Description, Characteristic And Algorithm For Creation Of A Dictionary Of Cell Types And Tissues In The Gtrd Database[1]

Michael A Kulyashov[1,2], Sergei K. Kolmykov[1,2,3], Ivan S. Evshin[1,2], Fedor A. Kolpakov [1,2]
[1]Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[2] Biosoft.ru, Novosibirsk, Russia, m.kulyashov@developmentontheedge.com
[3] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Abstract. The GTRD database (http://gtrd.biouml.org) contains information on transcription factor binding sites and open chromatin. The cell types and tissues presented in the GTRD: 1) were arranged in a single dictionary (3954 entries); 2) were divided into 90 unique clusters; 3) 3225 records were compared with main databases of cell types; 4) were used to match experiments with the databases of transcription regulation: 588 for FANTOM 5, 5962 for ENCODE and 21720 for GTEx.

Keywords: GTRD, database structure, dictionary of cell types and tissues, cell type onthology

1 Introduction

The GTRD database (Gene Transcription Factors Database, http://gtrd.biouml.org) contains information on transcription factor binding sites and open chromatin sites that have been experimentally identified in various cell types and tissues using high-performance ChIP-seq and DNase-seq methods respectively [1]. GTRD currently contains the largest number of uniformly processed experiments of the corresponding types in the world (http://wiki.biouml.org/index.php/GTRD_comparison).

A separate ChIP-seq experiment allows all binding sites for a single transcription factor to be identified for a single cell type (with or without treatment) or a tissue fragment excreted from the body. A separate DNase-seq experiment will allow to identify all sections of open chromatin also only for a single cell type (with or without treatment) or a tissue fragment. Therefore, to integrate data on a given cell type or tissue, it is necessary to maintain a dictionary of cell types and tissues.

To understand the regulation of transcription the important step is integration of GTRD with the main specialized transcription regulation databases, the main of which are:

- ENCODE [2] is a project whose purpose is to analyze the functional elements of the genome;
- FANTOM5 [3] is a project aimed at studying the regulation of transcription in various cell types and tissues of a human and mouse;
- GTEx [4] is a project aimed at studying tissue-specific expression and mechanisms of gene regulation of human.

At the stage of developing these databases, the issue of systematization of cell types and lines was also raised. To solve this problem, specialized databases are used to compare cell types and tissues (Table 1).

Table 1. Description of the databases used to identify cell types and tissues.

Database name	Database description	Prefix	Databases which uses this ontology database
EFO (Experimental factor ontology)[5]	An ontology that describes various cellular states, experimental conditions, and developmental stages.	EFO	ENCODE, FANTOM 5, GTEx

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR-WS.org/Vol-2569/short3.pdf

UBERON (UBER anatomy atlas)[6]	Ontology of various anatomical structures of animals, which is composed taking into account all modern ideas about the anatomical structure, functioning and stages of development.	UBERON	ENCODE, FANTOM 5, GTEx
Cell ontology[7]	This ontology describes various types of cells.	CL	ENCODE, FANTOM 5, GTEx
BRENDA tissue ontology[8]	In this anthology, a huge amount of various data is collected, cell types, tissues, cell cultures for various taxonomic groups, such as animals, plants, fungi, protozoa.	BTO	ENCODE, FANTOM 5, GTEx
Plant ontology[9]	Ontology with a description of various anatomical and morphological structures of plants, as well as data on the growth and development of plants.	PO	
Cellosaurus [10]	A resource that contains a description of a large number of cell lines, as well as data on the relationship of cell lines that are used in many biomedical studies.	CVCL	ENCODE

2 Materials and Methods

To create a dictionary of cell types and tissues and then compare experiments between GTRD and transcription regulation bases, a plan of work was developed as shown in Figure 1, where each stage of development is numbered.

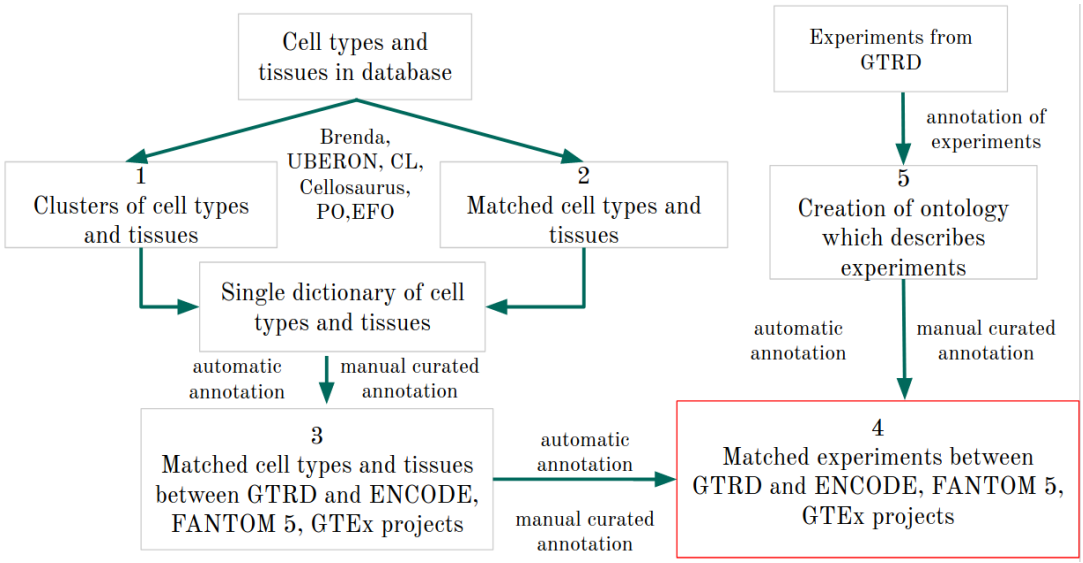

Figure 1. Pipeline for creating a dictionary of cell types and tissues and subsequent comparisons with databases for transcription regulation.

2.1 The division of cell types and tissues into clusters.

For this work, the following principle was used: the original tissue and developmental stage on which cell type or tissue was extracted. To determine the type of cells and tissues that failed to determine the organ, anatomical or morphological system in which they are located, was made a separate group “Others”. Data was recorded according to the principle of “key” - “value”, where the key is the name of the cluster and the value is name of cell type or tissue from GTRD.

2.2 Comparison of cell types and tissues with specialized databases.

For creation of a dictionary of cell types and tissues the following databases were selected: EFO, Brand Ontology, UBERON, Cell Ontology, Plant Ontology (Table 1). The comparison of the cell type or tissue was carried out according to the following principle: the assignment of the most accurate data that would allow all the information to be obtained, but at the same time they would not give characteristics that contradict this cell type and tissue. This work is done by using semi-automatic and manual annotation methods:

1) Semi-automatic annotation - for this was developed the program in Python 3.6, which for each cell type and tissue located in the GTRD performed a search on the databases described in table 1, using the API ontology search service [11]. Results that satisfy the above points have been added to the GTRD.

2) manual annotation was carried out for cell types and tissues,for which didn’t find any matches.

The data obtained are recorded on the principle of “key” - “value”, where the key corresponds to the name of the cell type or tissue from GTRD, and value is a matched record from specialized databases.

2.3 Comparison of dictionaries of cell types and tissues between GTRD and the transcription regulation database.

For this work was developed a program in Python 3.6, which checked for matching entries from the GTRD dictionary with available transcription regulation databases (ENCODE, FANTOM5, GTEX). After that, the list was checked for errors.

2.4 Comparison of experiments between GTRD and the transcriptional regulation database.

Comparison of experiments between GTRD and the transcriptional regulation database. To do this, we used a program written in Python 3.6, which compares all the experimental data in the database with each cell type and tissue. Then, on the basis of the comparison obtained earlier, between the dictionaries of cell types and tissues of GRTD and transcription regulation databases, experiments were compared. After this, list has been checked manually to eliminate errors and also for each experiments for which the results of automatic annotation didn’t find any matches or which are comparable only for cluster to which cell type or tissue from experiment belong.

2.5 Creating additional keys to describe the experiments.

When describing the experimental data in the GTRD, in addition to the cell type or tissues, it is necessary to indicate conditions, some of which are given in Table 2. These conditions are formed in the form of a key-value set and are attached to the results of the GTRD experiments.

Key	Meaning
Treatment	This key describes whether the treatment was carried out with various chemical or biological compounds and in what quantities
Genotype	The key that describes the presence of genetic modifications other than the normal (“wild type”) genotype
Sex	This key describes the biological sex of the object of research from which the sample was taken.
Age	The key describes the age at which the sample from object of research was taken
Developmental stage	The key that describes the stage of the life cycle of an object of research at the time of taking a sample from it

Strain	Key describing the strain to which the sample belongs.
Source	A key that describes various characteristics of the sample, such as from what area the sample was taken from, the sample freshly isolated or frozen, e.t.c.

3 Results

The constructed dictionary of cell types and tissues of GTRD currently contains 3954 entries. All cell types and tissues in the GTRD dictionary were divided into 90 clusters (Figure 2).

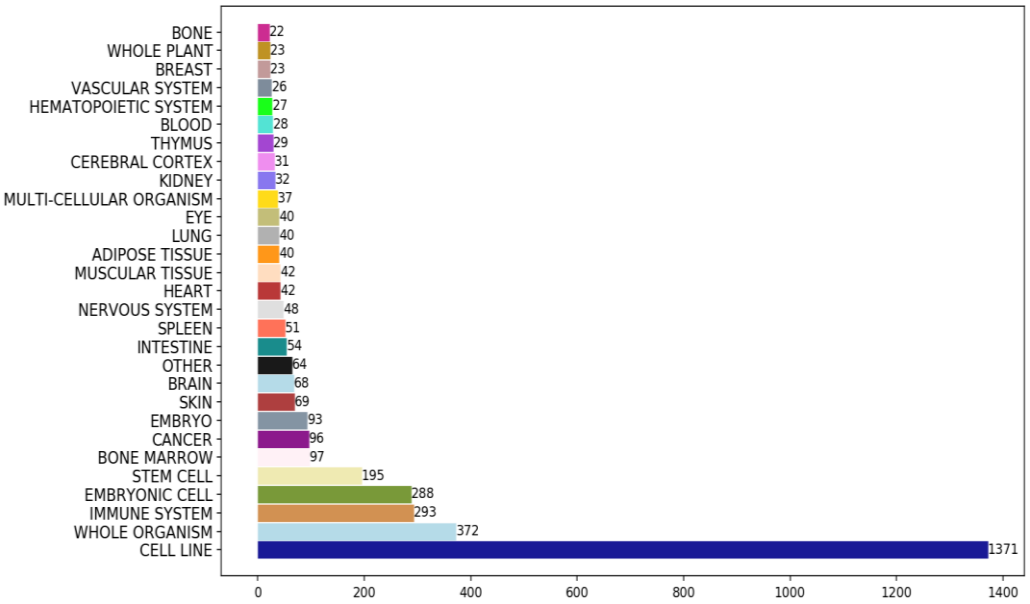

Figure 2. The number of cell types and tissues included in large clusters, with more than 20 elements.

Allocation to clusters make it possible to systematize cell types and tissues in the GTRD database according to anatomical, morphological, and other characteristics. The presence of large clusters, shown in Figure 1, can significantly simplify the verification of cell types and tissues, speeds up the work with them, and further allows us to analyze experimental data for experiments from anatomically close areas. For a small number of cell types, there was not enough information to determine their belonging to any cluster (Other category, Fig. 2).

For 3225 entries, correspondence was established with the main databases of cell types and tissues presented in Table 1 and this represents more than 80% of all entries in the dictionary. This correspondence was found for most cell types and tissues (Table 3).

Table 3. Number of matched cell types and tissues with main databases.

Species	CVCL	CL	UBERON	BTO	EFO	PO	Unmatched
Arabidopsis thaliana	0	0	0	57	0	7	2
Caenorhabditis elegans	0	1	47	0	0	0	0

Danio rerio	0	1	6	1	0	0	39
Drosophila melanogaster	13	4	86	6	0	0	45
Homo sapiens	685	511	189	103	11	0	306
Mus musculus	124	653	281	83	14	0	332
Rattus norvegicus	8	15	18	3	0	0	5
Saccharomyces cerevisiae	0	0	184	7	0	0	0
Schizosaccharomyces pombe	0	0	124	0	0	0	0
Total	830	1185	935	259	25	7	729

No matches were found for 729 records, i.e. these cell types or tissues are found only in the experiments described in the GTRD.

As result of the comparison of GTRD cell types and tissues with specialized transcription regulation databases we can show number of matches between GTRD and transcription regulation databases, which presented in Table 4.

Table 4. The result of comparing GTRD with transcription regulation databases.

Database	Number of cell types and tissues		Количество экспериментов	
	in database	matched with GTRD	in database	matched with GTRD
FANTOM 5	856[2]	293	1458[2]	588
ENCODE v92	450[1]	432	6753[1]	5962
GTEx v8	54	36	25713	21720

Thanks to the previous comparison it became possible to integrate transcriptional regulation experiments to GTRD from ENCODE, FANTOM5 for each experiments, for which was founded match of cell type or tissue.This comparison made it easy to integrate data for various analyzes, which was demonstrated by us on the data from FANTOM 5 in the work presented at the MCCMB-2019 conference [12].

The next step, after compiling the dictionary and comparing thanks to it the cellular types and tissues, was the development of a system for describing experiments. The introduced new keys, as well as the reworked old ones, allow to analyze experiments on various grounds, such as sex, age, and others, which opens up new possibilities for data analysis.

3 Conclusion

The dictionary of cell types and tissues that we have developed currently has 82% of the matched cell types and tissues of all records presented in the GTR, which sets us the task of minimizing the number of unmatched cell types and tissues.

[1] for ChIP-seq and DNase-seq experiments

[2] excluding time course experiments

Also an important step will be to introduce for cell types and tissues, a hierarchical system in the future, which will increase accuracy when comparing experiments from various databases. In addition, the number of experiments in GTRD increases every year, and them will allow to match more cell types and tissues and as a consequence to integrate more experiments from transcriptional regulation databases.

Acknowledgements. This work was supported by the Russian Science Foundation (grant No. 19-14-00295)

References

[1] *Yevshin Y., Sharipov., Kolmykov S., Kondrakhin Y.,Kolpakov F*. GTRD: a database on gene transcription regulation—2019 update // Nucleic Acids Research. 2019, Volume 47, Issue D1, Pages D100–D105.

[2] *Landt S.G., Marinov G.K., Kundaje A., et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia// Genome Res. 2012, 22(9):1813–1831.

[3] *GTEx Consortium*. The Genotype-Tissue Expression (GTEx) project// Nat Genet. 2013, 45(6):580–585.

[4] *Lizio M., et al*. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. 2015, 16: 22.

[5] *Malone J., Holloway E., Adamusiak T., et al*. Modeling Sample Variables with an Experimental Factor Ontology// Bioinformatics. 2010, 26(8):1112-1118.

[6] *Mungall C.J., Torniai C., Gkoutos G.V., et al*. Uberon, an integrative multi-species anatomy ontology// Genome Biol. 2012, 13, R5.

[7] *Bard J., Rhee S.Y., Ashburner M.* An ontology for cell types//Genome Biol. 2005;6(2):R21.

[8] *Gremse M, Chang A, Schomburg I, et al*. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources// Nucleic Acids Res. 2011, D507–D513.

[9] *Cooper L., Walls R., Elser J., et al.* The Plant Ontology as a Tool for Comparative Plant Anatomy and Genomic Analyses// Plant and Cell Physiology. 2013, Volume 54, Issue 2, Page e1.

[10] *Bairoch A.* The Cellosaurus, a Cell-Line Knowledge Resource// J Biomol Tech. 2018, 29(2):page 25–38.

[11] *Jupp S., et al*. A new Ontology Lookup Service at EMBL-EBI// Proceedings of SWAT4LS International Conference 2015, Cambridge.

[12] *Kondrakhin Y., Kolmykov S., Yevshin I., Sharipov R., Ryabova A., Kulyashov., Kolpakov F*. Combining GTRD ChIP-Seq datasets with FANTOM5's transcription start sites for prediction of gene expression levels // Proceedings of MCCMB International Conference 2019, Moscow: Lomonosov Moscow State University