D104–D111 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 24 November 2020*
doi: 10.1093/nar/gkaa1057

GTRD: an integrated view of transcription regulation

Semyon Kolmykov[1,2,3], Ivan Yevshin[1,2], Mikhail Kulyashov[1,2,4], Ruslan Sharipov[1,2,4], Yury Kondrakhin[1,2], Vsevolod J. Makeev[5,6,7,8], Ivan V. Kulakovskiy[5,8,9], Alexander Kel[1,10,11] and Fedor Kolpakov[1,2,*]

[1]BIOSOFT.RU, LLC, Novosibirsk 630090, Russian Federation, [2]Federal Research Center for Information and Computational Technologies, Novosibirsk 630090, Russian Federation, [3]Federal Research Center Institute of Cytology and Genetics SB RAS, Novosibirsk 630090, Russian Federation, [4]Novosibirsk State University, Novosibirsk 630090, Russian Federation, [5]Vavilov Institute of General Genetics RAS, Moscow 119991, Russian Federation, [6]Moscow Institute of Physics and Technology (State University), Dolgoprudny 141700, Russian Federation, [7]NRC ≪Kurchatov Institute≫ - GOSNIIGENETIKA, Kurchatov Genomic Center, Moscow 123182, Russian Federation, [8]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russian Federation, [9]Institute of Protein Research, Russian Academy of Sciences, Pushchino 142290, Russian Federation, [10]geneXplain GmbH, 38302 Wolfenbüttel, Germany and [11]Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk 630090, Russian Federation

Received September 15, 2020; Revised October 18, 2020; Editorial Decision October 19, 2020; Accepted November 03, 2020

ABSTRACT

The Gene Transcription Regulation Database (GTRD; http://gtrd.biouml.org/) contains uniformly annotated and processed NGS data related to gene transcription regulation: ChIP-seq, ChIP-exo, DNase-seq, MNase-seq, ATAC-seq and RNA-seq. With the latest release, the database has reached a new level of data integration. All cell types (cell lines and tissues) presented in the GTRD were arranged into a dictionary and linked with different ontologies (BRENDA, Cell Ontology, Uberon, Cellosaurus and Experimental Factor Ontology) and with related experiments in specialized databases on transcription regulation (FANTOM5, ENCODE and GTEx). The updated version of the GTRD provides an integrated view of transcription regulation through a dedicated web interface with advanced browsing and search capabilities, an integrated genome browser, and table reports by cell types, transcription factors, and genes of interest.

INTRODUCTION

Transcriptional regulation is a complex process that depends on multiple factors (1,2), including:

- transcription factors (TFs), which bind the DNA sites in the gene regulatory regions and activate or repress gene expression through interaction with various cofactors;
- histone modifications (methylation, phosphorylation, acetylation, etc.) that define the state of chromatin and make particular regions active or inactive;
- DNA methylation of cytosine or adenine, which changes the activity of individual promoters or even large loci, mainly by suppressing gene expression and promoting the formation of heterochromatin.

High-throughput sequencing provides a way to assess those factors at the genome-wide scale. Large-scale international collaborations such as ENCODE (3) and FANTOM (4) serve as the gold standard for systematic acquisition, integrative analysis, and sharing of high-quality experimental data. In particular, the ENCODE Encyclopedia brings together the most salient analytical products and provides tools for searching and visualizing these data (5). However, despite being the largest single source, the ENCODE data provide only a limited contribution to the total pool of the data on gene regulation produced in different research laboratories. Such data are available in GEO (6) and SRA (7) repositories but lack uniform annotation and processing pipelines, thus limiting the possibilities for large-scale integrated analysis. This motivates the development and application of standardized workflows and databases to allow for efficient usage of the vast but unsystematic published data.

In pursuing this goal, we started developing the Gene Transcription Regulation Database (GTRD) in 2011. We started with uniform annotation and analysis of key components of transcription regulation – transcription factor binding sites (TFBSs) – identified in ChIP-seq experiments for humans and mice, the top two species by the

*To whom correspondence should be addressed. Tel: +7 383 363 68 29; Email: fedor@biouml.org

number of ChIP-seq experiments in GEO (8). In the next major release, we extended the GTRD content by seven additional species using the most available TF ChIP-Seq experiments (the complete list: *Homo sapiens, Mus musculus, Rattus norvegicus, Danio rerio, Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae, Schizosaccharomyces pombe* and *Arabidopsis thaliana*), and included the data on the next major level of transcription regulation – the open chromatin regions identified in DNase-seq experiments (9).

The current goal of the GTRD database is to provide uniform annotation and integrative analyses of all next-generation sequencing (NGS) data from GEO and SRA that meet the following criteria: (1) related to transcription regulation; (2) already widespread and actively generated by the research community. With these ideas in mind, we included the following experimental data types:

- ChIP-exo (10) and ChIP-nexus (11) for precise identification of TFBSs,
- ChIP-seq (12) for identification of the histone modifications,
- ATAC-seq (13) and FAIRE-seq (14) for identification of open chromatin regions,
- MNase-seq (15) for assessing nucleosome landscapes,
- RNA-seq from cells with altered TF activity through (knock-out, knock-down), or various types of activation to reveal the TF target genes (16).

We also provide results of extended analysis of the previously included data, such as updated meta-clusters (8), master-TFBS tracks and allele-specific binding sites derived from TF ChIP-Seq data. Below, we briefly describe the current workflow allowing for uniform annotation and analysis of the experimental data in the GTRD database as well as integration with other resources related to transcription regulation (Figure 1), highlighting changes in the current release relative to previous GTRD publications (8,9).

MATERIALS AND METHODS

GTRD content is divided into four sections:

(1) Raw data downloaded from GEO, SRA, ENCODE and modENCODE.
(2) Meta-data generated by automated annotation and expert curation of the raw data. The meta-data contains the uniform annotation of experiments (cell source, experimental conditions, control experiment) and information specific to the data type (e.g. for ChIP-seq experiments, the target TF, and the used antibody). We have developed automation tools that download meta-information from ENCODE, modENCODE, and KnockTF (16) and reduce the efforts for manual annotation. Particularly, a special geominer software (Supplement 1) is used for proper conversion and annotation of the original GEO metadata with control vocabularies and ontologies.
(3) Results of uniform analysis of the raw data. For each type of experiment, we developed specialized workflows for uniform analysis (http://wiki.biouml.org/index.php/GTRD_Workflow). These workflows are executed within our in-house distributed computing solution, e-grid, which allows parallel data processing on multiple computational nodes.
(4) Downloads, the plain text files that contain results of the uniform annotation and analysis of the data described above, which can be used for automated downstream analysis of gene regulation by external tools.

Previous versions of the GTRD used the MySQL database as a backend to store the meta-data and all generated tracks (except for the read alignment BAM files). As the data volume increased, the backend became a bottleneck, and in the recent release, the MySQL database is used only for meta-data. The generated genomic signal tracks are stored directly in the file system as bigBed (17) files, providing multiple advantages:

- bigBed processing is significantly more efficient in e-grid in parallel mode;
- the storage space requirements are much lower due to bigBed internal compression;
- bigBed files can be directly downloaded from the GTRD without additional conversion and directly visualized in UCSC (17) or Ensembl (18) genome browsers.

The alignment data is also available for download in bigWig format (17) suitable for visualization in external genome browsers. Particularly, bigWig files are provided for ChIP-seq peaks coverage, DNase-seq and ATAC-seq coverage of open chromatin regions, and gene expression estimates from RNA-seq data.

We have updated the links to the classification of human and mouse TFs with the most recent TFClass data (19), which allows the use of the GTRD with the database of transcription factors and their motifs, TRANSFAC (20). Furthermore, in this GTRD release, we also linked the TFs with CIS-BP (21), which covers all model organisms of GTRD (except *Schizosaccharomyces pombe*).

An essential task for integrative analysis is the accurate annotation of the experimental data, including experimental conditions with suitable control vocabularies and ontologies. In the current version of the GTRD, we created a single dictionary for cell types that includes 3954 entries (22). For further analyses and visualization, the cell types were divided into 90 clusters corresponding to the main organs and tissues of corresponding organisms. Whenever possible, all entries were linked with the existing cell type dictionaries and ontologies: UBERON (23), Cell Ontology (24), BRENDA tissue ontology (25), Plant Ontology (26), and Cellosaurus (27). A dictionary of experimental conditions was also created and linked with the Experimental Factor Ontology (EFO) (28). Links to these ontologies allow accurate mapping of GTRD data with data from other databases on gene expression regulation, taking into account cell specificity and experimental conditions. This way (8), the GTRD data was arranged with the relevant data on gene expression from FANTOM5 (29) and GTEx (30).

We have developed new methods of integrative analysis, as described below.

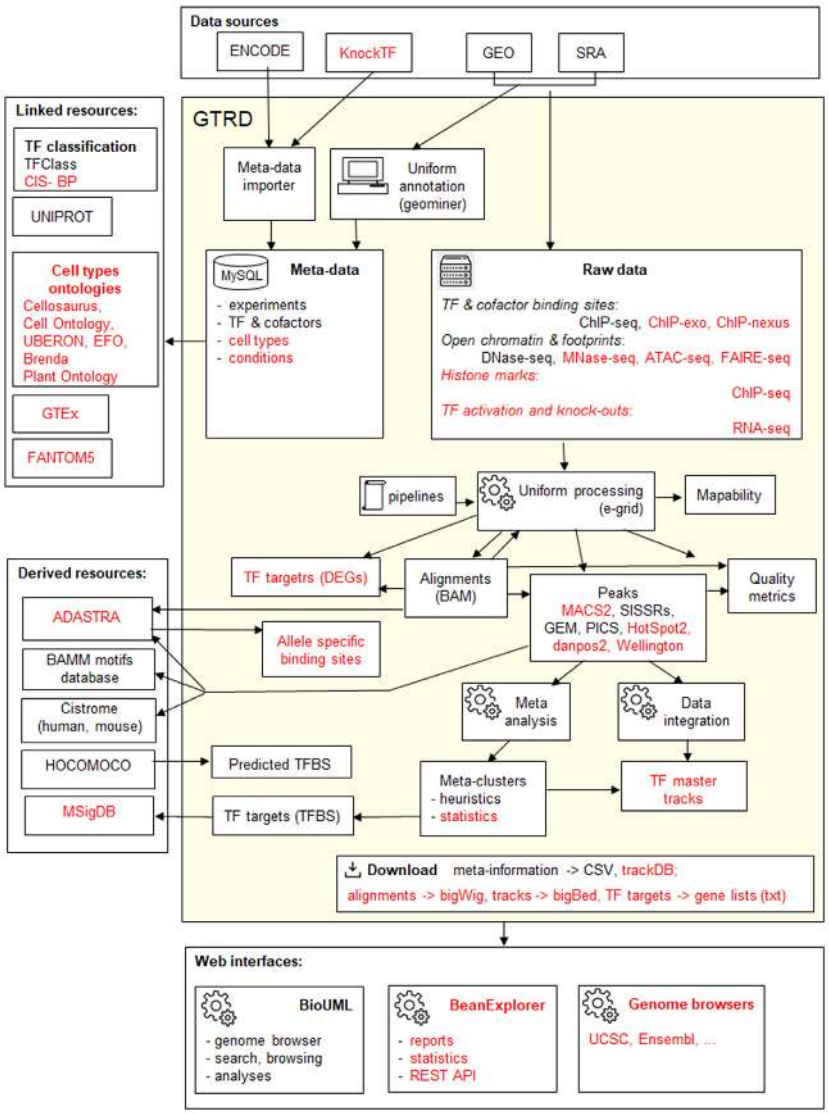

Figure 1. Current workflow for uniform annotation and analysis of experimental data in the GTRD. New data and tools added in the most recent version compared to the previous publication are highlighted in red. DEGs – differentially expressed genes.

Construction of meta-clusters with statistical methods

The meta-clusters are sets of non-redundant TFBSs for each TF that are obtained by merging TFBSs identified by different peak callers in different experiments for the same TF. Previously, we used a heuristic approach (8). Here, we applied a novel algorithm based on quantitative quality metrics (31) and rank aggregation (32) (see Supplement 2 for algorithm details). This not only allows the generation of new versions of meta-clusters but also provides reliability measures in the form of rank aggregation scores, allowing the most reliable meta-clusters to be picked up. Two approaches are suitable for selection of the most reliable meta-clusters (Supplement 3). The first approach is based on a two-component normal mixture, the second uses logistic regression. In general, the most reliable meta-clusters are in good concordance with meta-clusters found with the previous heuristic version (Supplement 4), but have shorter lengths. However, truly reliable meta-clusters require taking into account not only statistical scores but also biological data (location within open chromatin regions, DNase footprints, sequence motifs, etc). To include this information, we introduced the concept of the master track.

Construction of master tracks

The track of the master sites is a further extension of the concept of meta-clusters. For a given TF, the meta-cluster defines the boundaries of the binding site (binding region). The master site for such a region integrates all information from the GTRD: set of peaks used for meta-cluster construction (ChIP-seq, ChIP-exo); motif hits from sequence scanning with position weight matrices (PWMs); cell types where corresponding peaks were identified; overlapping open chromatin regions and DNase or ATAC-seq footprints; allele-specific binding events. This information provides important data on TF binding in different cell types and conditions. The BioUML

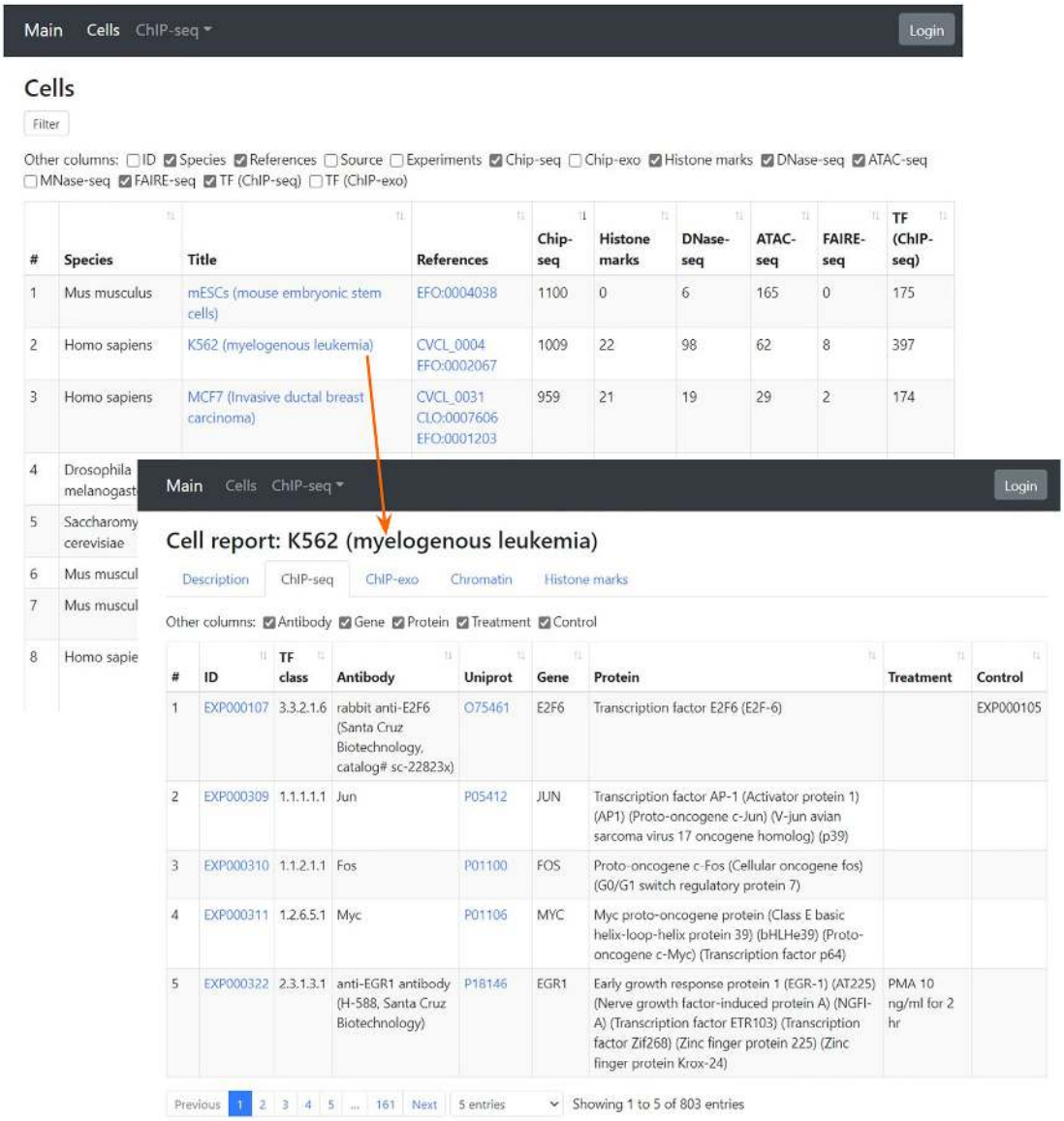

Figure 2. Example of a summary report and an overview of data for a selected cell type.

genome browser provides visualization of this information (Supplement 5).

Identification of the allele-specific binding (ASB) events occurs where a TF demonstrates differential binding to alternating alleles of a single-nucleotide variant (33).

Another important practical task is the identification of TF target genes. A common approach is to consider a gene to be regulated by a given TF if it has a corresponding binding site in the promoter region. For all TFs in the GTRD, we generated lists of genes with binding sites in the promoter region, which is defined in three variants—[−5000; +500], [–1000; +100], [–500; +50] nt relative to the transcription start sites (TSSs)—and in the whole gene [−5000, +5000] relative to gene boundaries. These lists are also included in MSigDB (34) for gene set enrichment analysis. It is noteworthy that only a limited fraction of TFBSs (typically <15%) directly affects gene expression; thus, while being useful for exploratory analysis, these data should be used with caution in other scenarios, such as reconstruction of regulatory networks.

The new GTRD version also provides TF targets revealed by another common approach: analysis of RNA-seq data from cells with down- or upregulated TF activity (knock-out, knock-down, or activation experiments). Depending on experiment conditions, such data can still reveal indirect TF targets—e.g., if RNA-seq is performed later than one hour after corresponding TF activation.

The new GTRD release features a significantly updated web interface provided by the BioUML platform (9): it provides a data browser, advanced search capabilities, and an integrated genome browser for all data types included in the GTRD. A novel 'Track finder' panel was added to the BioUML genome browser for advanced search of genomic tracks by different criteria (see Supplement 6).

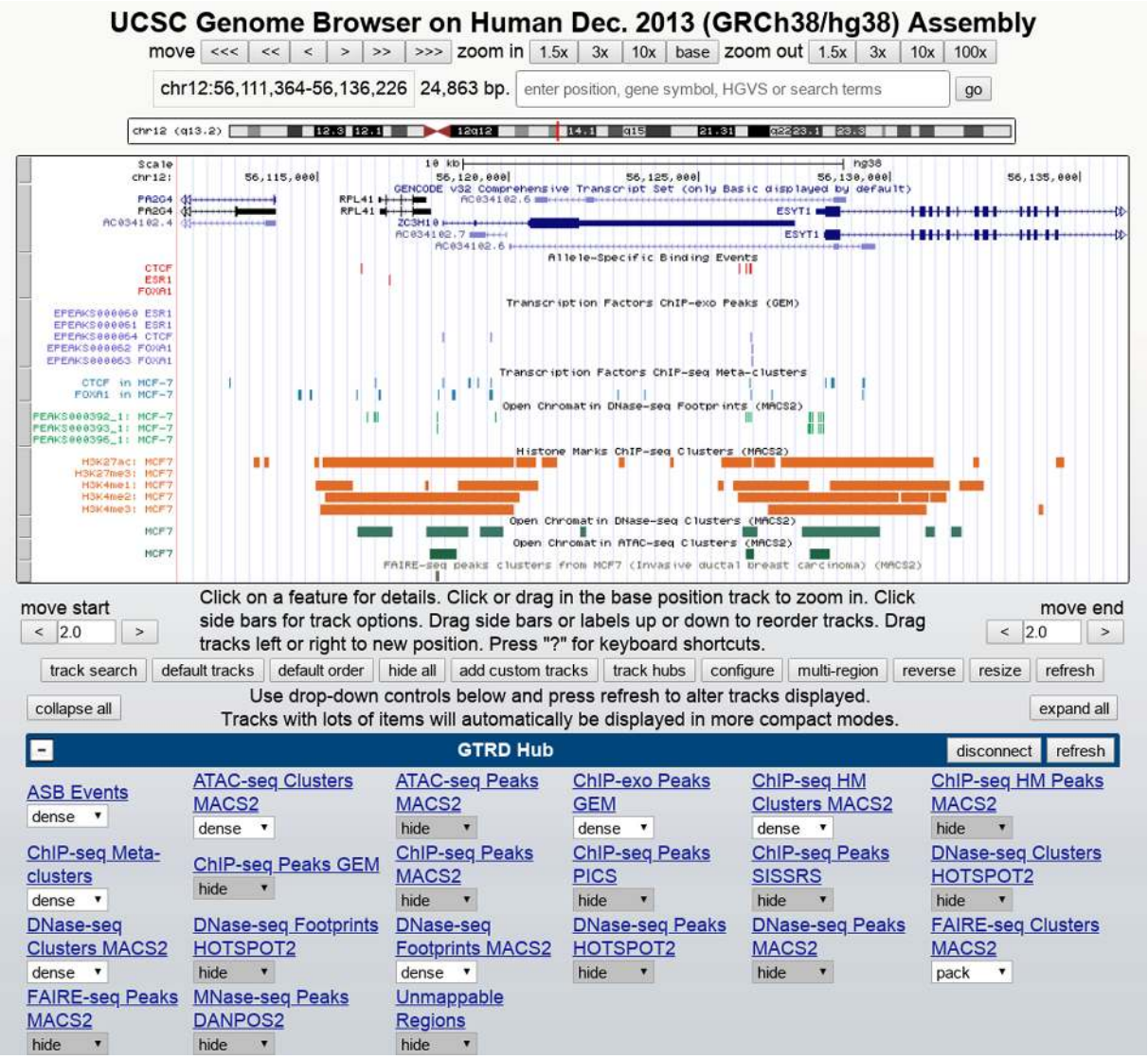

Figure 3. Visualization of GTRD tracks in the UCSC genome browser.

We also present a completely new web interface for generating summary statistics and reports for a given cell type, TF, and experiment from GTRD meta-data (Figure 2). Here, we applied the BeanExplorer technology (https://github.com/DevelopmentOnTheEdge/beanexplorer) that easily allows to generate a web interface based on information from any relational database.

The bigBed and Big files from the GTRD website can be directly visualized by the UCSC (36) and Ensembl (18) genome browsers (Figure 3) with the dedicated track hubs created for each species. Each track hub contains main tracks from the GTRD database in bigBed and bigWig formats and the necessary meta-information following the UCSC track hub standards (37).

We would like to highlight that information from the GTRD was successfully used in several derived resources, such as: (i) the HOCOMOCO collection of TFBS models (38), which is based on GTRD ChIP-Seq peaks and provides the resulting models for usage within the GTRD for genome scanning and TFBS recognition within ChIP-Seq peaks; (ii) the BaMM motifs database (39) and the BaMM server (40) for the recognition of TFBSs; (iii) human and mouse cistromes, the maps of putative cis-regulatory regions bound by TFs, as an early approach to the construction of meta-clusters (41); (iv) MSigDB (34) includes a GTRD subset of TF targets—genes contain TFBSs identified in ChIP-seq experiments in the region [−1000,+500] nt around the TSS; (v) ADASTRA database of allele-specific binding sites (https://adastra.autosome.ru/).

DISCUSSION

TF ChIP-seq was the first type of experimental data reprocessed and provided in the GTRD, which remains the largest database by the number of TFs for which available NGS data (ChIP-seq, ChIP-exo, ChIP-nexus) were uniformly annotated and processed (Table 1). The number of ChIP-seq experiments in the GTRD has doubled since our

Table 1. Comparison of the GTRD with other databases on ChIP-seq experiments

Database	Number of TF ChIP-seq experiments	Number of TFs	ChIP-seq peak callers	Species
GTRD v20.06	total: 35 719 *H. sapiens:* 15 982	total: 3 599 *H. sapiens:* 1391	MACS2, MACS, GEM, PICS, SISSRs	*H. sapiens, M. musculus, R. norvegicus,D. melanogaster, C. elegans, D. rerio, S. pombe, S. cerevisiae, A. thaliana*
ChIP-Atlas	total: 30 495* *H. sapiens:* 13 558*	total: 1 781** *H. sapiens:* 1 020**	MACS2	*H. sapiens, M. musculus, R. norvegicus, D. melanogaster, C. elegans, S. cerevisiae*
CistromeDB	total: 24 065** *H. sapiens:* 13 976**	total: 1654** *H. sapiens:* 1470**	MACS2	*H. sapiens, M. musculus*
ENCODE	total: 3 816 *H. sapiens:* 2 632	total: 2 160 *H. sapiens:* 964	SPP	*H. sapiens, M. musculus, D. melanogaster, C. elegans*
ChIPBase	total: 4300** *H. sapiens:* 2498**	total: 870** *H. sapiens:* 480**	no uniform pipeline, each ChIP-seq is processed by different peak caller	*H. sapiens, M. musculus, R. norvegicus,D. rerio, X. tropicalis, C. elegans,D. melanogaster, S. cerevisiae, A. thaliana, G. gallus*
ReMap 2020 3rd release	total: 6307** *H. sapiens:* 5798**	total: 1507** *H. sapiens:* 1135**	MACS2	*H. sapiens, A. thaliana*
Factorbook	total: 1886** *H. sapiens:* 1813**	total: 682** *H. sapiens:* 682**	None	*H. sapiens, M. musculus*

*'TF and others' according to ChIP-Atlas info. 7 374 input files for human and 17 914 input files in total given in a separate category in ChIP-Atlas are not taken into account in the table as they are not assigned to particular ChIP-Seq data.
**TFs and other DNA binding proteins excluding polymerases and histones.

Table 2. Coverage of known TFs by ChIP-seq, ChIP-exo, and ChIP-nexus experiments in the GTRD database

Specie	CIS-BP (TF)	GTRD (TF & cofactors)	TF - intersecton CIS-BP - GTRD	%
Homo sapiens	1639	1535	1032	63
Mus musculus	1513	856	473	31
Rattus norvegicus	1362	47	19	1.4
Danio rerio	2350	27	13	0.5
Caenorhabditis elegans	766	338	218	28.5
Drosophila melanogaster	719	583	360	50
Saccharomyces cerevisiae	239	198	-	21
Schizosaccharomyces pombe	115	66	-	8.6
Arabidopsis thaliana	1749	135	88	5

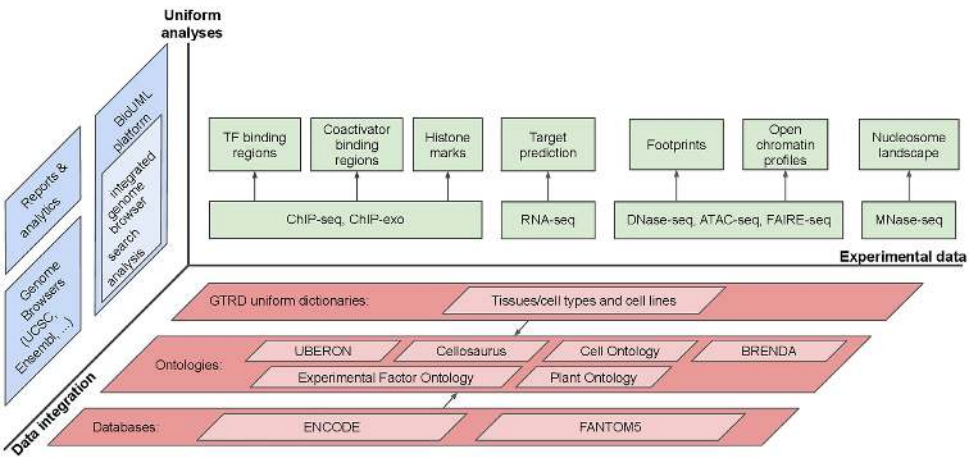

Figure 4. GTRD – integrated view of transcription regulation.

previous publication (35 719 reported in the current release versus 17 485 in the previous one). Importantly, the GTRD is continuously updated, and all ChIP-seq experiments from GEO and SRA for nine taxa within the GTRD scope of interest are usually included within six months after being deposited in public access.

There is a practically important question: What is the total fraction of known TFs covered by ChIP-seq, ChIP-exo, or ChIP-nexus experiments? For this purpose, we linked TFs in the GTRD with CIS-BP, which contains the most comprehensive TF lists across species (Table 2). Notably, about two-thirds of human TFs are covered by at least one ChIP-Seq experiment, so in a few years, we can expect that genomic binding-site data will be available for all human TFs experimentally studied in at least one cell type. This holds also for *D. melanogaster* and *S. cerevisiae*. A special

section in the GTRD is under construction to highlight TFs without published ChIP-seq, ChIP-exo, or ChIP-nexus experiments, so scientists can share their research status for a given TF – whether it is under investigation now or planned for the near future, and when results will be publicly available.

The current version of the GTRD is not limited to TF ChIP-seq data, and the contribution of other data is increasing (see summary statistics in Supplements 7, 8). Figure 4 illustrates our vision of gene transcription regulation data and its integrative analysis and representation within the GTRD framework. We also plan to integrate data on methylation of DNA into the GTRD. The MethMotif database (42) demonstrates how ChIP-seq data can be integrated with cell type-specific CpG methylation information.

Integration of the GTRD with other collections of experimental data on transcription regulation (FANTOM5, ENCODE, and GTEx) can act as a starting point for studying specific questions of mechanisms of transcription regulation of specific genes. For example, in our study of connections between TSS activities and the TFBSs (Sharipov *et al*., 2020, accepted in *PLoS One*), we developed an algorithm that predicts gene expression from TFBSs using FANTOM5 gene expression data and TF binding data from the GTRD. The algorithm utilizes precise TFBS locations and their arrangements, correlating them with TSS activity identified in the FANTOM5 project, thus yielding TFs contributing most to the control of gene expression and the location of their binding sites as referred to TSS.

Single-nucleotide variants (SNV) in gene regulatory regions can alter gene expression and contribute to phenotypes of individual cells and the whole organism, including disease susceptibility and progression. The Genotype-Tissue Expression (GTEx) project provides information about associations of SNV and gene expression for 54 non-diseased tissues (eQTL). Using clusters of cell types described above, we can associate eQTL data with corresponding TFBSs, taking into account the cellular context to study possible mechanisms of SNVs' alternations of gene activity.

Thus, with this latest release, the GTRD database becomes the largest integrated resource of data on transcription regulation in eukaryotes. It is noteworthy that GTRD contains not only uniformly annotated and processed NGS data, but also the results of the meta-analysis, presented in the form of meta-clusters, the sets of non-redundant and reproducible TFBSs for each TF, obtained by merging TFBSs identified in different experiments for the same TF. The meta-clusters can be considered as the first step towards complete cistrome for the corresponding organisms. The track of the master sites that integrates all relevant information from the GTRD database for a given TFBS further extends the concept of meta-clusters. This information can be used both for understanding how a particular site is involved in the transcription regulation of a gene as well as and for the development of new methods for identification of the most reliable TFBS from both statistical and biological evidence.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Russian Science Foundation [19-14-00295 to S.K., I.Y., M.K, R.Sh., Y.K., F.K.]; ASB identification was supported by the Russian Science Foundation [20-74-10075 to I.V.K.]; V.J.M. was supported by the Ministry of Science and Higher Education of the Russian Federation [075-15-2019-1658]. Funding for open access charge: Russian Science Foundation [19-14-00295 to F.K.].
Conflict of interest statement. None declared.

REFERENCES

1. Yáñez-Cuna,J.O., Kvon,E.Z. and Stark,A. (2013) Deciphering the transcriptional cis-regulatory code. *Trends Genet.*, **29**, 11–22.
2. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.
3. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
4. Forrest,A.R., Kawaji,H., Rehli,M., Baillie,J.K., De Hoon,M.J., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., Itoh,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
5. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K. *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.*, **48**, D882–D889.
6. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
7. Kodama,Y., Shumway,M. and Leinonen,R. (2012) International Nucleotide Sequence Database C. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
8. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**: D61–D67.
9. Yevshin,I., Sharipov,R., Kolmykov,S., Kondrakhin,Y. and Kolpakov,F. (2019) GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.*, **47**, D100–D105.
10. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
11. He,Q., Johnston,J. and Zeitlinger,J. (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotech.*, **33**, 395–401.
12. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
13. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
14. Gaulton,K.J., Nammo,T., Pasquali,L., Simon,J.M., Giresi,P.G., Fogarty,M.P., Panhuis,T.M., Mieczkowski,P., Secchi,A., Bosco,D. *et al.* (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.
15. Kuan,P.F., Huebert,D., Gasch,A. and Keles,S. (2009) A non-homogeneous hidden-state model on first order differences for automatic detection of nucleosome positions. *Stat. Appl. Genet. Mol. Biol.*, **8**, 29.
16. Feng,C., Song,C., Liu,Y., Qian,F., Gao,Y., Ning,Z., Wang,Q., Jiang,Y., Li,Y., Li,M. *et al.* (2020) KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Res.*, **48**, D93–D100.
17. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
18. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G.,

Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
19. Wingender,E., Schoeps,T., Haubrock,M., Krull,M. and Dönitz,J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.*, **46**, D343–D347.
20. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
21. 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.
22. Kulyashov,M.A., Kolmykov,S.K., Evshin,I.S. and Kolpakov,F.A. (2020) Description, characteristic and algorithm for creation of a dictionary of cell types and tissues in the GTRD database. In: *CEUR Workshop Proceedings*, Vol. **2569**, pp. 13–18.
23. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
24. Diehl,A.D., Meehan,T.F., Bradford,Y.M., Brush,M.H., Dahdul,W.M., Dougall,D.S., He,Y., Osumi-Sutherland,D., Ruttenberg,A., Sarntivijai,S. *et al.* (2016) The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.*, **7**, 44.
25. Jeske,L., Placzek,S., Schomburg,I., Chang,A. and Schomburg,D. (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.*, **47**, D542–D549.
26. Cooper,L., Walls,R.L., Elser,J., Gandolfo,M.A., Stevenson,D.W., Smith,B., Preece,J., Athreya,B., Mungall,C.J., Rensing,S. *et al.* (2013) The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.*, **54**, e1.
27. Bairoch,A. (2018). The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech.*, **29**, 25.
28. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
29. Lizio,M., Abugessaisa,I., Noguchi,S., Kondo,A., Hasegawa,A., Hon,C.C., De Hoon,M., Severin,J., Oki,S., Hayashizaki,Y. *et al.* (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res.*, **47**, D752–D758.
30. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
31. Kolmykov,S.K., Kondrakhin,Y.V., Yevshin,I.S., Sharipov,R.N., Ryabova,A.S. and Kolpakov,F.A. (2019) Population size estimation for quality control of ChIP-Seq datasets. *PLoS One*, **14**, e0221760.
32. Lin,S. (2010) Rank aggregation methods. *Wiley Interdiscip. Rev. Comput. Stat.*, **2**, 555–570.
33. Abramov,S., Boytsov,A., Bykova,D., Penzar,D., Yevshin,I., Kolmykov,S., Fridman,M., Favorov,A., Vorontsov,I., Baulin,E. *et al.* (2020) Landscape of allele-specific transcription factor binding in the human genome. bioRxiv doi: https://doi.org/10.1101/2020.10.07.327643, 13 October 2020, preprint: not peer reviewed.
34. Liberzon,A., Birger,C., Thorvaldsdóttir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The molecular signatures database hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
35. Kolpakov,F., Akberdin,I., Kashapov,T., Kiselev,L., Kolmykov,S., Kondrakhin,Y., Kutumova,E., Mandrik,N., Pintus,S., Ryabova,A. *et al.* (2019) BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data. *Nucleic Acids Res.*, **47**, W225–W233.
36. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
37. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
38. 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.
39. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
40. Kiesel,A., Roth,C., Ge,W., Wess,M., Meier,M. and Söding,J. (2018) The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.*, **46**, W215–W220.
41. Vorontsov,I.E., Fedorova,A.D., Yevshin,I.S., Sharipov,R.N., Kolpakov,F.A., Makeev,V.J. and Kulakovskiy,I.V. (2018) Genome-wide map of human and mouse transcription factor binding sites aggregated from ChIP-Seq data. *BMC Res. Notes*, **11**, 756.
42. Lin,Q.X.X., Sian,S., An,O., Thieffry,D., Jha,S. and Benoukraf,T. (2019) MethMotif: an integrative cell specific database of transcription factor binding motifs coupled with DNA methylation profiles. *Nucleic Acids Res.*, **47**, D145–D154.