

# GEOsearch: Extendable Search Engine for Gene Expression Omnibus

Zhicheng Ji

Hongkai Ji

Johns Hopkins University,  
Baltimore, Maryland, USA  
zji4@jhu.edu

Johns Hopkins University,  
Baltimore, Maryland, USA  
hji@jhsp.h.edu

April 24, 2017

## Contents

<b>1</b>	<b>Introductions</b>	<b>1</b>
<b>2</b>	<b>Find Term Alias</b>	<b>2</b>
<b>3</b>	<b>Perform Searching</b>	<b>3</b>
<b>4</b>	<b>Frequencies of Common Biology Keywords</b>	<b>5</b>
<b>5</b>	<b>Details of GSM Samples Given GSE Accession ID.</b>	<b>6</b>
<b>6</b>	<b>SEPA GUI</b>	<b>9</b>
<b>7</b>	<b>Reference</b>	<b>9</b>

## 1 Introductions

As the largest and most used public repositories for genomics data, the NCBI Gene Expression Omnibus (GEO [1], <http://www.ncbi.nlm.nih.gov/geo/>) is an indispensable tool for researchers to search and explore various kind of genomics data. However, the default search function of GEO is not comprehensive and powerful enough. For example, if the search term contains gene name such as Oct4, GEO will ignore samples or experiments related to its alias Pou5f1, which makes the search results incomprehensive. In addition, GEO does not provide second-round search functions for users to further narrow down the experiments or samples of interest after an initial search. For example, users may want to focus on experiments related to specific biological contexts such as

different cell types, tissues and diseases. It is tedious and inefficient for users to perform such second-round search with default search functions provided by GEO. GEOmetadb [2] was previously proposed as an alternative search engine for GEO to facilitate the query of GEO metadata. However, GEOmetadb does not provide the function of searching gene alias or performing second-round search. In addition, GEOmetadb depends on a static GEO metadata database which requires frequent updating and prevents it from obtaining the most up-to-dated results. To address these problems, we developed GEOsearch which is an expandable search engine available online. GEOsearch can provide more comprehensive search results by automatically searching all alias of the gene names contained in the search term. The search results are then integrated and displayed in a compact and editable table. After an initial search, GEOsearch summarize the biological contexts of the search results and allows users to perform second-round search to further narrow down the search results. Unlike GEOmetadb, GEOsearch takes advantage of the programmatic search portal provided by GEO and does not depend on any external database. As a result GEOsearch does not require routine updating and maintaining. GEOsearch can be primarily accessed by directly visiting the online user interface <https://zhiji.shinyapps.io/GEOsearch>. In this way GEOsearch can be used as an alternative of NCBI GEO's default search engine. GEOsearch can also be used programmatically by running R commands.

## 2 Find Term Alias

The function `termalias` first picks gene names from the search term. It then searches alias for all the gene names contained in the search term. It next queries GEO and retain alias that appear frequently enough in GEO database. Finally it returns a combinatory results of retained alias.

```
##
##
## Warning: Package 'GEOsearch' is deprecated and will be removed
from
## Bioconductor version 3.6
## Loading required package: AnnotationDbi
## Loading required package: stats4
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##   clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##   clusterExport, clusterMap, parApply, parCapply, parLapply,
##   parLapplyLB, parRapply, parSapply, parSapplyLB
```

```

## The following objects are masked from 'package:stats':
##
##   IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##   Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##   as.data.frame, cbind, colMeans, colSums, colnames, do.call,
##   duplicated, eval, evalq, get, grep, grepl, intersect,
##   is.unsorted, lapply, lengths, mapply, match, mget, order,
##   paste, pmax, pmax.int, pmin, pmin.int, rank, rbind, rowMeans,
##   rowSums, rownames, sapply, setdiff, sort, table, tapply,
##   union, unique, unsplit, which, which.max, which.min
## Loading required package: Biobase
## Welcome to Bioconductor
##
##   Vignettes contain introductory material; view with
##   'browseVignettes()'. To cite Bioconductor, see
##   'citation("Biobase)", and for packages 'citation("pkgname)".
## Loading required package: IRanges
## Loading required package: S4Vectors
##
## Attaching package: 'S4Vectors'
## The following object is masked from 'package:base':
##
##   expand.grid

```

```

Oct4alias <- TermAlias("Oct4 RNA-seq")

## 'select()' returned 1:1 mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:1 mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns

Oct4alias

## [1] "oct3 rna-seq" "oct4 rna-seq" "oct-3 rna-seq" "oct-4 rna-seq"
## [5] "pou5f1 rna-seq"

```

In this example, `termalias` picks the gene name (Oct4) and finds all its alias (e.g. Pou5f1). It returns the combinatory results of all alias.

### 3 Perform Searching

The function `GEOsearchterm` searches the input terms one by one in NCBI GEO database and returns an integrated table of search results. The returned results

should contain exactly the same information as the results returned by directly searching in <http://www.ncbi.nlm.nih.gov/geo/>. The input of GEOsearchterm is typically the direct output of the function termalias.

```
Oct4searchres <- GEOSearchTerm(Oct4alias)
head(Oct4searchres)

##      Series      Organism
## 1 GSE79636 Homo sapiens
## 2 GSE85854 Homo sapiens
## 3 GSE86790 Mus musculus
## 4 GSE56568 Homo sapiens
## 5 GSE45338 Mus musculus
## 6 GSE70713 Mus musculus
##
## 1
## 2
## 3 Three-dimensional retinal organoids from mouse pluripotent stem cells mimic in vivo dev
## 4 Critical role of transient activation of human
## 5
## 6 Single-cell analysis reveals lineage segregation
##
## 1 Expression profiling
## 2 Expression profiling
## 3 Expression profiling
## 4 Non-coding RNA profiling
## 5 Genome binding/occupancy profiling by high throughput sequencing; Expression profiling
## 6 Expression profiling
##
##      Platform Sample.Number
## 1 GPL16791 317
## 2 GPL11154 4
## 3 GPL17021 16
## 4 GPL11154 15
## 5 4 related Platforms 13
## 6 GPL17021 GPL13112 236
##
## 1
## 2
## 3 Generation of three-dimensional (3D) organoids with optic cup like structures from plur
## 4 We recently sho
## 5
## 6 Here we re
##
##      Term
## 1 oct3 rna-seq;oct4 rna-seq
## 2 oct3 rna-seq
## 3 oct3 rna-seq
```

```
## 4          oct3 rna-seq
## 5          oct4 rna-seq
## 6          oct4 rna-seq
```

The example above lists not only search results for the term "Oct4 RNA-seq" but also results for the term "Pou5f1 RNA-seq". The combined results are more comprehensive than the results of searching either one of the term. In principal the search results returned by GEOsearch should always be as or more comprehensive than the search results returned by GEO default search functions.

## 4 Frequencies of Common Biology Keywords

The function `keywordfreq` calculates the frequencies of each common biology keyword appearing in the given search table. The list of common biology keywords is compiled from <http://www.atcc.org/>. The list contains three categories: cell types, diseases and tissues. Users can specify which category to be used. The function also returns log fold change and FDR of fisher test to check whether each keyword has significantly more appearance compared to base frequency. The base frequency is defined as the number of appearance of the key word in all samples (roughly 40000 samples) included in GEO database.

```
Oct4keywordfreq <- KeyWordFreq(Oct4searchres)
head(Oct4keywordfreq)

##                                     term frequency
## pluripotent                        pluripotent    26
## embryonic                          embryonic     32
## stem cell                          stem cell     15
## embryonic stem cell                embryonic stem cell    8
## pluripotent embryonic stem cell pluripotent embryonic stem cell    1
## inner cell mass                    inner cell mass    2
##                                     logfoldchange  FDR
## pluripotent                        3.208894 1.731386e-27
## embryonic                          1.843194 7.093860e-17
## stem cell                          1.885798 7.217017e-08
## embryonic stem cell                2.610058 1.139724e-06
## pluripotent embryonic stem cell    6.260716 1.601257e-02
## inner cell mass                    3.012281 2.148884e-02
```

In this example embryonic stem cell is among the most frequently appeared biology keyword, which is consistent with known biology [3]. This function allows users to quickly explore the cell types, tissues or organs strongly related to the search term. Users can easily perform second-round search to filter out unrelated records using the online user interface.

## 5 Details of GSM Samples Given GSE Accession ID.

The function `sampledetail` returns an integrated table containing details of all GSM samples for a list of GSE accession ID. This function is especially useful if users want to compare samples from different experiments (GSE).

```
SampleDetail(c("GSE69322", "GSE64008"))
```

##	Experiment	Sample	Title	Type
## 1	GSE69322	GSM1697711	G6L1 FLAG-ChIP 36h+d Line1	ChIP-Seq
## 2	GSE69322	GSM1697712	G6L1 IP control 36h+d Line1	ChIP-Seq
## 3	GSE69322	GSM1697713	G6L1 FLAG-ChIP 36h+d Line2	ChIP-Seq
## 4	GSE69322	GSM1697714	G6L1 IP control 36h+d Line2	ChIP-Seq
## 5	GSE69322	GSM1697715	Gata6 ChIP XEN1	ChIP-Seq
## 6	GSE69322	GSM1697716	IP control XEN1	ChIP-Seq
## 7	GSE69322	GSM1697717	Gata6 ChIP XEN2	ChIP-Seq
## 8	GSE69322	GSM1697718	IP control XEN2	ChIP-Seq
## 9	GSE69322	GSM1697719	G6 K1	RNA-Seq
## 10	GSE69322	GSM1697720	G6 K2	RNA-Seq
## 11	GSE69322	GSM1697721	G6 M1	RNA-Seq
## 12	GSE69322	GSM1697722	G6 M2	RNA-Seq
## 13	GSE69322	GSM1697723	K1	RNA-Seq
## 14	GSE69322	GSM1697724	K2	RNA-Seq
## 15	GSE69322	GSM1697725	M1	RNA-Seq
## 16	GSE69322	GSM1697726	M2	RNA-Seq
## 17	GSE64008	GSM1562325	Nup153-DamID-r1	OTHER
## 18	GSE64008	GSM1562326	Nup153-DamID-r2	OTHER
## 19	GSE64008	GSM1562327	Oct4-DamID-r1	OTHER
## 20	GSE64008	GSM1562328	Oct4-DamID-r2	OTHER
## 21	GSE64008	GSM1562329	GFP-DamID-r1	OTHER
## 22	GSE64008	GSM1562330	GFP-DamID-r2	OTHER
## 23	GSE64008	GSM1562331	Oct4-ChIPseq-shCtrl	ChIP-Seq
## 24	GSE64008	GSM1562332	Oct4-ChIPseq-shNup153-2	ChIP-Seq
## 25	GSE64008	GSM1562333	Ring1B-ChIPseq-shCtrl	ChIP-Seq
## 26	GSE64008	GSM1562334	Ring1B-ChIPseq-shNup153-2	ChIP-Seq
## 27	GSE64008	GSM1562335	Ring1B-ChIPseq-shNup153-4	ChIP-Seq
## 28	GSE64008	GSM1562336	Ring1B-ChIPseq-NeuP	ChIP-Seq
## 29	GSE64008	GSM1562337	Cbx7-ChIPseq-shCtrl	ChIP-Seq
## 30	GSE64008	GSM1562338	Input-ChIPseq	ChIP-Seq
## 31	GSE64008	GSM1562339	ES-RNaseq-shCtrl	RNA-Seq
## 32	GSE64008	GSM1562340	ES-RNaseq-shNup153-2	RNA-Seq
## 33	GSE64008	GSM1562341	ES-RNaseq-shNup153-4	RNA-Seq
## 34	GSE64008	GSM1562342	NeuP-RNaseq-r1	RNA-Seq
## 35	GSE64008	GSM1562343	NeuP-RNaseq-r2	RNA-Seq
##			Source	Organism

```

## 1      Mouse embryonic stem cell line1 Mus musculus
## 2      Mouse embryonic stem cell line1 Mus musculus
## 3      Mouse embryonic stem cell line2 Mus musculus
## 4      Mouse embryonic stem cell line2 Mus musculus
## 5      Extraembryonic endoderm stem cell line1 Mus musculus
## 6      Extraembryonic endoderm stem cell line1 Mus musculus
## 7      Extraembryonic endoderm stem cell line2 Mus musculus
## 8      Extraembryonic endoderm stem cell line2 Mus musculus
## 9              Human embryonic stem cells Homo sapiens
## 10             Human embryonic stem cells Homo sapiens
## 11             Human embryonic stem cells Homo sapiens
## 12             Human embryonic stem cells Homo sapiens
## 13             Human embryonic stem cells Homo sapiens
## 14             Human embryonic stem cells Homo sapiens
## 15             Human embryonic stem cells Homo sapiens
## 16             Human embryonic stem cells Homo sapiens
## 17      mouse embryonic stem cells E14 Mus musculus
## 18      mouse embryonic stem cells E14 Mus musculus
## 19      mouse embryonic stem cells E14 Mus musculus
## 20      mouse embryonic stem cells E14 Mus musculus
## 21      mouse embryonic stem cells E14 Mus musculus
## 22      mouse embryonic stem cells E14 Mus musculus
## 23      mouse embryonic stem cells E14 Mus musculus
## 24      mouse embryonic stem cells E14 Mus musculus
## 25      mouse embryonic stem cells E14 Mus musculus
## 26      mouse embryonic stem cells E14 Mus musculus
## 27      mouse embryonic stem cells E14 Mus musculus
## 28              Neural progenitor cells Mus musculus
## 29      mouse embryonic stem cells E14 Mus musculus
## 30      mouse embryonic stem cells E14 Mus musculus
## 31      mouse embryonic stem cells E14 Mus musculus
## 32      mouse embryonic stem cells E14 Mus musculus
## 33      mouse embryonic stem cells E14 Mus musculus
## 34              Neural progenitor cells Mus musculus
## 35              Neural progenitor cells Mus musculus
##

```

```

## 1      cell type: embryonic stem cell; treatment: ChIP for FLAG-tagged Gata6 in embryonic stem cells
## 2              cell type: embryonic stem cell; treatment: ChIP for FLAG-tagged Gata6 in embryonic stem cells
## 3      cell type: embryonic stem cell; treatment: ChIP for FLAG-tagged Gata6 in embryonic stem cells
## 4              cell type: embryonic stem cell; treatment: ChIP for FLAG-tagged Gata6 in embryonic stem cells
## 5              cell type: extraembryonic endoderm stem cell; treatment: Gata6 ChIP crosslink
## 6              cell type: extraembryonic endoderm stem cell; treatment: Gata6 ChIP crosslink
## 7              cell type: extraembryonic endoderm stem cell; treatment: Gata6 ChIP crosslink
## 8              cell type: extraembryonic endoderm stem cell; treatment: Gata6 ChIP crosslink
## 9              cell type: extraembryonic endoderm stem cell; treatment: Gata6 ChIP crosslink

```





```

## 19 DamID-Seq localization of Oct4 (positive control) replicate 1; Oct4.DamID.bed
## 20 DamID-Seq localization of Oct4 (positive control) replicate 2; Oct4.DamID.bed
## 21 DamID-Seq localization of GFP (negative control) replicate 1; Nup153.DamID.bed
## 22 DamID-Seq localization of GFP (negative control) replicate 2; Nup153.DamID.bed
## 23             ChIP-Seq localization of Oct4 with shCtrl; Oct4.ctrl.bed
## 24             ChIP-Seq localization of Oct4 with shNup153 #2; Oct4.shNup153.bed
## 25             ChIP-Seq localization of Ring1B with shCtrl; Ring1B.shCtrl.bed
## 26             ChIP-Seq localization of Ring1B with shNup153 #2; Ring1B.shNup153-2.bed
## 27             ChIP-Seq localization of Ring1B with shNup153 #4; Ring1B.shNup153-4.bed
## 28             ChIP-Seq localization of Ring1B in NeuP; Ring1B.NeuP.bed
## 29             ChIP-Seq localization of Cbx7 with shCtrl; Cbx7.shCtrl.bed
## 30             ChIP-Seq input control; Oct4.ctrl.bed
## 31             strand specific polyA+ RNA; RNA-Seq of mESC with shCtrl; fpkm.txt
## 32             strand specific polyA+ RNA; RNA-Seq of mESC with shNup153 #2; fpkm.txt
## 33             strand specific polyA+ RNA; RNA-Seq of mESC with shNup153 #4; fpkm.txt
## 34             strand specific polyA+ RNA; RNA-Seq of NeuP; fpkm.txt
## 35             strand specific polyA+ RNA; RNA-Seq of NeuP; fpkm.txt

```

## 6 SEPA GUI

In addition to the basic command lines tools discussed above, SEPA provides a powerful which provides more comprehensive and convenient functions for gene expression pattern analysis. For example, users can easily transform raw gene expression data and convert gene identifiers before the analysis; save the result tables and plots of publication quality; identify genes with different expression patterns on true time and pseudo-time axis. Users are encouraged to use SEPA GUI.

## 7 Reference

- [1]Ron E., Michael D., and Alex E. L. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1), 207-210.
- [2]Yuelin Z., Sean D., Robert S., Paul S. M. and Yidong C. (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics.*, 24(23), 2798-2800.
- [3]Zaehres H., Lensch M.W., Daheron L., Stewart S.A., Itskovitz-Eldor J. and Daley GQ. (2005) High-efficiency RNA interference in human embryonic stem cells. *Stem Cells.*, 23(3), 299-305.