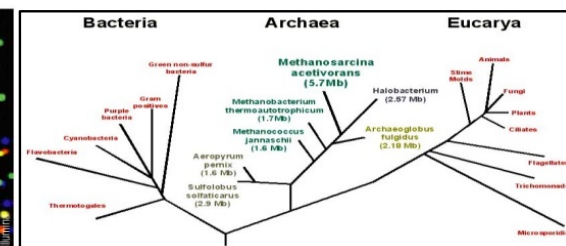
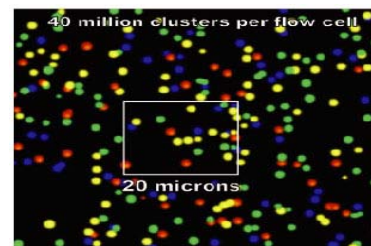






TAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA  
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCC  
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAAC  
 AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTA  
 ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC  
 CTACCCTAACCCCTAACCCCTAACCCCTAACCCCTA  
 ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA

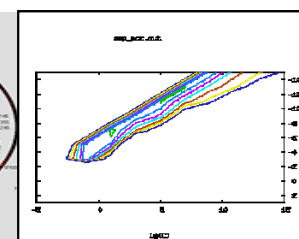
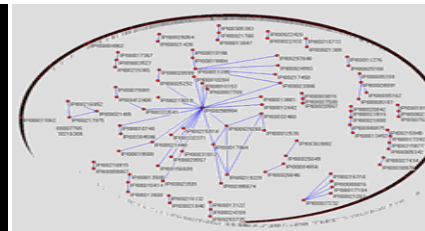
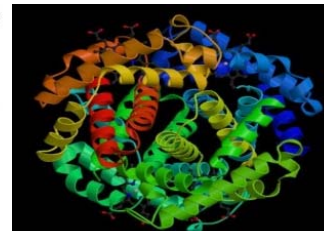
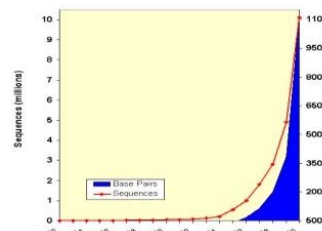
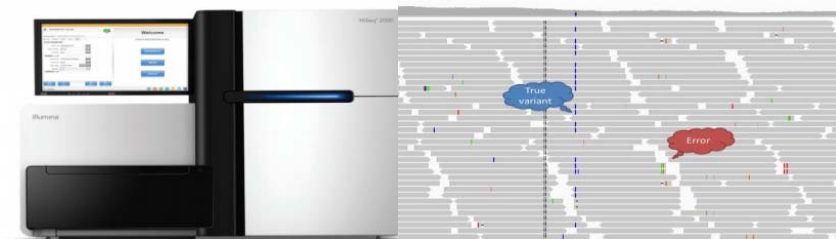


# Unit 4: Pathway Identification

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?



*Databases and ontologies***Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary**Xizeng Mao<sup>1,†</sup>, Tao Cai<sup>1,†</sup>, John G. Olyarchuk<sup>1</sup>, and Liping Wei<sup>1,2,\*</sup><sup>1</sup>Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P.R. China and <sup>2</sup>Biomedical Informatics, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USAReceived on January 18, 2005; revised on March 26, 2005; accepted for publication April 7, 2005  
Advance Access publication April 7, 2005**ABSTRACT****Motivation:** High-throughput technologies such as DNA sequencing and microarrays have created the need for automated annotation of large sets of genes, including whole genomes, and automated identification of pathways. Ontologies, such as the popular Gene Ontology (GO), provide a common controlled vocabulary for these types of automated analysis. Yet, while GO offers tremendous value, it also has

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



**KO Entries**

**Extracting ID, name, involved pathways of each KO term**

*koid, name, [pathway], ..., [pathway]*

...

*koid, name, [pathway], ..., [pathway]*

**Gene Entries**

**Extracting ID, name, related pathways, related KO terms of each gene**

*koid, name, [pathway], ..., [pathway], [ko term], ..., [ko term]*

...

*koid, name, [pathway], ..., [pathway], [ko term], ..., [ko term]*

**KOBAS Relational Database**

**Table KoPathways**

*koid, pid*

...

*koid, pid*

**Table GenePathways**

*gid, pid*

...

*gid, pid*

**Table Pathways**

*pid, db, id, name*

...

*pid, db, id, name*

**Table KOs**

*koid, name*

...

*koid, name*

**Table KoGenes**

*koid, gid*

...

*koid, gid*

**Table Genes**

*gid, name*

...

*gid, name*

# 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

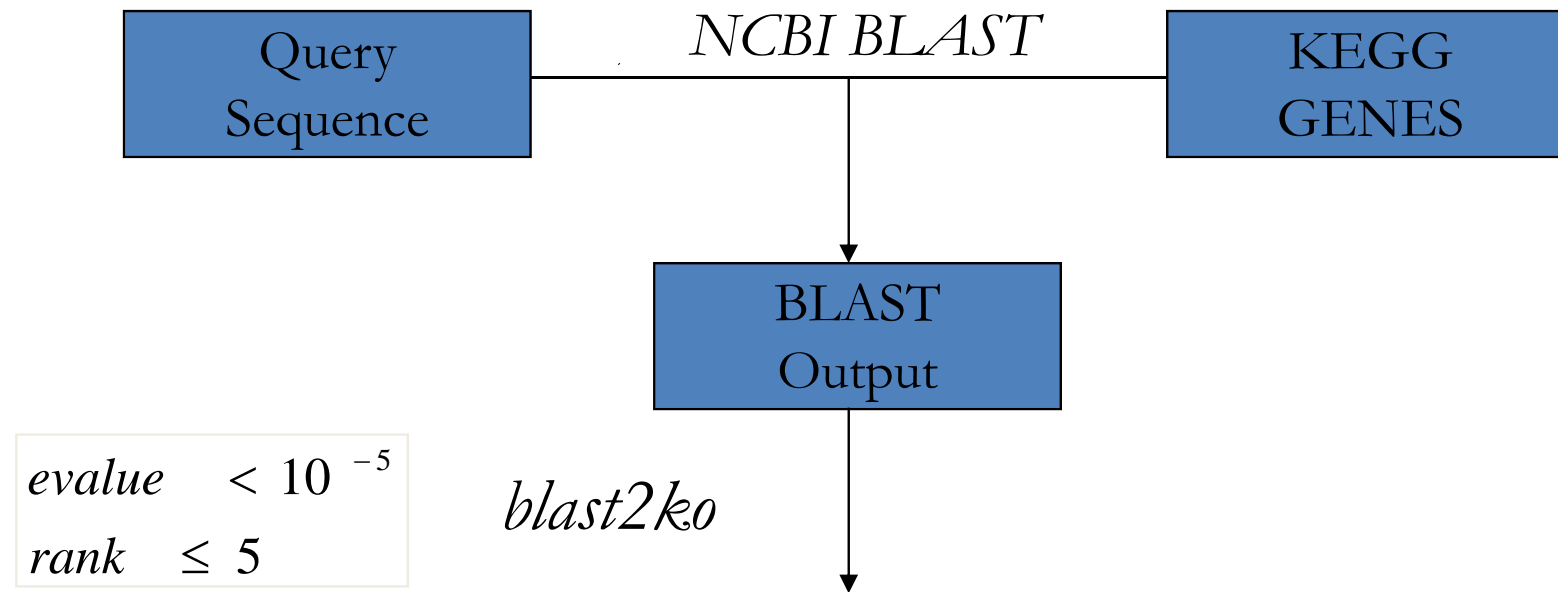
- genes in a poorly annotated species

# Sequence Mapping

Query  
Sequence

KEGG  
GENES

# Sequence mapping





# BLAST Hits

eco:b1216 chaA; sodium-calcium/proton antiporter  
 eco:b3322 pio0, pin0, gspB; calcium-binding protein  
 ece:Z1991 chaA; sodium-calcium/proton antiporter  
 ecs:ECs1721 sodium-calcium/proton antiporter; K07300  
 ecc:cl676 chaA; calcium/proton antiporter; K07300  
 sty:STY1281 chaA; putative calcium/proton antiporter  
 stt:tl680 chaA; putative calcium/proton antiporter  
 spt:SPA1102 chaA; putative calcium/proton antiporter  
 sec:SC1765 chaA; CaCA family, sodium-calcium/proton  
 stm:STM1771 chaA; CaCA family, sodium-calcium/proton  
 ype:YP01958 chaA; calcium/proton antiporter; K07300  
 ype:YP03576 yrbG; putative sodium/calcium exchanger  
 ype:YPCD1.30c lcrH, sycD; low calcium response protein  
 ypk:y2352 chaA; sodium-calcium/proton antiporter  
 ypm:YP1703 chaA; calcium/proton antiporter; K07300  
 ypm:YP3831 putative sodium/calcium exchanger protein  
 yps:YPTB1956 chaA; calcium/proton antiporter; K07300  
 yps:YPTB3520 yrbG; putative sodium/calcium exchanger  
 yps:pYV0056 lcrH, sycD; low calcium response protein  
 sfl:SF1219 chaA; sodium-calcium-proton antiporter  
 sfx:S1303 chaA; sodium-calcium/proton antiporter  
 ssn:SS0\_1961 chaA; sodium-calcium/proton antiporter  
 ssn:SS0\_3463 pin0; calcium-binding protein required for  
 eca:ECA0294 putative sodium/calcium exchanger protein  
 eca:ECA2022 chaA; putative calcium/proton antiporter  
 hdu:HD0810 putative sodium/calcium exchange protein  
 xfa:XF0668 hemolysin-type calcium binding protein  
 xfa:XF1011 hemolysin-type calcium binding protein  
 xfa:XF2759 hemolysin-type calcium binding protein  
 xft:PD0305 frpC; hemolysin-type calcium binding protein  
 xft:PD1506 hemolysin-type calcium binding protein  
 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

<b>Entry</b>	b1216	CDS	E.coli
<b>Gene name</b>	chaA		
<b>Definition</b>	<del>sodium-calcium/proton antiporter</del>		
<b>KO</b>	KO: K07300 Ca2+:H+ antiporter		
	<a href="#">OC search</a> <a href="#">OC viewer</a>		
<b>Class</b>	<a href="#">Gene catalog</a>		
<b>SSDB</b>	<a href="#">Ortholog</a> <a href="#">Paralog</a> <a href="#">Gene cluster</a>		
<b>Motif</b>	Pfam: <a href="#">DUF1538</a> <a href="#">Na_Ca_ex</a> <a href="#">Motif</a>		
<b>Other DBs</b>	Wisconsin: <a href="#">b1216</a> Colibri: <a href="#">chaA</a> RegulonDB: <a href="#">ECK120001216</a> NCBI-GI: <a href="#">16129179</a> NCBI-GeneID: <a href="#">945790</a> UniProt: <a href="#">P31801</a>		
<b>LinkDB</b>	<a href="#">PDB</a> <a href="#">All DBs</a>		
<b>Position</b>	complement(1269972..1271072) <a href="#">Genome map</a>		
<b>AA seq</b>	366 aa <a href="#">AA seq</a> <a href="#">DB search</a>		
	MSNAQEAVKTRHKETSLIFPVLAIVLFLWGSSQTLFVVIAINLLALIGILSSAFSVVRH ADVLAHRLGEPYGSLLSLSVVILEVSLISALMATGDAAPTLMRDTLYSIIMIVTGGLVG FSLLLGGRKFATQYMNLFGIKQYLIALFPLAIIIVLVFPMALPAANFSTGQALLVALISAA MYGVFLLIQTKTHQSLFVYEHEDDSDDDDPHHGKPSAHSSLWHAIIWLIHLLIAVIAVTKM NASSLETLLDSMNAPVAFVTGFLVALLILSPEGLGALKAVLNNQVQRAMNLFFGSVLATIS LTVPVVTLIAFMTGNELQFALGAPEMVVMVASLVLCHISFSTGRITNVLNGAAHLALFAAY LMTIFA		

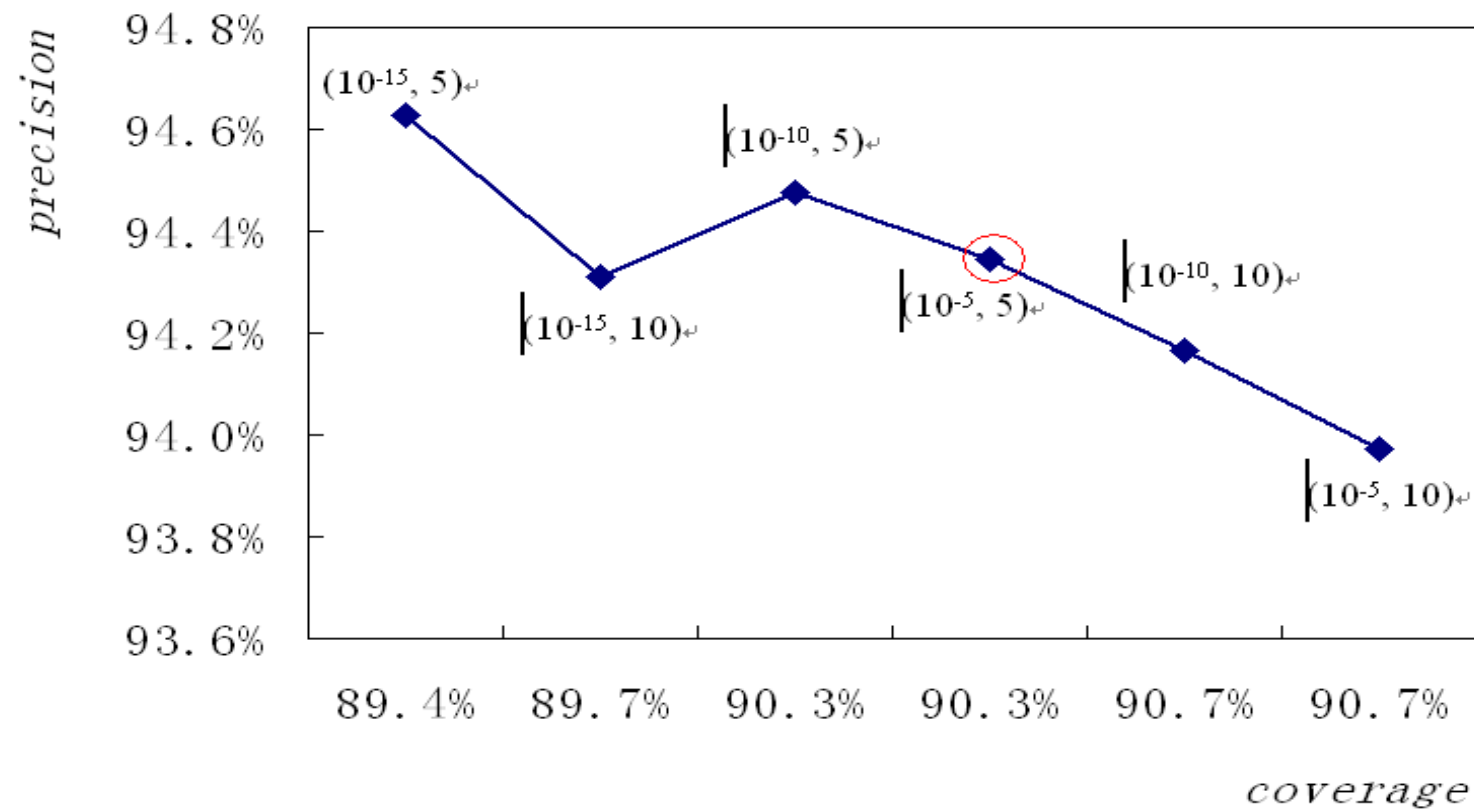
Map to KO  
 and then to pathway

# 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.,  $\leq 1/100$ ), 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  $\frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$

$$\begin{aligned} p\text{-value} &= \sum_{i=m}^M \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \\ &= 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \end{aligned}$$

# Multiple testing correction after hypergeometric test

		Test Outcome	
		Test Positive	Test Negative
Truth ("Gold standard")	Positive	<b>True Positive</b> (hit)	<b>False Negative</b> (miss)
	Negative	<b>False Positive</b> (false alarm)	<b>True Negative</b> (correct rejection)
		<b>Positive predictive value (PPV) =</b> <b>Precision =</b> $TP / (TP+FP)$	<b>Negative predictive value (NPV) =</b> $TN / (TN+FN)$
		<b>False discovery rate (FDR) =</b> $1 - \text{precision} =$ $FP / (TP+FP)$	

Family wise error rate (FWER)

$$\Pr(FP \geq 1)$$

very conservative

False discovery rate (FDR)

$$E\{FP/(TP + FP)\}$$

much less conservative

# KOBAS web server

## KOBAS 2.0

### Run KOBAS 2.0

Annotate

Identify

### Advanced KOBAS

Login (free registration)

User Space

Analysis history

### Download

Download

### Help

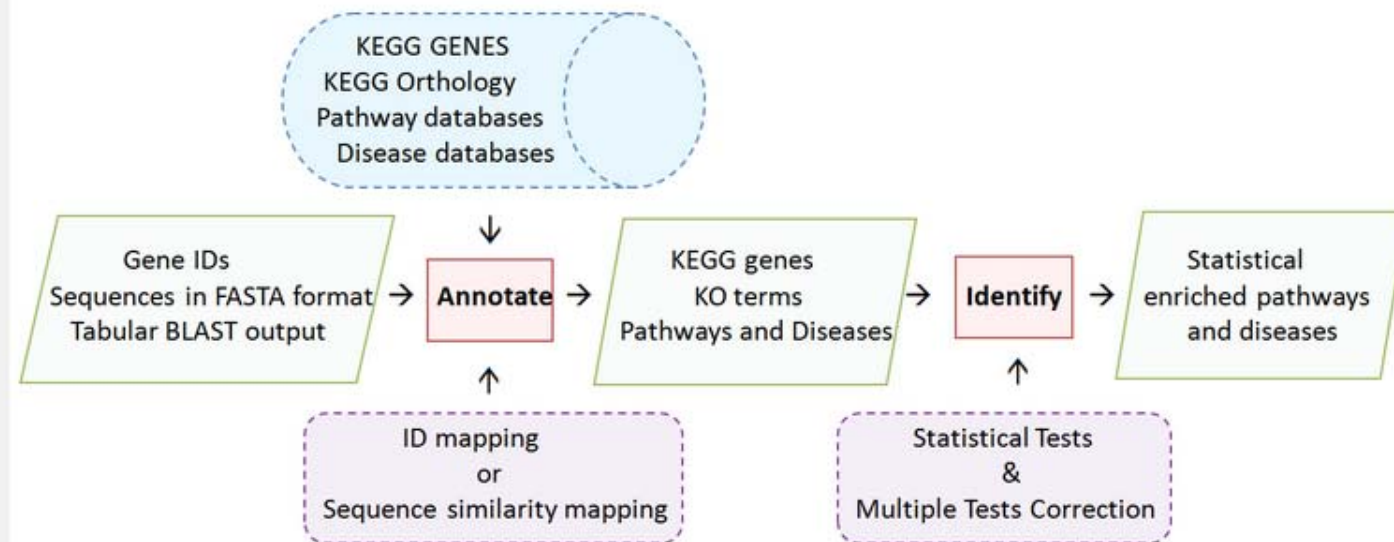
Tutorial

Feedback

Contact

Welcome to KOBAS 2.0

✦ KOBAS 2.0 is an update of KOBAS (KEGG Orthology Based Annotation System). Its purpose is to identify statistically enriched related pathways and diseases for a set of genes or proteins, using pathway and disease knowledge from multiple commonly used databases.



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

Copyright © Peking University

# databases integrated in KOBAS

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

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

# KOBAS 2.0

## Run KOBAS 2.0

Annotate

Identify

## Advanced KOBAS

User Space

Analysis history

## Download

Download

## Help

Tutorial

Feedback

Contact

## Users

My Account

Logout

**Annotate** - Annotates queries with KEGG GENES or KO terms, also annotates with pathway and disease information

The types of queries can be a set of protein or nucleotide sequences in FASTA format, the tabular format output of blast program, or a list of IDs (can be Entrez Gene ID, UniProtKB AC, or GI). Given queries inputted by users, Annotate assigns KEGG GENES or KO terms based on sequence similarity search, parsing blast output or ID mapping. If the type of input file is sequence, the maximum number of sequences is 500. If you want to annotate more sequences, you need to download sequences file of the desired species and run BLAST locally. Then do annotation on KOBAS with the tabular output file of BLAST. And you also can download standalone version to run locally.

After clicking the "Example" hyperlink, all the form will be filled automatically, and you can execute the program by simply clicking the 'Run' button.

[FASTA example](#)

[Blasttab example](#)

[Gene id example](#)

### Input

Use which file source?



A file on the server



Paste from clipboard



Upload a file from local disk

File type:

Entrez Gene ID



Which file contains the genes you want to annotate? :

-core.list



To which species do you want to map these genes? (If you want to map the genes to KO terms, please select 'KO') :

H.sapiens



Available database(s) for this species:

Pathway databases	KEGG PATHWAY	PID Curated	PID BioCarta	PID Reactome	
	BioCyc	Reactome	Panther		
Disease databases	KEGG DISEASE	GAD	FunDO	OMIM	NHGRI

### Output

Save result in directory:

- PTN



Output file:

core.annotate

Run

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

Displaying 1 ~ 20 of 434 Page 1 of 22



# KOBAS 2.0

## Run KOBAS 2.0

Annotate

Identify

## Advanced KOBAS

User Space

Analysis history

## Download

Download

## Help

Tutorial

Feedback

Contact

## Users

My Account

Logout

### Identify - Identifies enriched pathway or human disease terms

Frequently occurring or statistically significantly enriched pathway or human disease terms are identified by frequencies of terms or statistical significance of terms. The sample file is the output of Annotate. The background can be either whole gene set of a species or Annotate result of another gene set (e.g. all probe sets on a microarray).

After clicking the "Example" hyperlink, all the fields will be filled automatically, and you can execute the program simply by clicking the "Run" button.

[Example](#) (This input in text field is the output of Annotate Gene id example, except we truncated part not needed for computation)

Sample file (result of Annotate):

Show available databases according to the species used in Sample Input

#### Species and Databases

The species' name is H.sapiens

Please select the database(s) as your search range of pathways/diseases.

☒ KEGG PATHWAY ☐ PID Curated ☐ PID BioCarta ☐ PID Reactome  
☐ BioCyc ☐ Reactome ☐ Panther

**Note: The Corrected PValue in the result will be affected by the number of databases selected.**

☐ KEGG DISEASE ☐ GAD ☐ FunDO ☐ OMIM ☐ NHGRI

#### Background (defined by user or default)

Annotate result of another gene set (If no annotate file provided, KOBAS will use genes from whole genome as default background):

#### Options for statistics

##### Output

Save result in directory:

Output file:



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	<a href="#">hsa04726</a>	<a href="#">22</a>	122	0.00000394033941975	0.000764425847432
Neuroactive ligand-receptor interaction	KEGG PATHWAY	<a href="#">hsa04080</a>	<a href="#">36</a>	273	0.00000997207448195	0.000967291224749
Long-term potentiation	KEGG PATHWAY	<a href="#">hsa04720</a>	<a href="#">14</a>	70	0.0000684136276058	0.00442408125184
Calcium signaling pathway	KEGG PATHWAY	<a href="#">hsa04020</a>	<a href="#">24</a>	179	0.000222917061964	0.0108114775053
Retrograde endocannabinoid signaling	KEGG PATHWAY	<a href="#">hsa04723</a>	<a href="#">16</a>	99	0.000299191080095	0.0116086139077
GABAergic synapse	KEGG PATHWAY	<a href="#">hsa04727</a>	<a href="#">14</a>	89	0.000921165963741	0.029784366161
Cocaine addiction	KEGG PATHWAY	<a href="#">hsa05030</a>	<a href="#">9</a>	45	0.0013384436111	0.0370940086504
Endometrial cancer	KEGG PATHWAY	<a href="#">hsa05213</a>	<a href="#">9</a>	52	0.00379647012837	0.0848666881688
Type I diabetes mellitus	KEGG PATHWAY	<a href="#">hsa04940</a>	<a href="#">8</a>	43	0.00393711439958	0.0848666881688
Neurotrophin signaling pathway	KEGG PATHWAY	<a href="#">hsa04722</a>	<a href="#">16</a>	127	0.00441093846396	0.0855722062009
Thyroid cancer	KEGG PATHWAY	<a href="#">hsa05216</a>	<a href="#">6</a>	29	0.00708572206527	0.123766429065
Cholinergic synapse	KEGG PATHWAY	<a href="#">hsa04725</a>	<a href="#">14</a>	112	0.00797344258111	0.123766429065
Long-term depression	KEGG PATHWAY	<a href="#">hsa04730</a>	<a href="#">10</a>	70	0.00946765606078	0.123766429065
Renal cell carcinoma	KEGG PATHWAY	<a href="#">hsa05211</a>	<a href="#">10</a>	70	0.00946765606078	0.123766429065
Glutamatergic synapse	KEGG PATHWAY	<a href="#">hsa04724</a>	<a href="#">15</a>	126	0.00956956925758	0.123766429065
Leishmaniasis	KEGG PATHWAY	<a href="#">hsa05140</a>	<a href="#">10</a>	74	0.0137971438577	0.167290369275
Non-small cell lung cancer	KEGG PATHWAY	<a href="#">hsa05223</a>	<a href="#">8</a>	54	0.0157065468709	0.179239417233
Prostate cancer	KEGG PATHWAY	<a href="#">hsa05215</a>	<a href="#">11</a>	89	0.0189089008286	0.203795931152
Acute myeloid leukemia	KEGG PATHWAY	<a href="#">hsa05221</a>	<a href="#">8</a>	57	0.0212795025642	0.217274920918
Melanoma	KEGG PATHWAY	<a href="#">hsa05218</a>	<a href="#">9</a>	71	0.0276469485597	0.268175401029

Displaying 1 - 20 of 194 Page 1 of 10

# Using KOBAS standalone programs

## KOBAS 2.0

### Run KOBAS 2.0

Annotate

Identify

### Advanced KOBAS

User Space

Analysis history

### Download

Download

### Help

Tutorial

Feedback

Contact

### Users

My Account

Logout

### KOBAS 2.0 standalone command line version

You can download and install KOBAS 2.0 standalone command-line version on your computer. It runs on most linux based systems. You need to download three set of files: program package, backend database and sequence files (of KO, or a specific species)

[kobas2.0-20120208.tar.gz](#) This is the program package of KOBAS 2.0. Instructions for [installation](#) and usage are also in the package.

[kobas2.0-data-20120208.tar.gz](#) This is the backend database.

FASTA format protein sequence files of KO and all supported species can be downloaded [here](#).

### How to get tabular blast output

KOBAS 2.0 standalone command-line version is not easy to install, and backend database is very large to download. So we recommend users running BLAST locally and using the output as the input of KOBAS 2.0 if the number of sequences is larger than 500.

You need to run BLAST using NCBI BLAST Standalone Edition locally against [FASTA sequence file](#) of KO or a specific species.

Examples:

```
blastall -p blastp -d h.sapiens.pep.fasta -i protein.fasta -m 8 -o protein.blast.tab
```

```
blastall -p blastx -d ko.pep.fasta -i nucleotide.fasta -m 9 -o nucleotide.blast.tab
```

Sometimes, the blast output is too large, and it will take too much time for uploading. In this situation, you can use this [script](#) to slim the output.

Usage: `./slim_blast_output.py -i blast_output_file -e evalute > slimmed_file`

```
xiec@master ~ $ kobas/scripts/annotate.py -h
Usage: annotate.py [-l] -i infile [-t intype] -s species [-o outfile] [-e evaluate] [-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, --evaluate=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)
```

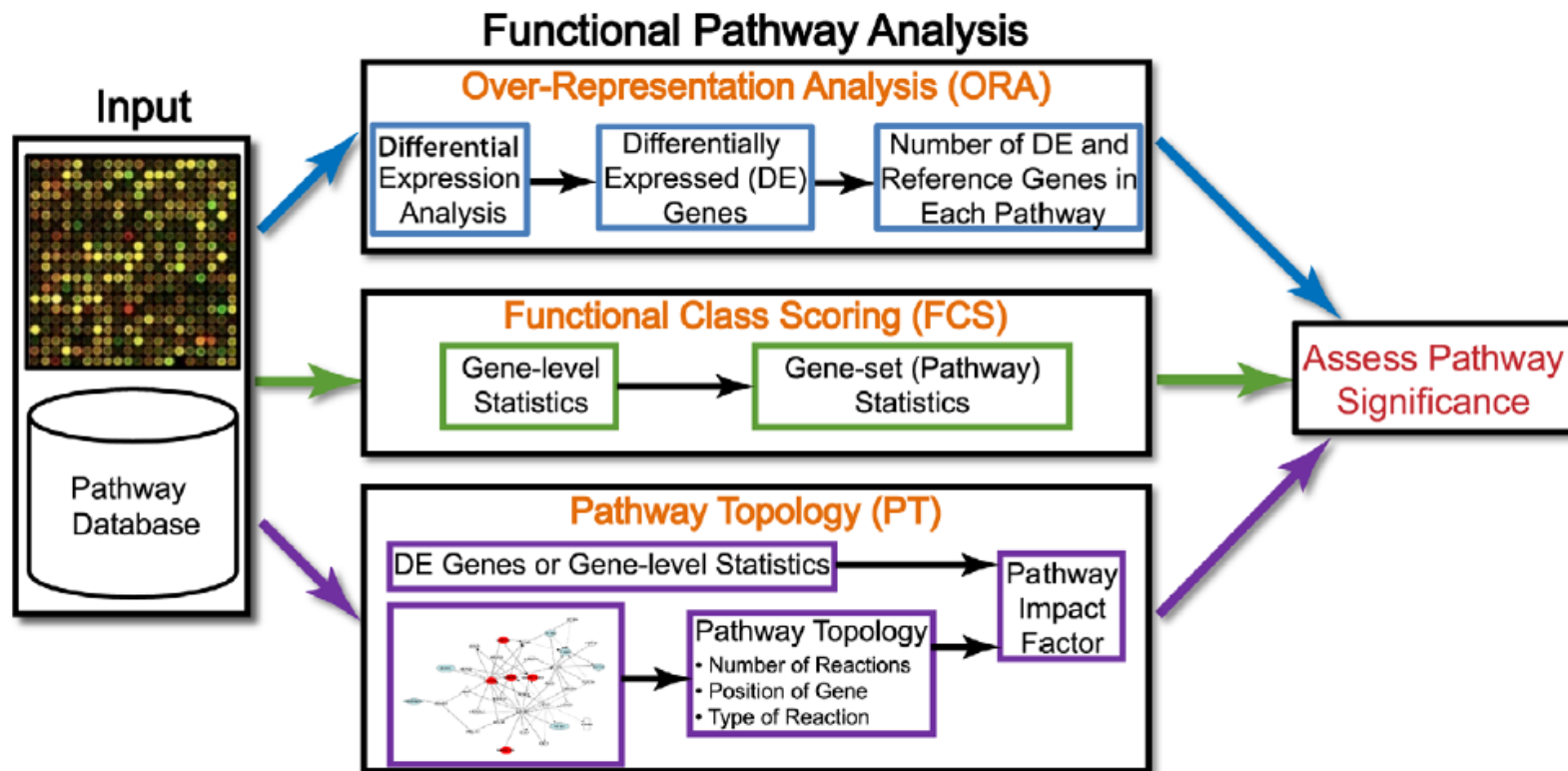
```

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

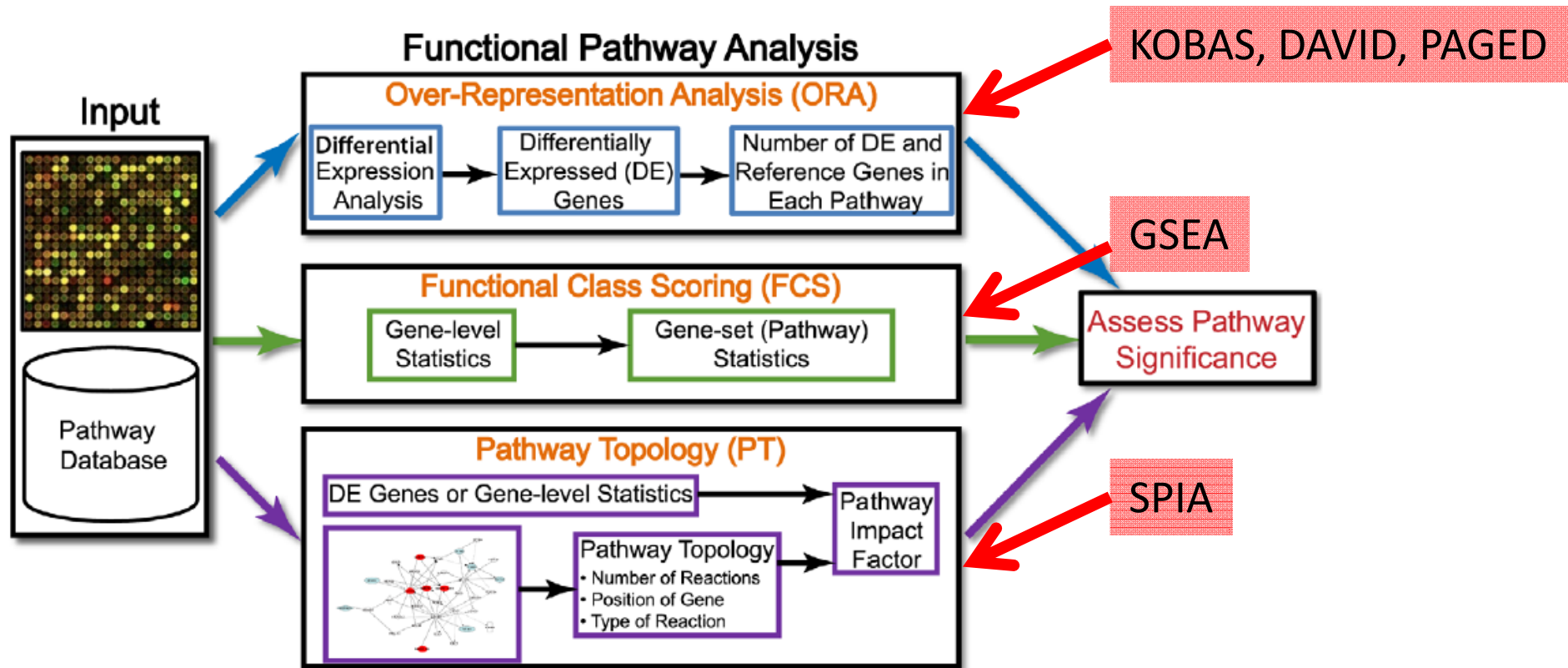
```

# 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



<https://www.coursera.org/course/pkubioinfo>