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **50**, D165–D173.
15. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
16. Pachkov,M., Balwierz,P.J., Arnold,P., Ozonov,E. and Van Nimwegen,E. (2012) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res.*, **41**, D214–D220.
17. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
18. Man,T.-K. and Stormo,G.D. (2001) Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
19. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*. pp. 28–37.
20. Mukherjee,S., Berger,M.F., Jona,G., Wang,X.S., Muzzey,D., Snyder,M., Young,R.A. and Bulyk,M.L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.
21. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
22. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
23. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Soboleva,A.V., Kasianov,A.S., Ashoor,H., Ba-Alawi,W., Bajic,V.B., Medvedeva,Y.A., Kolpakov,F.A. *et al.* (2016) HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, **44**, D116–D125.
24. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.

25. Siddharthan,R. (2010) Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS One*, **5**, e9722.
26. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
27. Coetzee,S.G., Coetzee,G.A. and Hazelett,D.J. (2015) motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics*, **31**, 3847–3849.
28. Zambelli,F., Pesole,G. and Pavesi,G. (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.*, **37**, W247–W252.
29. Zogopoulos,V.L., Spaho,K., Ntouka,C., Lappas,G.A., Kyranis,I., Bagos,P.G., Spandidos,D.A. and Michalopoulos,I. (2021) TFBSPred: A functional transcription factor binding site prediction webtool for humans and mice. *Int. J. Epigenet.*, **1**, 1–11.
30. Consortium,E.P. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.
31. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
32. Abugessaisa,I., Shimoji,H., Sahin,S., Kondo,A., Harshbarger,J., Lizio,M., Hayashizaki,Y., Carninci,P., Forrest,A., Kasukawa,T. *et al.* (2016) FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database*, **2016**, baw105.
33. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P. *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
34. G.P. Consortium2015) A global reference for human genetic variation. *Nature*, **526**, 68.
35. Manco,L., Ribeiro,M.L., Máximo,V., Almeida,H., Costa,A., Freitas,O., Barbot,J., Abade,A. and Tamagnini,G. (2000) A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. *Br. J. Haematol.*, **110**, 993–997.
36. Almeida,A.M., Murakami,Y., Layton,D.M., Hillmen,P., Sellick,G.S., Maeda,Y., Richards,S., Patterson,S., Kotsianidis,I., Mollica,L. *et al.* (2006) Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. *Nat. Med.*, **12**, 846–851.
37. Schliep,A. and Costa,I.G. (2022) General Hidden Markov Model library (GHMM). http://ghmm.sourceforge.net/.
38. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
39. Nishizaki,S.S., Ng,N., Dong,S., Porter,R.S., Morterud,C., Williams,C., Asman,C., Switzenberg,J.A. and Boyle,A.P. (2020) Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics*, **36**, 364–372.
40. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
41. Shannon,P. and Richards,M. (2022) MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs. R package version 1.36.0., https://doi.org/10.18129/B9.bioc.MotifDb.
42. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
43. Xie,Z., Hu,S., Blackshaw,S., Zhu,H. and Qian,J. (2010) hPDI: a database of experimental human protein–DNA interactions. *Bioinformatics*, **26**, 287–289.
44. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
45. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.