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

Introduction



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

 type
 Total experiments in GTRD

 in GTRD
 30670

 -nexus
 707

 1585
 953

 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

Cell line - this group includes cell lines, which (permanently established cultures) like MCF-7 cell line, or cell types, which were derived from cell lines, for example WA09 derived fibroblasts.

Multicellular organism - this cluster includes all tissues and cell types in which were used whole multicellular organism and all smaller clusters which describe different tissues of the organism, for example brain, spleen, bone-marrow and many others

Yeast - whole unicellular organisms, which are included in GTRD database, in different strains and developmental stages.

Embryo - are all cell types, which were described as whole embryos and this group includes a lower cluster embryonic cells in which presented all cell types and tissues which were derived from embryos, like mesendoderm, embryonic organs etc.

Clusters

Stem cell - stem cell of different stages (totipotent, pluripotent, multipotent, etc.), and all cell types which were derived from stem cells, for example embryonic stem cells derived epithelial cells.

Other - this cluster includes all cell types which now are not classified to any cluster.

Cancer - this group includes all cell types and tissues which contains cancer cells



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.

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:

🗅 organism	Homo sapiens						
Cluster							
Cell type	(no selection)						
Add experimenta factors							
Experimental factors	Add Remove						
E· 🛅 [1]							
Experimental factor	Conditions						
Experimental factors level 2	cells were hormone starved for 48 hours prior 1						
Ē· 🛅 [2]							
Experimental factor	Compound						
Experimental factors level 2	DMSO						
ChIP-seq experiments							
Chromatine experiments							
Histone marks experiments							
End Path to Folder	(select element)						
Run Interface of experiment filtering analysis							

Results

99% of all cell types and tissues which are presented in GTRD are in identified cluster. 'Cell type' was compared with specialized databases

Analysis run result

First Previous Page 1 of 4 Next Last Showing 1 to 50 of 164 entries										
ID 🔺	Cell type	Experimental factors	Antibody	Target class	Target aname	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	rlgG (sc-2027, SantaCruz)					Homo sapiens	GSM980659, GSE39879, 23269278	
EXP030495	VCaP (prostate carcinoma)	100 nM DEX	mlgG (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