# Advanced data curation in GTRD database: hierarchical dictionaries of cell types and experimental factors

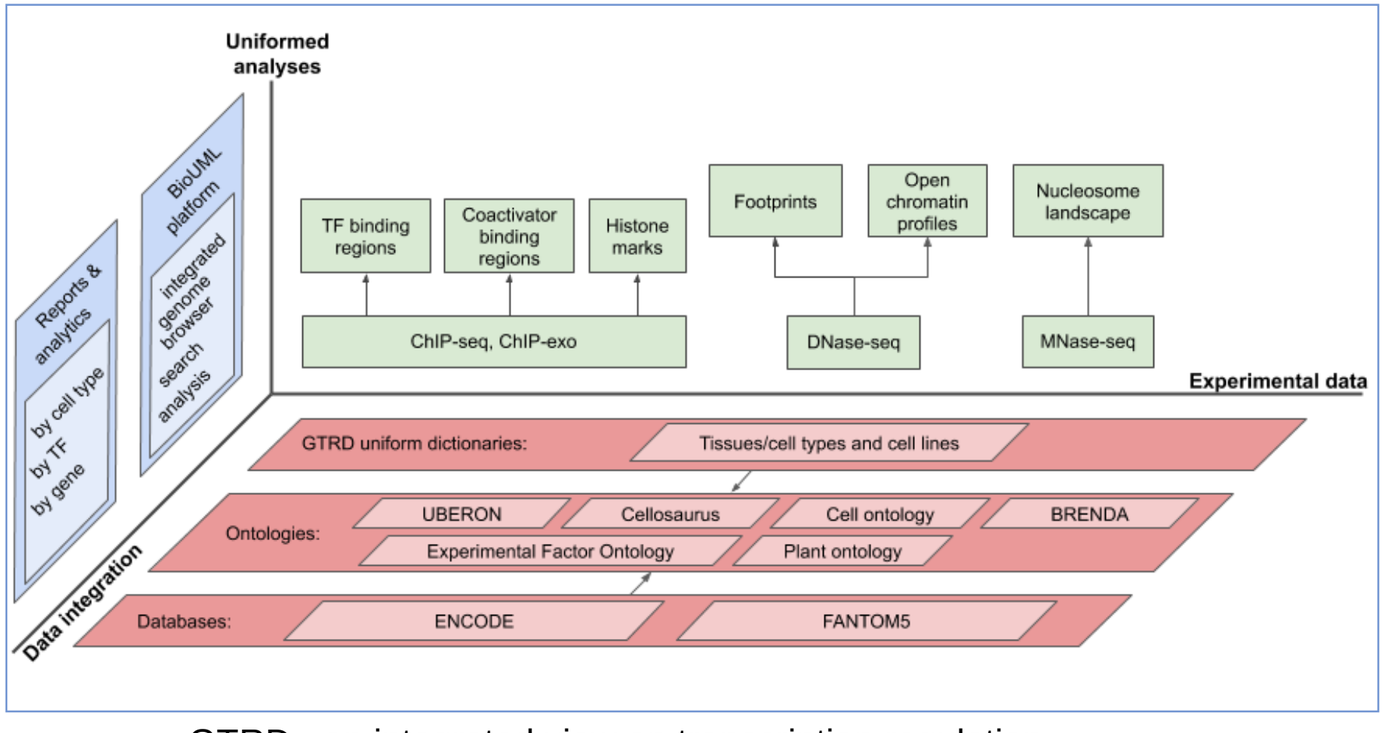

**Mikhail A. Kulyashov**
**BIOSOFT.RU,LLC**
**Novosibirsk State University**
**FRC Institute of Cytology and Genetics SB RAS**
**Institute of Computational Technologies SB RA**
**Novosibirsk, Russia m.kulyashov@mail.ru**

**Semyon K. Kolmykov**
**BIOSOFT.RU,LLC**
**Institute of Computational Technologies SB RAS**
**FRC Institute of Cytology and Genetics SB RAS**
**Novosibirsk, Russia kolmykovsk@gmail.com**

**Ivan S. Yevshin**
**BIOSOFT.RU,LLC**
**Institute of Computational Technologies SB RAS**
**Novosibirsk, Russia**
**ivan@developmentontheedge.com**

**Fedor A. Kolpakov**
**Institute of Computational Technologies SB RAS**
**BIOSOFT.RU,LLC**
**Novosibirsk, Russia fkolpakov@gmail.com**


# Introduction




GTRD - an integrated view on transcription regulation

| Experiment type | Total experiments in GTRD |
|---|---|
| ChIP-seq | 30670 |
| ChIP-exo, ChIP-nexus | 707 |
| DNase-seq | 1585 |
| MNase-seq | 953 |
| ChIP-seq HM | 2687 |

Number of experiments in GTRD

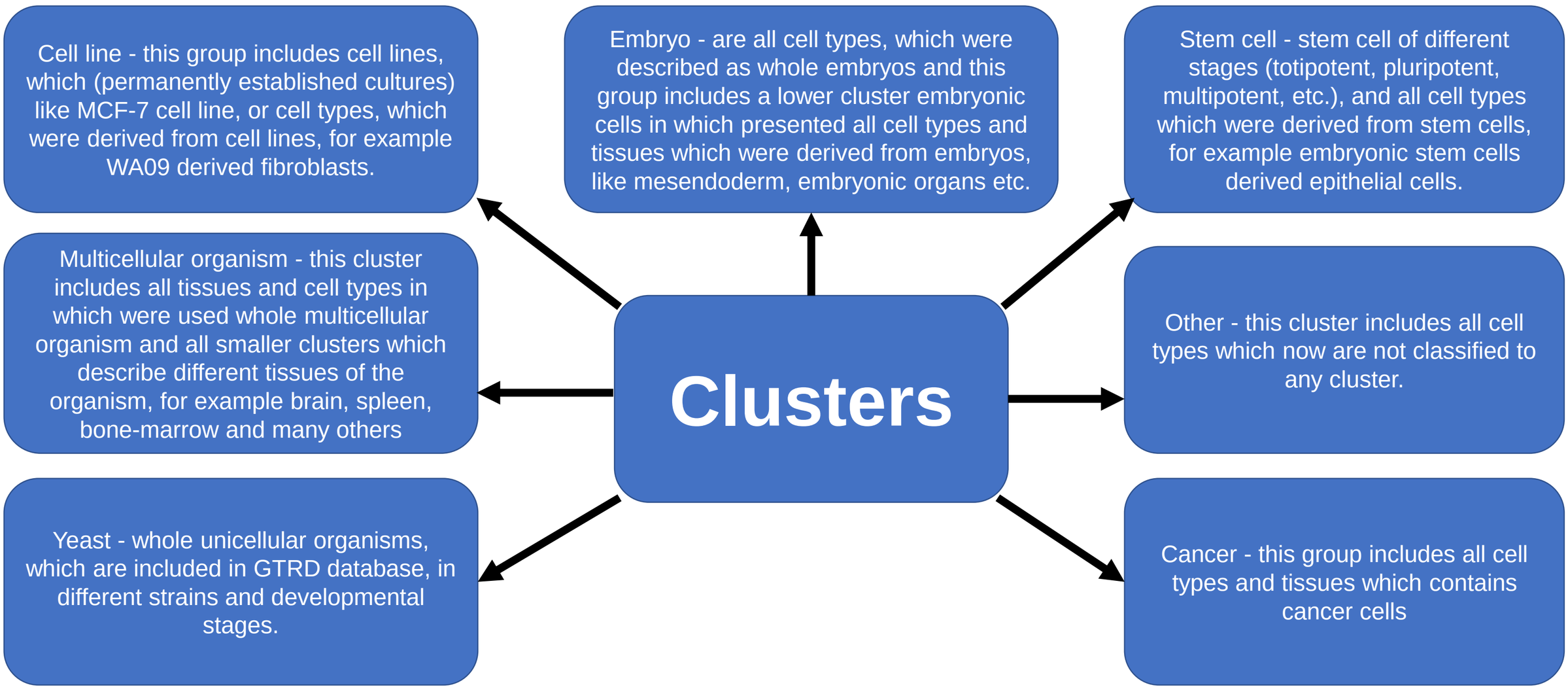
# Materials and methods

## Creation of a dictionary of cell types and tissues

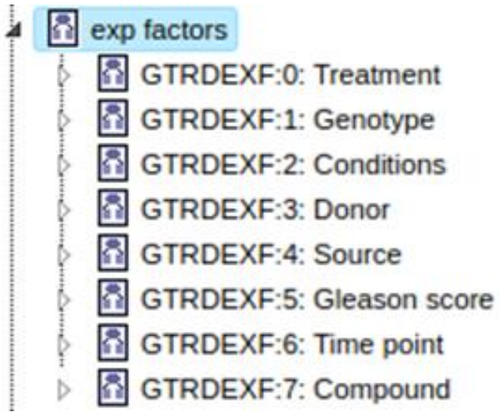
All cell lines, tissues and cell types which were used in experiments, described in GTRD, contain 2 fields that were used for building the dictionary:

- ***Cell type*** - cells on which experiment was made
- ***Source*** - tissues or cell types from which cells were isolated in experiments or system, which includes this cell, if isolation region was not indicated

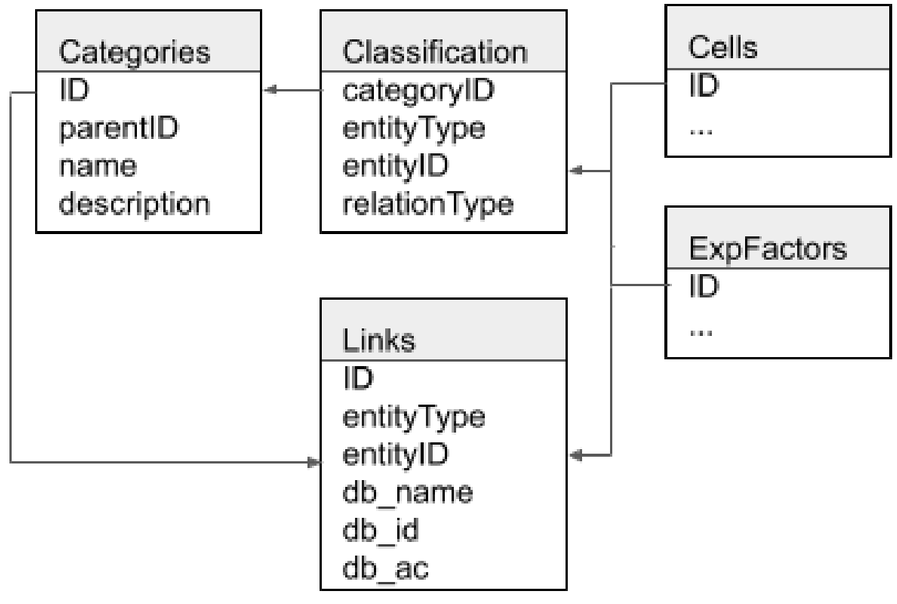
To describe cell type we use specialized databases. This approach can be very important for researches of transcriptional regulation. First step in process of creation a hierarchical dictionary was to separate sources to main ***clusters*** - a global groups which unify different sources, such organs, embryonic parts, cell lines and some other in union groups



# Creation of a dictionary of experimental factors



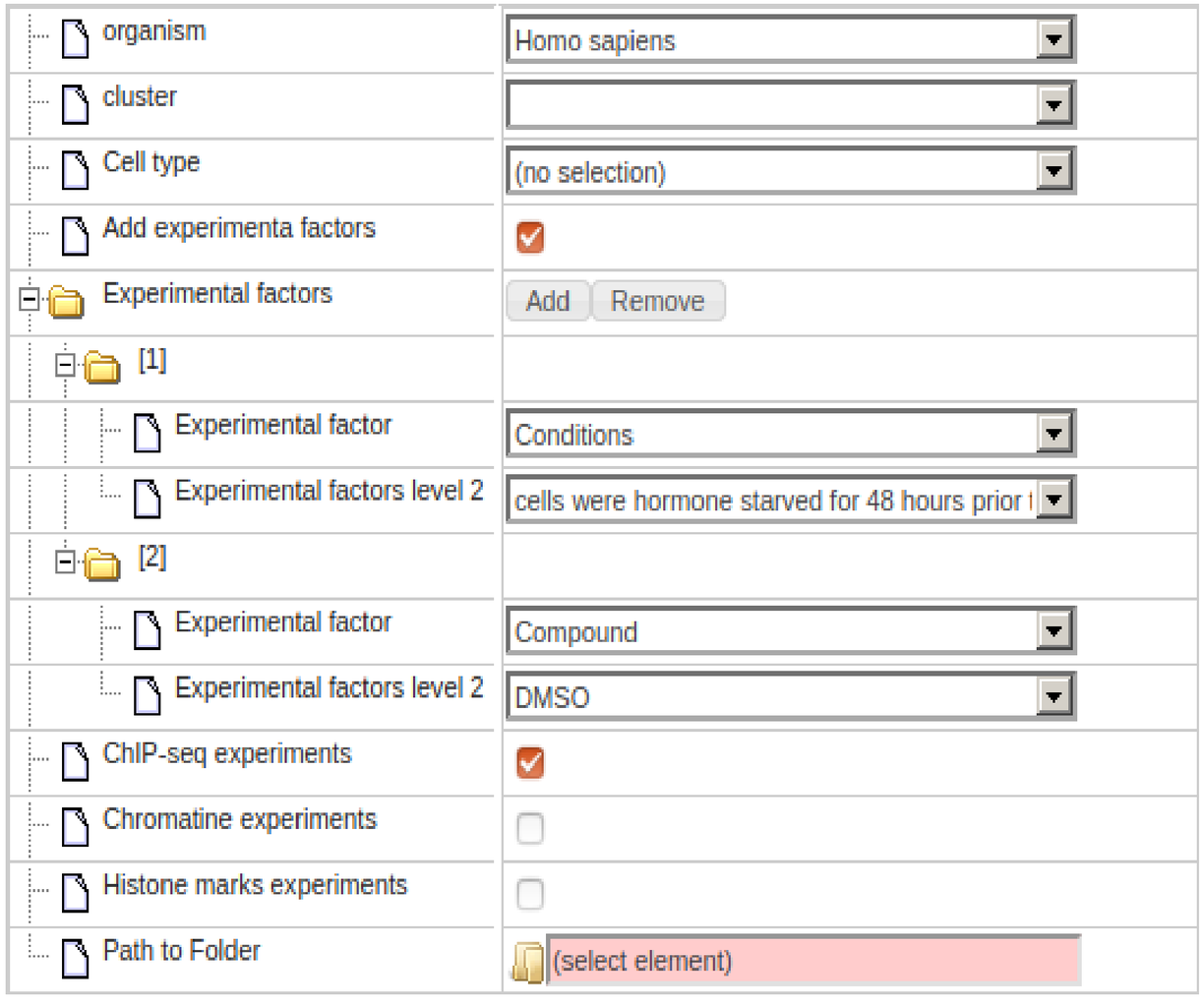

# GTRD database scheme for hierarchical classification of data




- *Categories* - describes classification tree where the clusters described above are the roots (parentID is null).
- *Classification* - links flat dictionary of cell types and experimental factors with categories tree.

# Results

- 99% of all cell types and tissues which are presented in GTRD are in identified cluster. ‘Cell type’ was compared with specialized databases for more than 80% of cell types and tissues;
- Were created hierarchical dictionaries for all cell types and tissues for annotated in GTRD database experiments;
- Created a hierarchical dictionary for experimental factors by analogy with the dictionary of cell types and tissues;
- Developed analysis for filtering experiments for cell types, tissues and experimental factors using BioUML platform, which was integrated into GTRD database;




Interface of experiment filtering analysis

# Analysis run result

First Previous Page 1 of 4 Next Last Showing 1 to 50 of 164 entries Show 50 entries

| ID | Cell type | Experimental factors | Antibody | Target class | Target name | Peaks | Control | Specie | External References |
|---|---|---|---|---|---|---|---|---|---|
| EXP000408 | VCaP (prostate carcinoma) | Vehicle | AR | 2.1.1.1.4 | AR | PEAKS033252 | EXP000411 | Homo sapiens | GSM696841, GSE28126, 21602788 |
| EXP000409 | VCaP (prostate carcinoma) | AR stimulated | AR | 2.1.1.1.4 | AR | PEAKS033253 | EXP000411 | Homo sapiens | GSM696842, GSE28126, 21602788 |
| EXP000411 | VCaP (prostate carcinoma) | AR stimulated | input | | | | | Homo sapiens | GSM696846, GSE28126, 21602788 |
| EXP000721 | VCaP (prostate carcinoma) | regular medium | FoxA1 | 3.3.1.1.1 | FOXA1 | PEAKS033384 | | Homo sapiens | GSM353630, GSE14092, 20478527 |
| EXP000723 | VCaP (prostate carcinoma) | NT | AR | 2.1.1.1.4 | AR | PEAKS033385 | | Homo sapiens | GSM353636, GSE14092, 20478527 |
| EXP000724 | VCaP (prostate carcinoma) | NT | ERG | 3.5.2.1.6 | ERG | PEAKS033382 | | Homo sapiens | GSM353637, GSE14092, 20478527 |
| EXP000725 | VCaP (prostate carcinoma) | siERG | AR | 2.1.1.1.4 | AR | PEAKS033164 | | Homo sapiens | GSM353638, GSE14092, 20478527 |
| EXP000726 | VCaP (prostate carcinoma) | siERG | ERG | 3.5.2.1.6 | ERG | PEAKS033165 | | Homo sapiens | GSM353639, GSE14092, 20478527 |
| EXP000732 | VCaP (prostate carcinoma) | ethanol | AR | 2.1.1.1.4 | AR | PEAKS033109 | | Homo sapiens | GSM353645, GSE14092, 20478527 |
| EXP000733 | VCaP (prostate carcinoma) | R1881 | AR | 2.1.1.1.4 | AR | PEAKS033110 | | Homo sapiens | GSM353646, GSE14092, 20478527 |
| EXP000734 | VCaP (prostate carcinoma) | regular medium | ERG | 3.5.2.1.6 | ERG | PEAKS033107 | | Homo sapiens | GSM353647, GSE14092, 20478527 |
| EXP000740 | VCaP (prostate carcinoma) | regular medium | ERG + AR | 2.1.1.1.4 | AR | PEAKS036855 | | Homo sapiens | GSM356767, GSE14092, 20478527 |
| EXP030494 | VCaP (prostate carcinoma) | 100 nM DHT | rIgG (sc-2027, SantaCruz) | | | | | Homo sapiens | GSM980659, GSE39879, 23269278 |
| EXP030495 | VCaP (prostate carcinoma) | 100 nM DEX | mIgG (sc-2025, Santa Cruz) | | | | | Homo sapiens | GSM980661, GSE39879, 23269278 |
| EXP030499 | VCaP (prostate carcinoma) | 100 nM DEX | GR (BuGr2, Millipore and Mab-010-050, Diagenode) | 2.1.1.1.1 | NR3C1 | PEAKS034496 | EXP030494 | Homo sapiens | GSM980660, GSE39879, 23269278 |
| EXP030507 | VCaP (prostate carcinoma) | 100 nM DHT | AR (Kang et al. 2004; Sahu et al. 2011) | 2.1.1.1.4 | AR | PEAKS034504 | EXP030494 | Homo sapiens | GSE39879, 23269278, GSM980658, GSM980657 |
| EXP030997 | VCaP (prostate carcinoma) | none | none | | | | | Homo sapiens | GSM717391, GSE28950, 22531786 |
| EXP030998 | VCaP (prostate carcinoma) | none | sc-815x | 2.1.1.1.4 | AR | PEAKS034672 | EXP030997 | Homo sapiens | GSM717392, GSE28950, 22531786 |

Result table for search all ChIP-seq experiments with all available treatments from GTRD for VCaP prostate cancer cell line