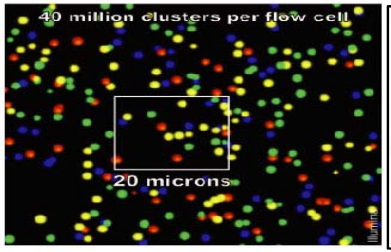
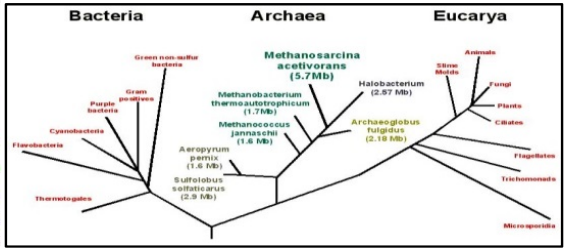
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



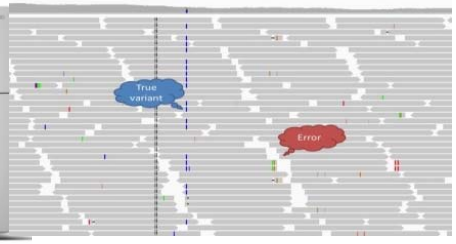


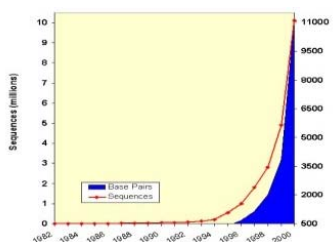
Ontology, and Identification of Molecular Pathways

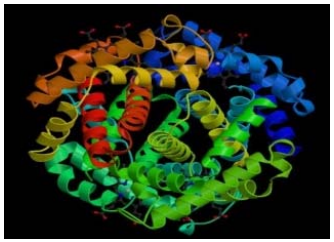
北京大学生物信息学中心 魏丽萍

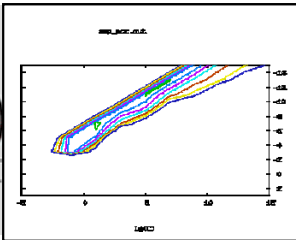
Liping Wei, Ph.D.

Center for Bioinformatics, Peking University

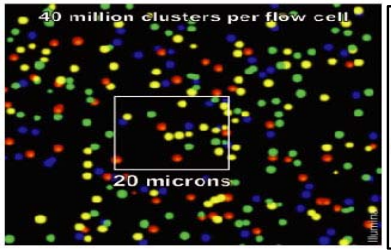


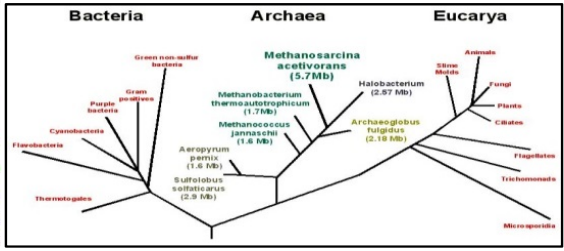






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



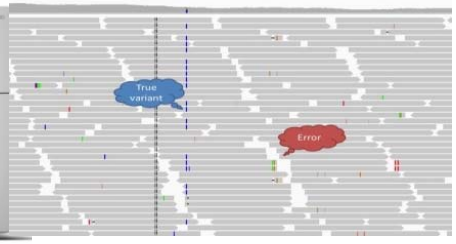


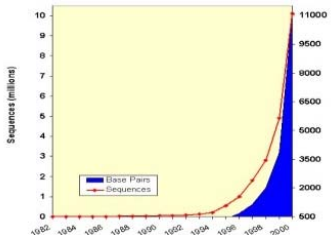
Unit 4: Pathway Identification

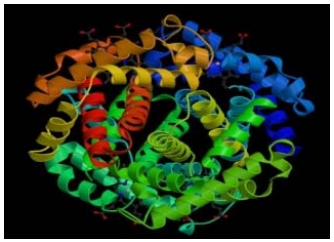
北京大学生物信息学中心 魏丽萍

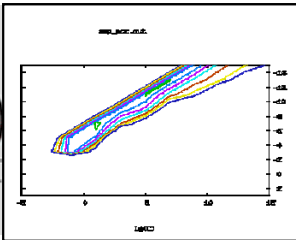
Liping Wei, Ph.D.

Center for Bioinformatics, Peking University









Questions

You have got a set of genes or proteins from your experiments.

How can you find out which pathways the proteins belong to?

How can you find out which the most significant pathways are?

BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 19 2005, pages 3787–3793
doi:10.1093/bioinformatics/bti430

Databases and ontologies

Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary

Xizeng Mao[1,†], Tao Cai[1,†], John G. Olyarchuk[1], and Liping Wei[1,2,*]

[1]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P.R. China and [2]Biomedical Informatics, Department of Medicine, Stanford University School of Medicine. Stanford. CA 94305. USA

Received on January 18, 2005; revised on March 26, 2005; ac

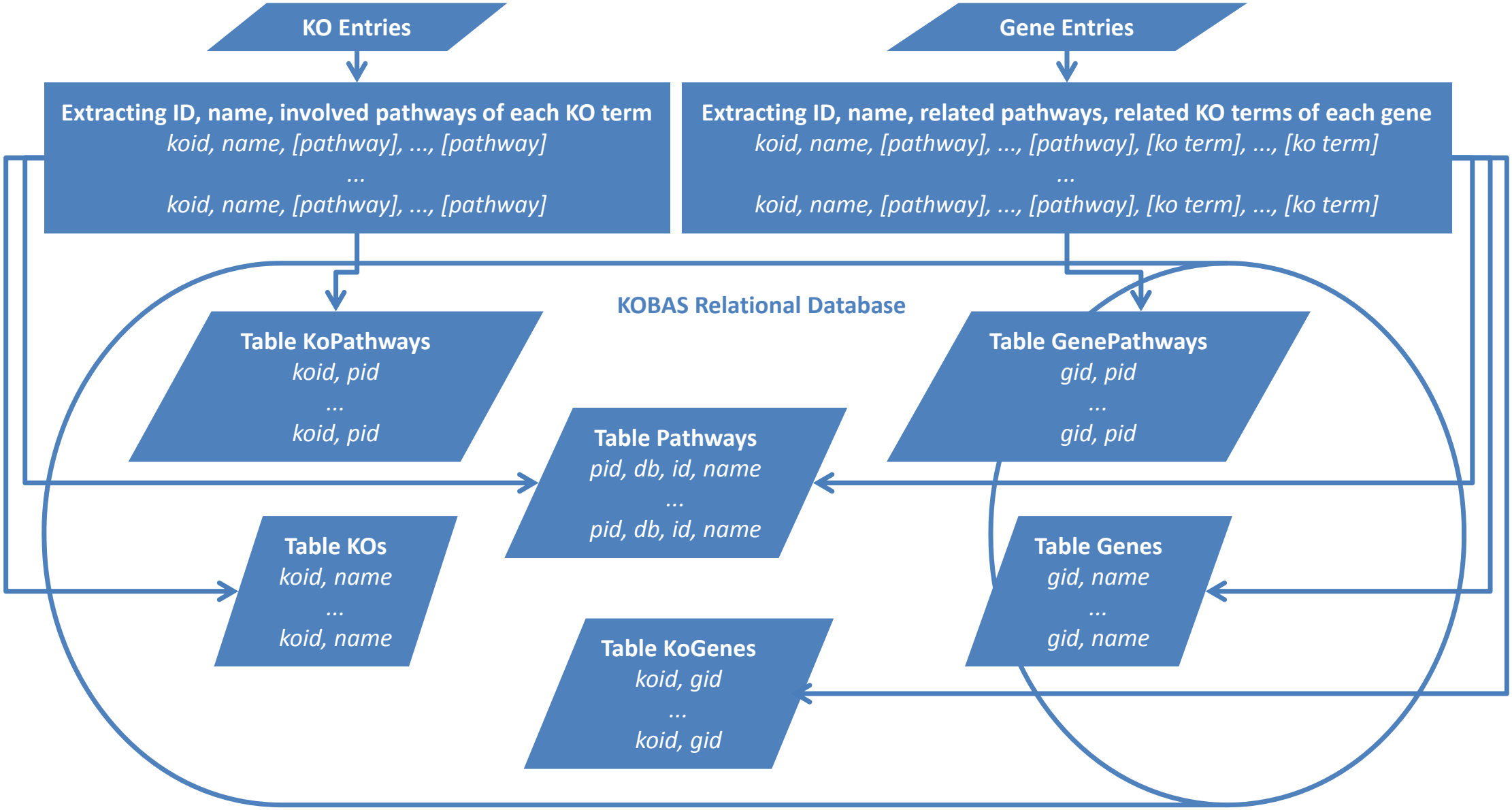
Advance Access publication April 7, 2005

ABSTRACT

Motivation: High-throughput technologies such as DNA seq
and microarrays have created the need for automated annot
large sets of genes, including whole genomes, and automate
fication of pathways. Ontologies, such as the popular Gene C
(GO), provide a common controlled vocabulary for these types
mated analysis. Yet, while GO offers tremendous value, it also has

is stru
have
logy
inclu
ormB
comb

manual curation to assign GO terms to genes in these genomes.



Mapping an input gene to pathway(s)

ID mapping

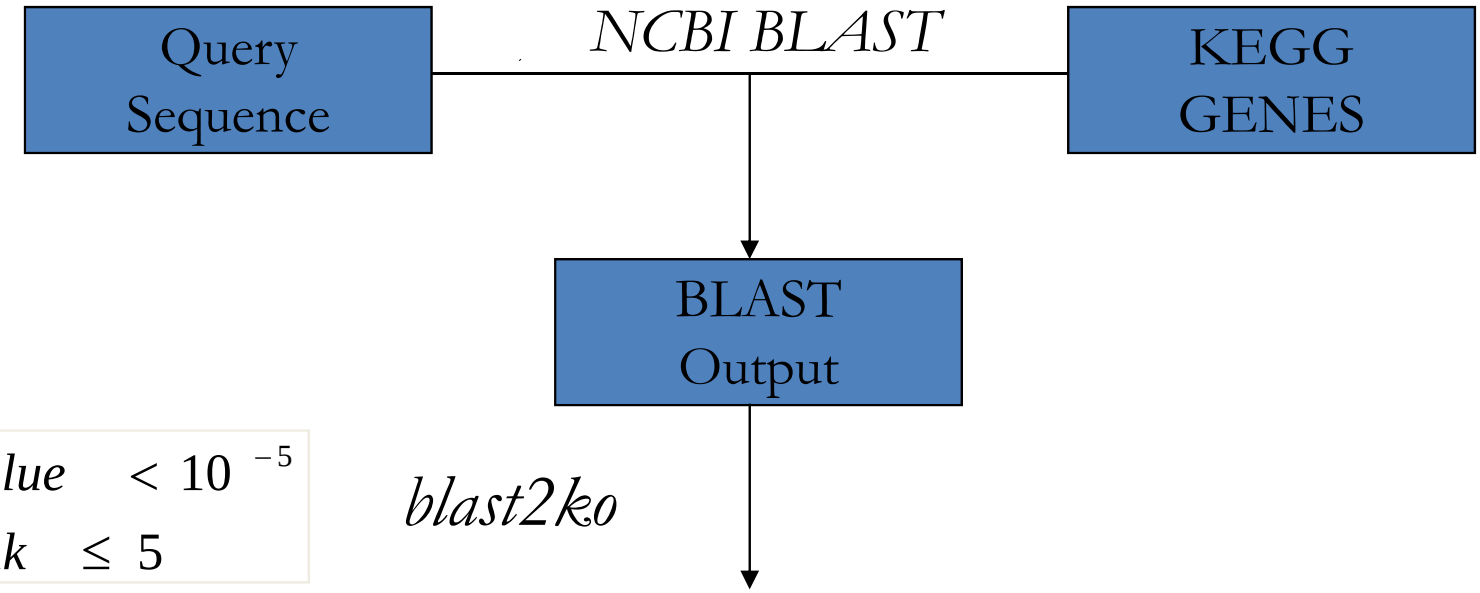
- Genbank GI
- Entrez Gene ID
- Ensembl Gene ID
- UniProtKB AC

Sequence similarity mapping

- newly discovered genes
- genes in a poorly annotated species

Sequence Mapping

Sequence mapping



$evalue < 10^{-5}$

$rank \leq 5$

BLAST Hits

```
eco:b1216 chaA; sodium-calcium/proton antiporte
eco:b3322 pioO, pinO, gspB; calcium-binding pro
ece:Z1991 chaA; sodium-calcium/proton antiporte
ecs:ECs1721 sodium-calcium/proton antiporter; K
ecc:c1676 chaA; calcium/proton antiporter; K073
sty:STY1281 chaA; putative calcium/proton antip
stt:t1680 chaA; putative calcium/proton antipor
spt:SPA1102 chaA; putative calcium/proton antip
sec:SC1765 chaA; CaCA family, sodium-calcium/pr
stm:STM1771 chaA; CaCA family, sodium-calcium/p
ype:YP01958 chaA; calcium/proton antiporter; KO
ype:YP03576 yrbG; putative sodium/calcium excha
ype:YPCD1.30c lcrH, sycD; low calcium response
ypk:y2352 chaA; sodium-calcium/proton antiporte
ypm:YP1703 chaA; calcium/proton antiporter; KO7
ypm:YP3831 putative sodium/calcium exchanger pr
yps:YPTB1956 chaA; calcium/proton antiporter; K
yps:YPTB3520 yrbG; putative sodium/calcium excl
yps:pYV0056 lcrH, sycD; low calcium response pr
sfl:SF1219 chaA; sodium-calcium-proton antiport
sfx:S1303 chaA; sodium-calcium/proton antiporte
ssn:SSO_1961 chaA; sodium-calcium/proton antipo
ssn:SSO_3463 pinO; calcium-binding protein requ
eca:ECA0294 putative sodium/calcium exchanger p
eca:ECA2022 chaA; putative calcium/proton antip
hdu:HD0810 putative sodium/calcium exchange pro
xfa:XF0668 hemolysin-type calcium binding prote
xfa:XF1011 hemolysin-type calcium binding prote
xfa:XF2759 hemolysin-type calcium binding prote
xft:PD0305 frpC; hemolysin-type calcium binding
xft:PD1506 hemolysin-type calcium binding prote
xft:PD2094 frpC; hemolysin-type calcium binding protein
xft:PD2097 frpC; hemolysin-type calcium binding protein
xac:XAC2197 hemolysin-type calcium binding protein
xac:XAC2198 hemolysin-type calcium binding protein
xac:XAC2949 calcium-binding protein
vvu:VV21571 calcium binding protein
vvy:VV0454 putative sodium/calcium exchanger protein
vvy:VVA0109 calcium/proton antiporter; K07300 Ca2+:H+ antiporter
vvy:VVA0384 putative calcium-binding protein
------
```

Entry	b1216 CDS E.coli
Gene name	chaA
Definition	sodium-calcium/proton antiporter
KO	KO: K07300 Ca2+:H+ antiporter OC search OC viewer
Class	Gene catalog
SSDB	Ortholog Paralog Gene cluster
Motif	Pfam: DUF1538 Na_Ca_ex Motif
Other DBs	Wisconsin: b1216 Colibri: chaA RegulonDB: ECK120001216 NCBI-GI: 16129179 NCBI-GeneID: 945790 UniProt: P31801
LinkDB	PDB All DBs
Position	complement(1269972..1271072) Genome map
AA seq	366 aa AA seq DB search MSNAQEAVKTRHKETSLIFPVLALVVLFLWGSSQTLPVVIAINLLALIGILSSAFSVVRH ADVLAHRLGEPYGSLILSLSVVILEVSLISALMATGDAAPTLMRDTLYSIIMIVTGGLVG FSLLLGGRKFATQYMNLFGIKQYLIALFPLAIIVLVFPMALPAANFSTGQALLVALISAA MYGVFLLIQTKTHQSLFVYEHEDDSDDDDPHHGKPSAHSSLWHAIWLIIHLIAVIAVTKM NASSLETLLDSMNAPVAFTGFLVALLILSPEGLGALKAVLNNQVQRAMNLFFGSVLATIS LTVPVVTLIAFMTGNELQFALGAPEMVVMVASLVLCHISFSTGRTNVLNGAAHLALFAAY LMTIFA

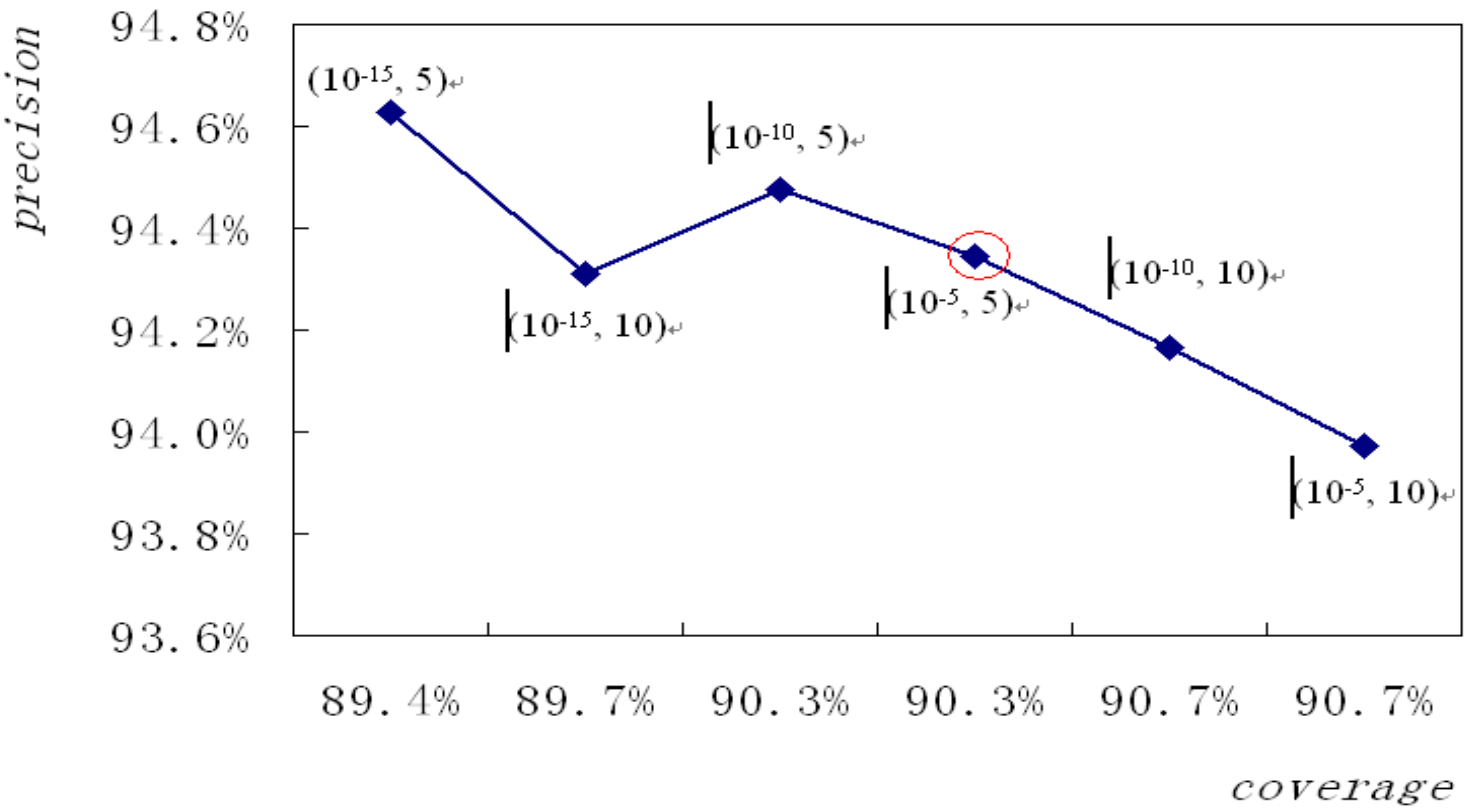
Map to KO
and then to pathway

Evaluation of Pathway Annotation by Sequence Similarity

$$precision = \frac{TP}{TP + FP}$$

$$coverage = \frac{TP}{N}$$

KOBAS Evaluation: *S.cerevisiae*



“New” annotations of yeast genes

Gene ID	Annotation in KEGG	Annotation by KOBAS	Annotation in SGD
YBL075C	None	K03283, TC.HSP70; heat shock protein 70, Hsp70 family	heat-inducible cytosolic member of the 70 kDa heat shock protein family
YCR068W	None	K01046, E3.1.1.3; triacylglycerol lipase	Lipase, required for intravacuolar lysis of autophagic bodies; ...
YDL160C	None	K01509, E3.6.1.3; adenosinetriphosphatase	Cytoplasmic DexD/H-box helicase, stimulates mRNA decapping, ...
YER103W	None	K03283,TC.HSP70; heat shock protein 70, Hsp70 family	member of 70 kDa heat shock protein family

Which pathways are significant?

- Most frequent pathways
- Most enriched pathways

For a specific pathway,

N: the total number of genes

For example, the whole genome

Often called "background"

Only consider genes mapped to pathways.

M: the number of genes in this pathway

n: the total number of query genes

Often called "foreground"

m: the number of query genes in this pathway

When we take n genes from all N background genes, what is the probability of getting m genes from a specific pathway of size M **just by chance**?

Null hypothesis

If this happens just by chance, then this pathway is **not special** for your experiment.

p-value: the probability that the data have occurred by chance assuming that the null hypothesis is true.

If the *p*-value is very small (e.g., <= 1/100), then your observation is unlikely to have occurred just by chance.

Then it is likely that this particular pathway is special for your experiment! You "reject" the null hypothesis.

The smaller the p-value, the more likely it is to reject the null hypothesis.

p-value: the probability of observing data at least as extreme as this, assuming that the null hypothesis is true.

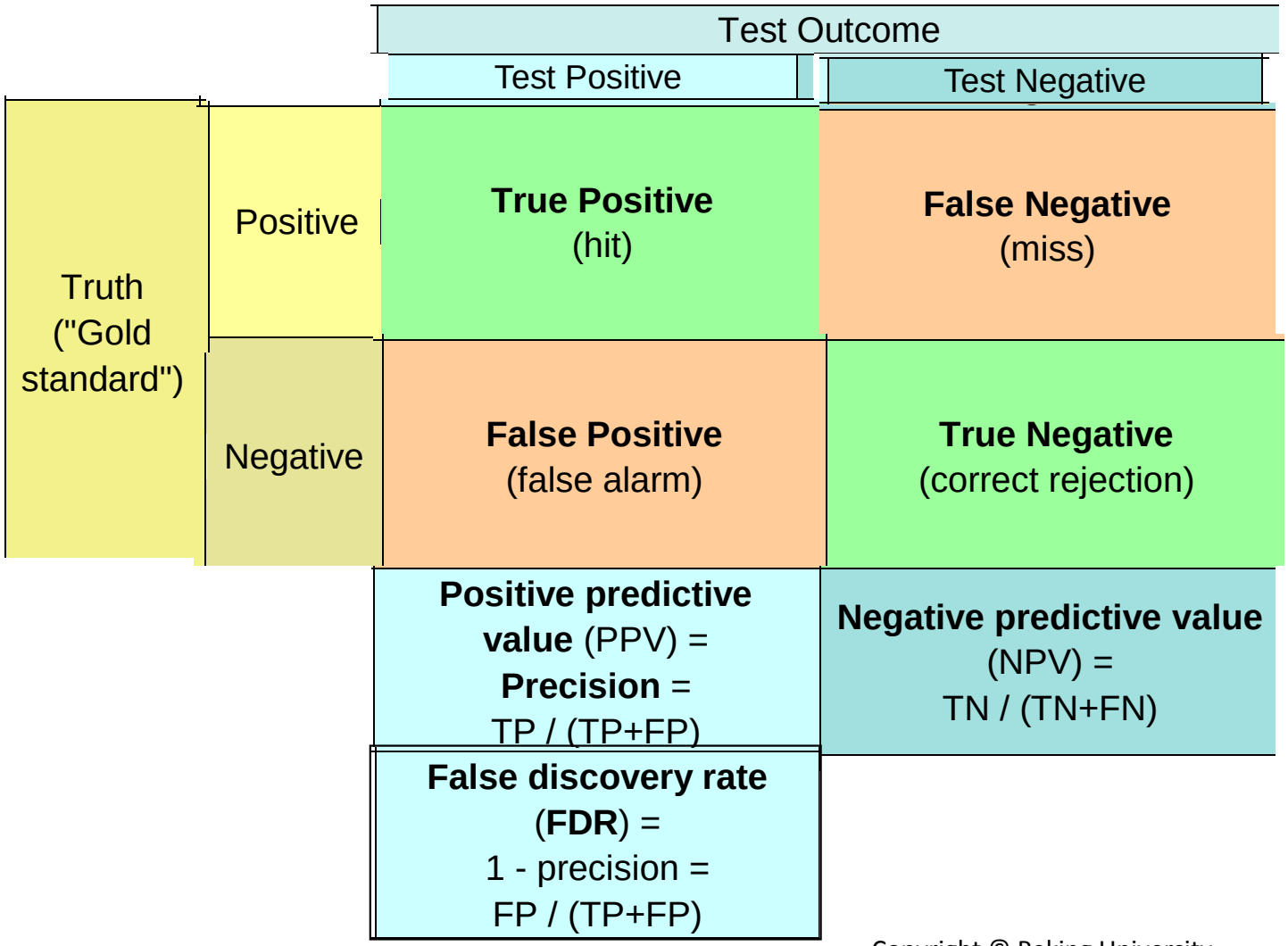
When you randomly draw n genes from a total set of N genes, what is the probability that i of the genes fall in a particular pathway of size M?

This is described by the **hypergeometric distribution** to be $\frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$

p-value = $\sum_{i=m}^{M} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$

= 1- $\sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$

Multiple testing correction after hypergeometric test



Family wise error rate (FWER)

Pr(FP ≥ 1)

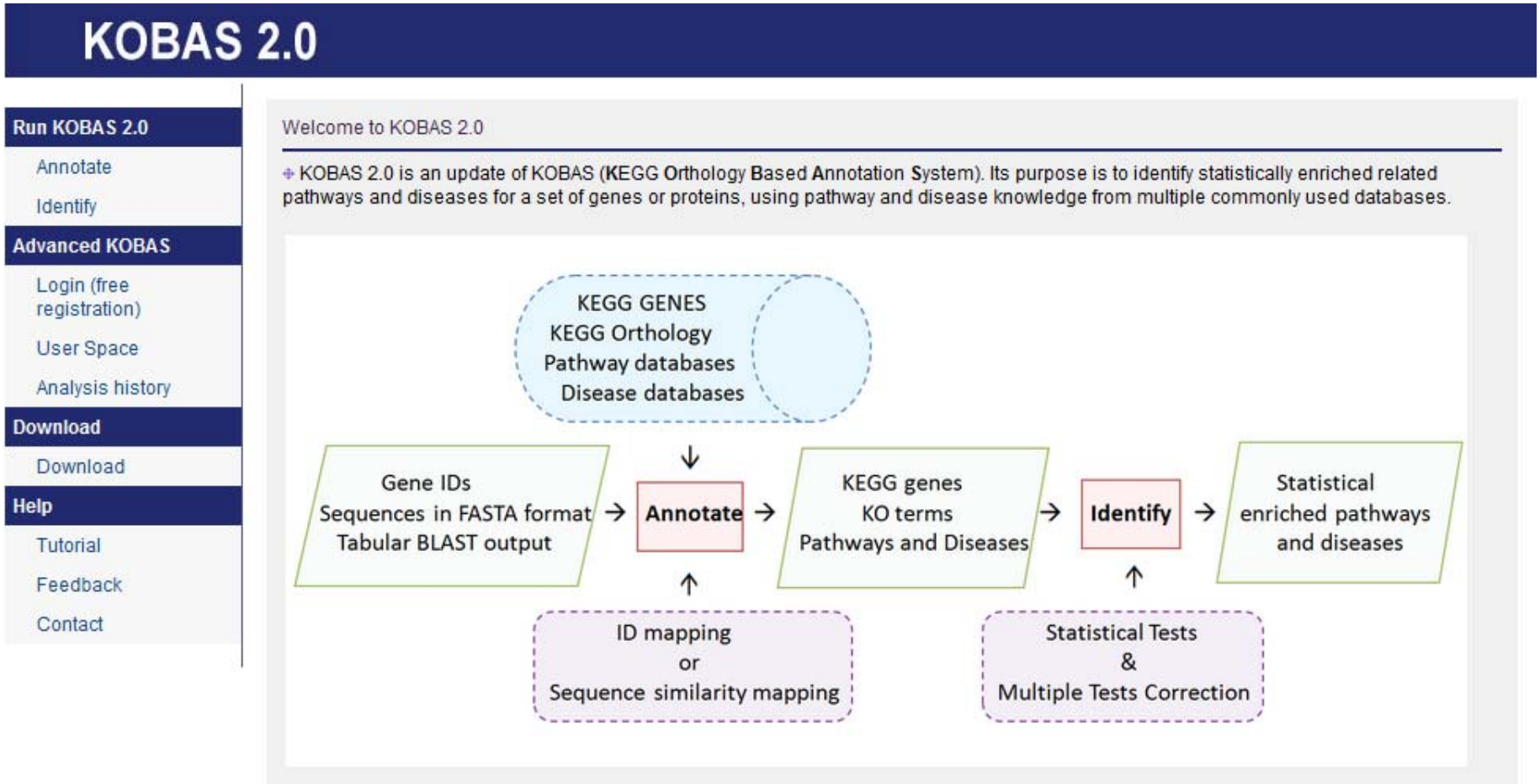
very conservative

False discovery rate (FDR)

E{FP/(TP + FP)}

much less conservative

KOBAS web server

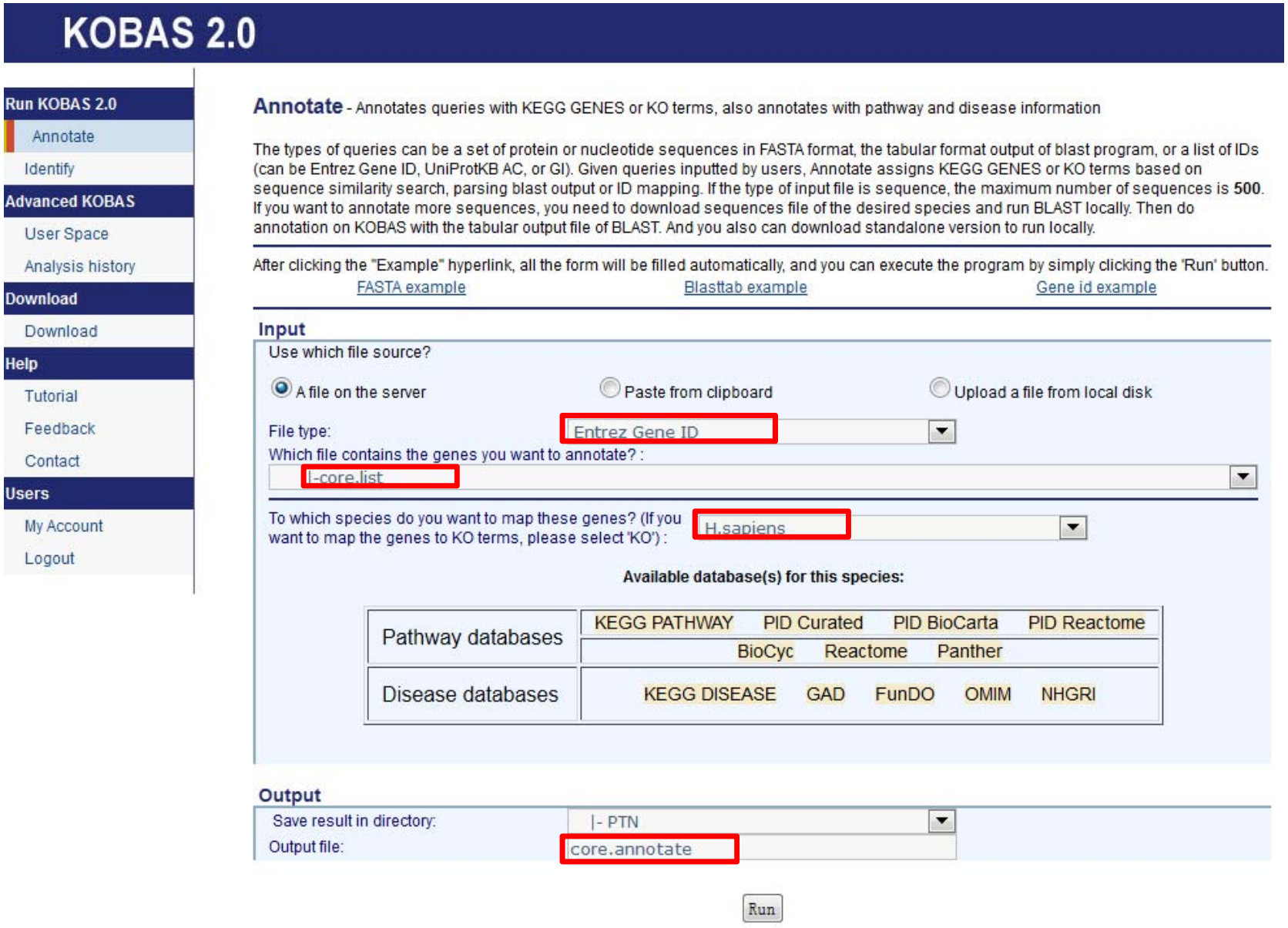


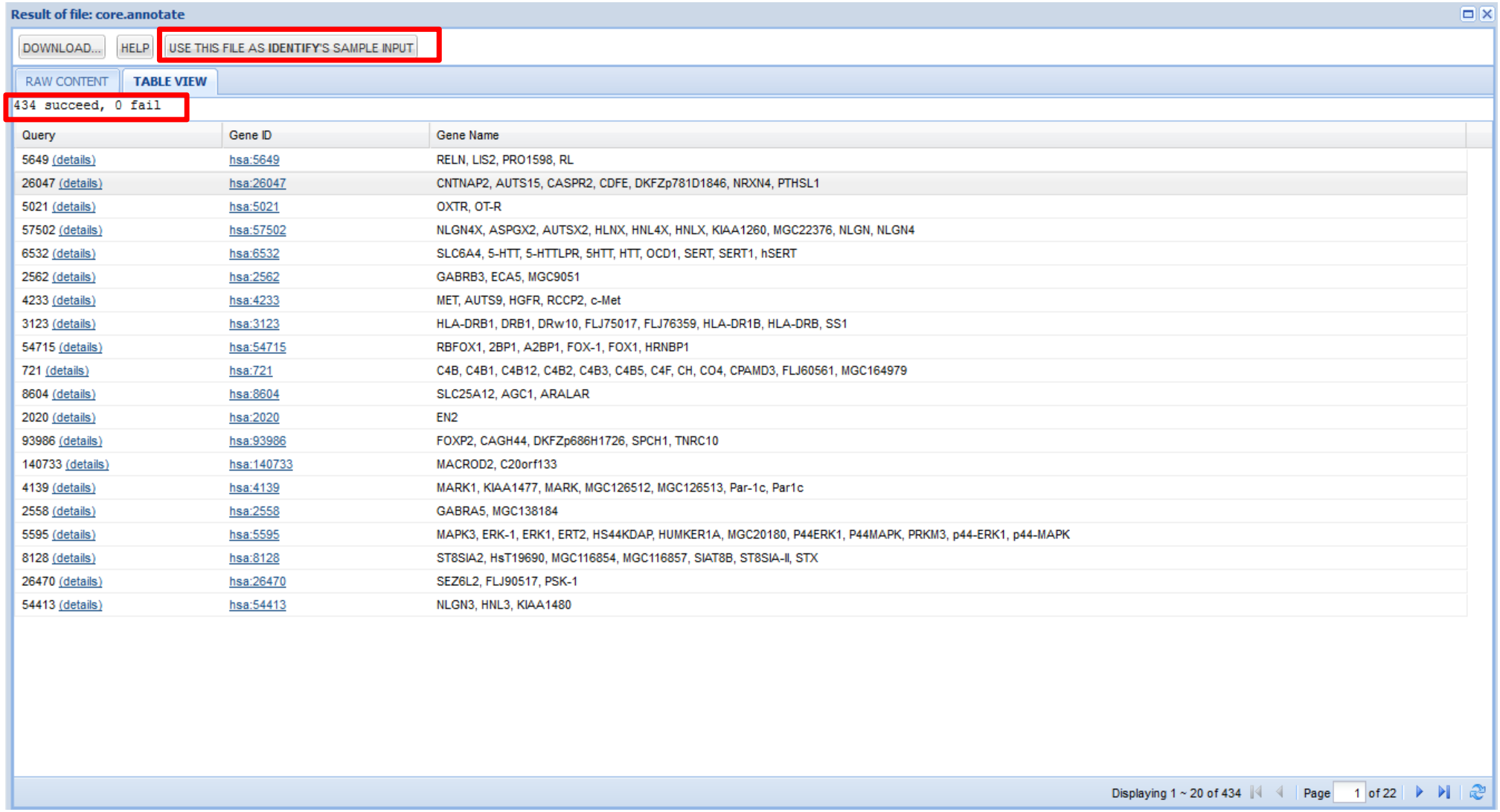
http://kobas.cbi.pku.edu.cn/

databases integrated in KOBAS

Database	URL
KEGG PATHWAY	http://www.genome.jp/kegg/pathway.html
PID	http://pid.nci.nih.gov/
BioCarta (from PID)	http://www.biocarta.com/
Reactome	http://www.reactome.org/
BioCyc	http://biocyc.org/
PANTHER	http://www.pantherdb.org/
Gene Ontology	http://www.geneontology.org/
OMIM	http://www.ncbi.nlm.nih.gov/omim/
KEGG DISEASE	http://www.genome.jp/kegg/disease/
FunDO	http://django.nubic.northwestern.edu/fundo/
GAD	http://geneticassociationdb.nih.gov/
NHGRI GWAS Catalog	http://www.genome.gov/gwastudies/

Xie *et al.*, *Nucleic Acids Res.*, 2011





Result of file: core.annotate

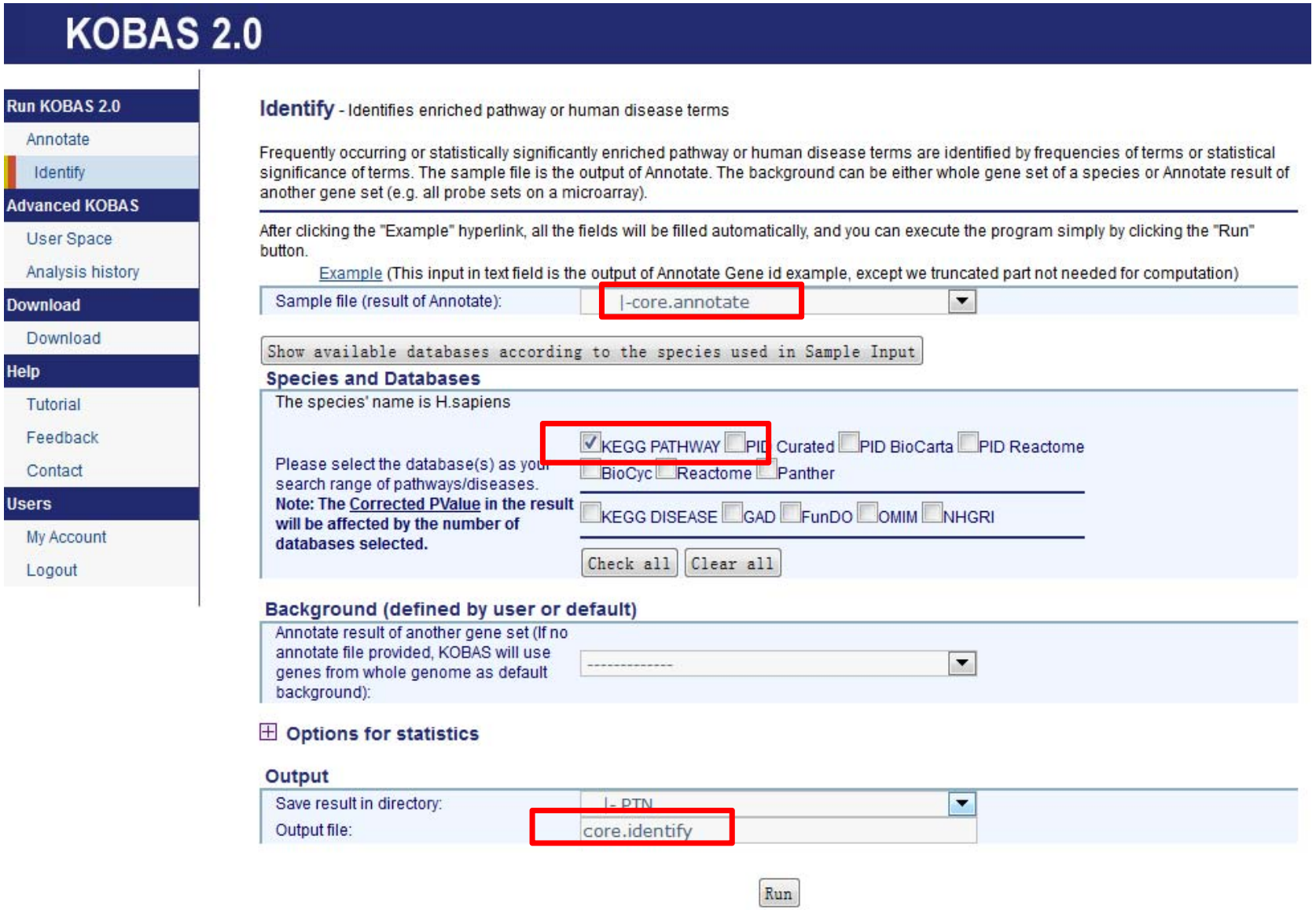
DOWNLOAD... HELP USE THIS FILE AS IDENTIFY'S SAMPLE INPUT

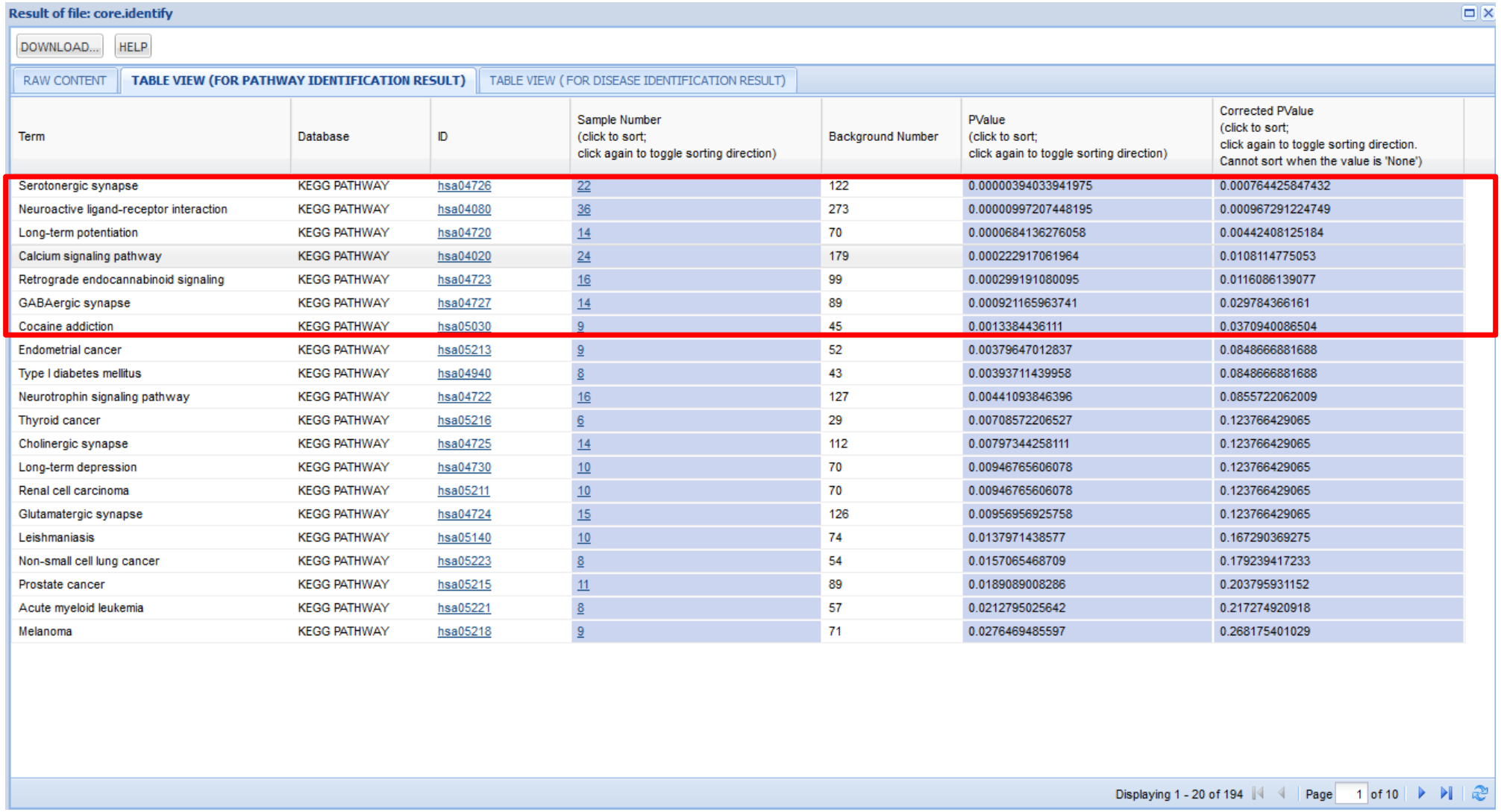
RAW CONTENT TABLE VIEW

434 succeed, 0 fail

Query	Gene ID	Gene Name
5649 (details)	hsa:5649	RELN, LIS2, PRO1598, RL
26047 (details)	hsa:26047	CNTNAP2, AUTS15, CASPR2, CDFE, DKFZp781D1846, NRXN4, PTHSL1
5021 (details)	hsa:5021	OXTR, OT-R
57502 (details)	hsa:57502	NLGN4X, ASPGX2, AUTSX2, HLNX, HNL4X, HNLX, KIAA1260, MGC22376, NLGN, NLGN4
6532 (details)	hsa:6532	SLC6A4, 5-HTT, 5-HTTLPR, 5HTT, HTT, OCD1, SERT, SERT1, hSERT
2562 (details)	hsa:2562	GABRB3, ECA5, MGC9051
4233 (details)	hsa:4233	MET, AUTS9, HGFR, RCCP2, c-Met
3123 (details)	hsa:3123	HLA-DRB1, DRB1, DRw10, FLJ75017, FLJ76359, HLA-DR1B, HLA-DRB, SS1
54715 (details)	hsa:54715	RBFOX1, 2BP1, A2BP1, FOX-1, FOX1, HRNBP1
721 (details)	hsa:721	C4B, C4B1, C4B12, C4B2, C4B3, C4B5, C4F, CH, CO4, CPAMD3, FLJ60561, MGC164979
8604 (details)	hsa:8604	SLC25A12, AGC1, ARALAR
2020 (details)	hsa:2020	EN2
93986 (details)	hsa:93986	FOXP2, CAGH44, DKFZp686H1726, SPCH1, TNRC10
140733 (details)	hsa:140733	MACROD2, C20orf133
4139 (details)	hsa:4139	MARK1, KIAA1477, MARK, MGC126512, MGC126513, Par-1c, Par1c
2558 (details)	hsa:2558	GABRA5, MGC138184
5595 (details)	hsa:5595	MAPK3, ERK-1, ERK1, ERT2, HS44KDAP, HUMKER1A, MGC20180, P44ERK1, P44MAPK, PRKM3, p44-ERK1, p44-MAPK
8128 (details)	hsa:8128	ST8SIA2, HsT19690, MGC116854, MGC116857, SIAT8B, ST8SIA-II, STX
26470 (details)	hsa:26470	SEZ6L2, FLJ90517, PSK-1
54413 (details)	hsa:54413	NLGN3, HNL3, KIAA1480

Displaying 1 ~ 20 of 434 Page 1 of 22





Result of file: core.identify

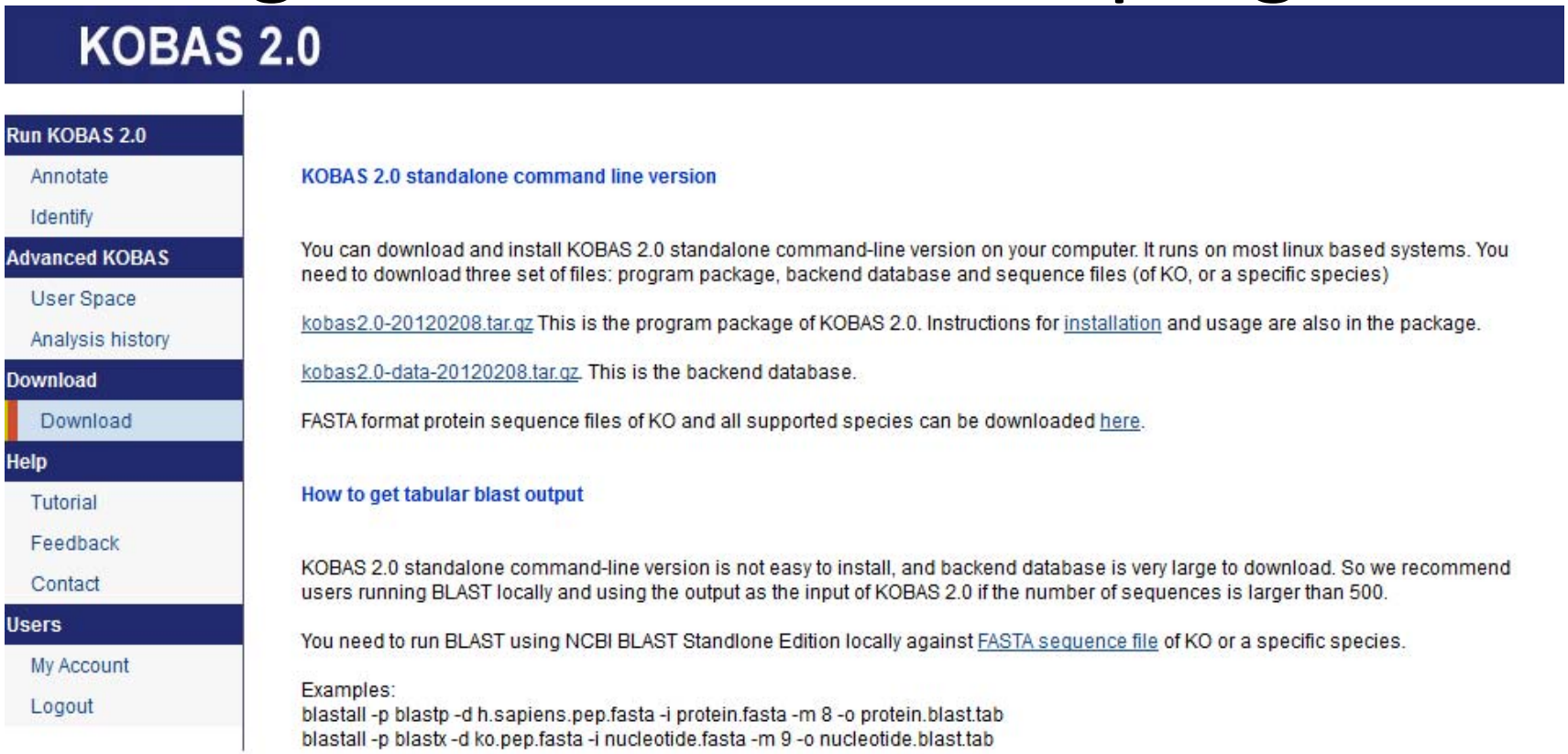
DOWNLOAD... HELP

RAW CONTENT | TABLE VIEW (FOR PATHWAY IDENTIFICATION RESULT) | TABLE VIEW (FOR DISEASE IDENTIFICATION RESULT)

Term	Database	ID	Sample Number (click to sort; click again to toggle sorting direction)	Background Number	PValue (click to sort; click again to toggle sorting direction)	Corrected PValue (click to sort; click again to toggle sorting direction. Cannot sort when the value is 'None')
Serotonergic synapse	KEGG PATHWAY	hsa04726	22	122	0.00000394033941975	0.000764425847432
Neuroactive ligand-receptor interaction	KEGG PATHWAY	hsa04080	36	273	0.00000997207448195	0.000967291224749
Long-term potentiation	KEGG PATHWAY	hsa04720	14	70	0.0000684136276058	0.00442408125184
Calcium signaling pathway	KEGG PATHWAY	hsa04020	24	179	0.000222917061964	0.0108114775053
Retrograde endocannabinoid signaling	KEGG PATHWAY	hsa04723	16	99	0.000299191080095	0.0116086139077
GABAergic synapse	KEGG PATHWAY	hsa04727	14	89	0.000921165963741	0.029784366161
Cocaine addiction	KEGG PATHWAY	hsa05030	9	45	0.0013384436111	0.0370940086504
Endometrial cancer	KEGG PATHWAY	hsa05213	9	52	0.00379647012837	0.0848666881688
Type I diabetes mellitus	KEGG PATHWAY	hsa04940	8	43	0.00393711439958	0.0848666881688
Neurotrophin signaling pathway	KEGG PATHWAY	hsa04722	16	127	0.00441093846396	0.0855722062009
Thyroid cancer	KEGG PATHWAY	hsa05216	6	29	0.00708572206527	0.123766429065
Cholinergic synapse	KEGG PATHWAY	hsa04725	14	112	0.00797344258111	0.123766429065
Long-term depression	KEGG PATHWAY	hsa04730	10	70	0.00946765606078	0.123766429065
Renal cell carcinoma	KEGG PATHWAY	hsa05211	10	70	0.00946765606078	0.123766429065
Glutamatergic synapse	KEGG PATHWAY	hsa04724	15	126	0.00956956925758	0.123766429065
Leishmaniasis	KEGG PATHWAY	hsa05140	10	74	0.0137971438577	0.167290369275
Non-small cell lung cancer	KEGG PATHWAY	hsa05223	8	54	0.0157065468709	0.179239417233
Prostate cancer	KEGG PATHWAY	hsa05215	11	89	0.0189089008286	0.203795931152
Acute myeloid leukemia	KEGG PATHWAY	hsa05221	8	57	0.0212795025642	0.217274920918
Melanoma	KEGG PATHWAY	hsa05218	9	71	0.0276469485597	0.268175401029

Displaying 1 - 20 of 194 | Page 1 of 10

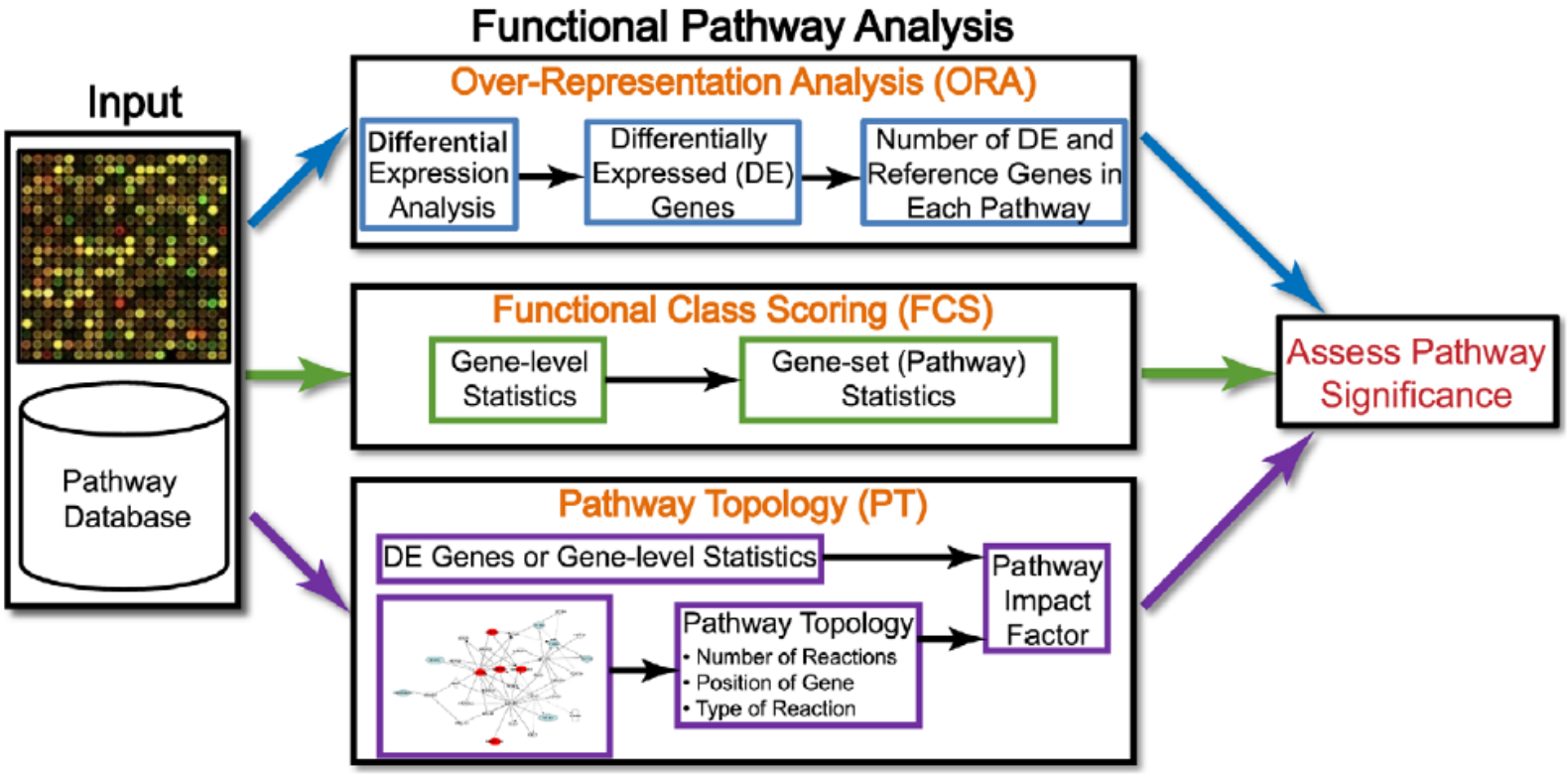
Using KOBAS standalone programs



```
xiec@master ~ $ kobas/scripts/annotate.py -h
Usage: annotate.py [-l] -i infile [-t intype] -s species [-o outfile] [-e evalue] [-r rank] [-n nCPUs

Options:
  -h, --help            show this help message and exit
  -l, --list            list available species, or list available databases
                        for a specific species
  -i INFILE, --infile=INFILE
                        input data file
  -t INTYPE, --intype=INTYPE
                        input type (fasta:pro, fasta:nuc, blastout:xml,
                        blastout:tab, id:ncbigi, id:uniprot, id:ensembl,
                        id:ncbigene), default fasta:pro
  -s SPECIES, --species=SPECIES
                        species abbreviation (for example: ko for KEGG
                        Orthology, hsa for Homo sapiens, mmu for Mus musculus,
                        dme for Drosophila melanogaster, ath for Arabidopsis
                        thaliana, sce for Saccharomyces cerevisiae and eco for
                        Escherichia coli K-12 MG1655)
  -o OUTFILE, --outfile=OUTFILE
                        output file for annotation result, default stdout
  -e EVALUE, --evalue=EVALUE
                        expect threshold for BLAST, default 1e-5
  -r RANK, --rank=RANK  rank cutoff for valid hits from BLAST result, default
                        5
  -n NCPUS, --nCPUs=NCPUS
                        number of CPUs to be used by BLAST, default 1
  -c COVERAGE, --coverage=COVERAGE
                        subject coverage cutoff for BLAST, default 0
  -z ORTHOLOG, --ortholog=ORTHOLOG
                        whether only use orthologs for cross-species
                        annotation or not, default NO (if only use orthologs,
                        please provide the species abbreviation of your input)
```

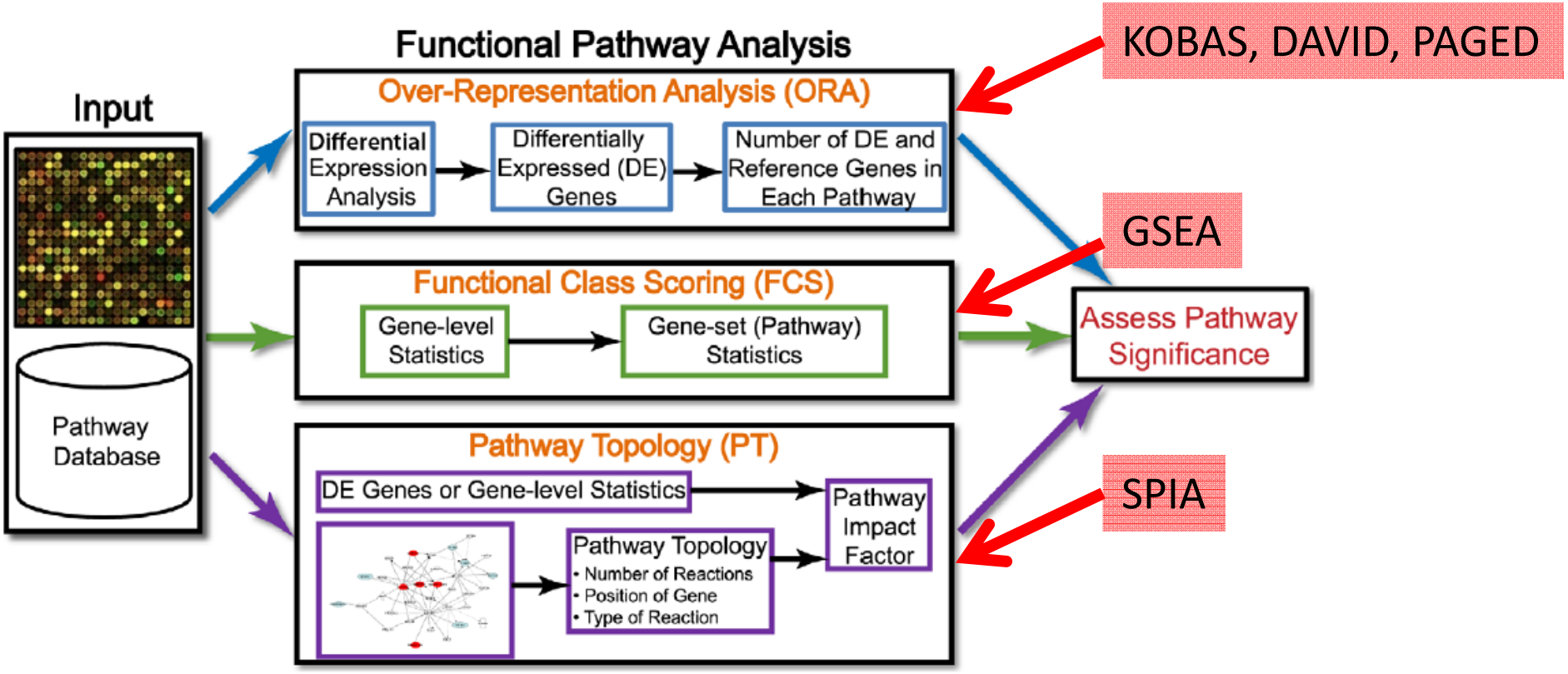
```
xiec@master ~ $ kobas/scripts/identify.py -h
Usage: identify.py -f fgfile [-b bgfile] [-d databases] [-m test] [-n fdr] [-o outfile] [-c cutoff]

Options:
  -h, --help            show this help message and exit
  -f FGFILE, --fgfile=FGFILE
                        foreground file, the output of annotate
  -b BGFILE, --bgfile=BGFILE
                        background file, the output of annotate (3 or 4-letter
                        file name is not allowed), or species abbreviation
                        (for example: hsa for Homo sapiens, mmu for Mus
                        musculus, dme for Drosophila melanogaster, ath for
                        Arabidopsis thaliana, sce for Saccharomyces cerevisiae
                        and eco for Escherichia coli K-12 MG1655), default
                        same species as annotate
  -d DB, --db=DB        databases for selection, 1-letter abbreviation
                        separated by "/": K for KEGG PATHWAY, n for PID, b for
                        BioCarta, R for Reactome, B for BioCyc, p for PANTHER,
                        o for OMIM, k for KEGG DISEASE, f for FunDO, g for
                        GAD, N for NHGRI GWAS Catalog and G for Gene Ontology,
                        default K/n/b/R/B/p/o/k/f/g/N/G
  -m METHOD, --method=METHOD
                        choose statistical test method: b for binomial test, c
                        for chi-square test, h for hypergeometric test /
                        Fisher's exact test, and x for frequency list, default
                        hypergeometric test / Fisher's exact test
  -n FDR, --fdr=FDR     choose false discovery rate (FDR) correction method:
                        BH for Benjamini and Hochberg, BY for Benjamini and
                        Yekutieli, QVALUE, and None, default BH
  -o OUTFILE, --outfile=OUTFILE
                        output file for identification result, default stdout
  -c CUTOFF, --cutoff=CUTOFF
                        terms with less than cutoff number of genes are not
                        used for statistical tests, default 5
```

Other pathway identification methods



Khatri *et al.*, *PLoS Comput. Biol.*, 2012

Other pathway identification methods



Khatri *et al.*, *PLoS Comput. Biol.*, 2012

Summary Questions

How can you map the genes from your experiments to pathways?

How can you find the most significant pathways?

How does KOBAS do these?

How would you do these?

生物信息学：导论与方法
Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University