Background

- We have a Solexa reads for yeast genome
- We want to know difference between mutant and wild type
- This time I focused on the details of mutated protein information

Methods

Overview

There are already numbers of softwares for this process

list of examples
http://seqanswers.com/wiki/Special:BrowseData

Mapping the reads to Genome

SNP detection

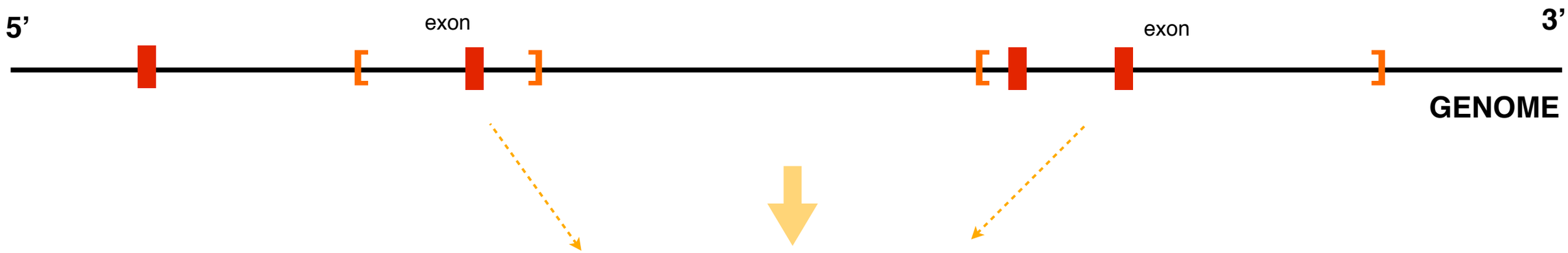
Coding !!

Find Affected Protein

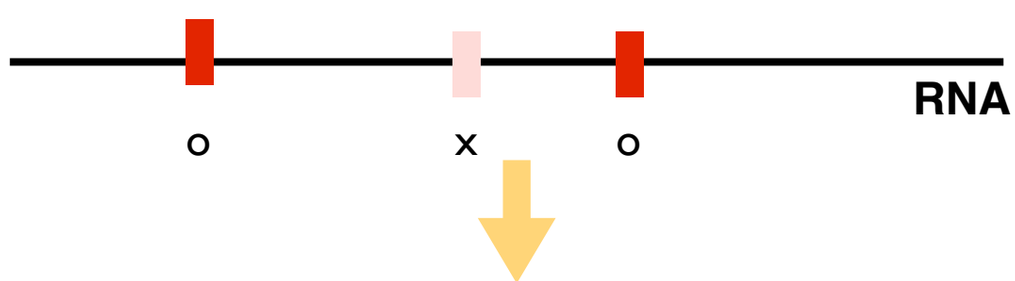
Give Protein Annotation

1. We have the Detected SNP positions on the genome

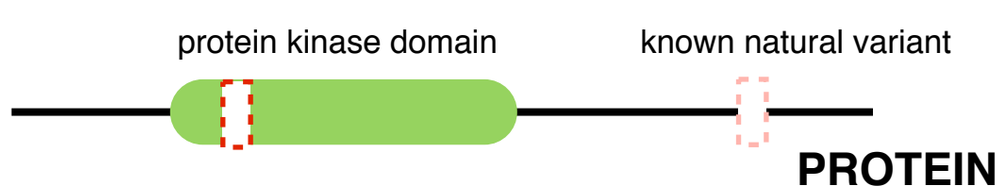
2. Only look at SNPs in the CDS region



3. Look only at SNPs that changes the AA sequence



4. look at features of protein at the mutation position

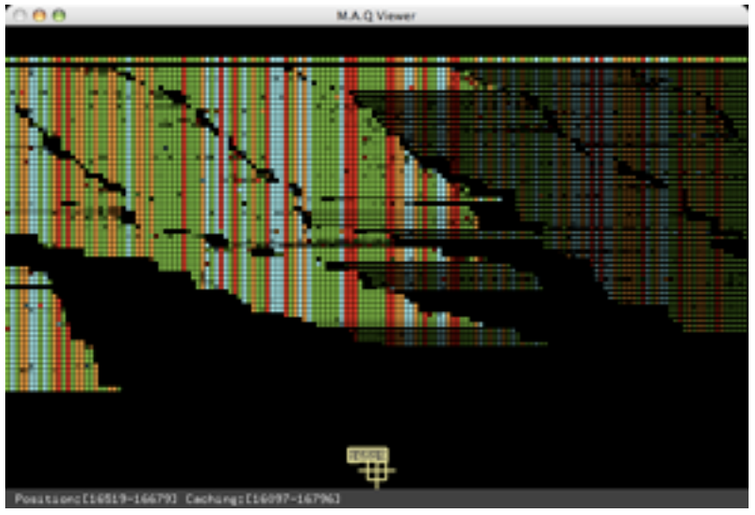


Methods

Software & Database

MAQ

- One of the most-used Mapping & SNP detection software for Solexa reads
- maqview, wich visualized the mapping result of MAQ is also available



http://maq.sourceforge.net/

Perl

- Useful when you handle string characters (like DNA, RNA, AA sequences)
- Easy to code (no need to compile)
- lots of modules in CPAN (http://search.cpan.org/)

http://www.perl.com/

bioperl

- Library of perl modules for bioinformatics
- Has many Classes for analysis
- Has lots of HOWTOs and documentation
- Installed in MSI server
- There are also biopython, bioruby, biojava,etc (FYI)

http://www.bioperl.org/wiki/Main_Page

SRA

- Powered by NCBI
- Database that collects nextgen sequencing data
- I used this for the dummy data
- sometime, it is called "short read archive"
  and sometimes it is called "sequence read archive"

SGD

- Database for yeast
- Has Pathway information, Genome information
- SGD id is the primary identifier for each gene in this database

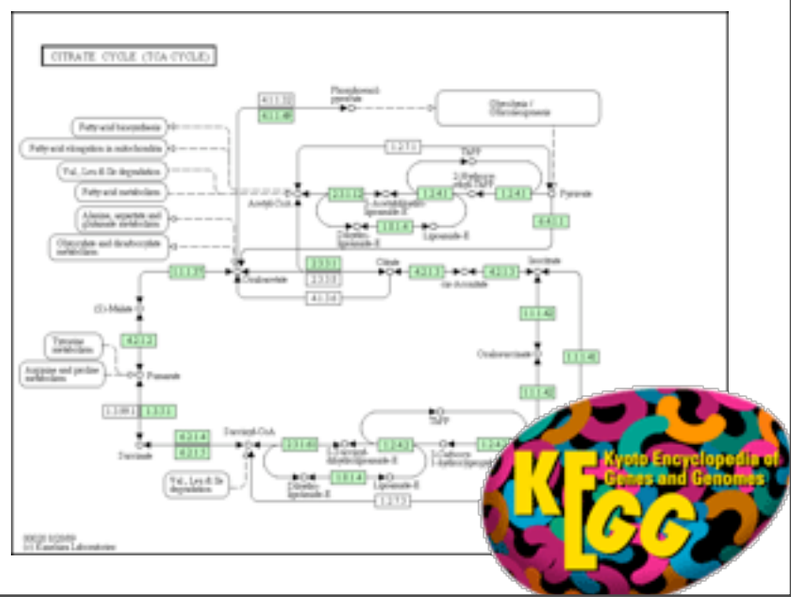
UniProt(Swiss-prot)

- Human curated protein database(Swiss-prot)
- Has lots of annotation like GO terms
- Has various links to database (SGD,KEGG,etc)

http://www.uniprot.org/

KEGG

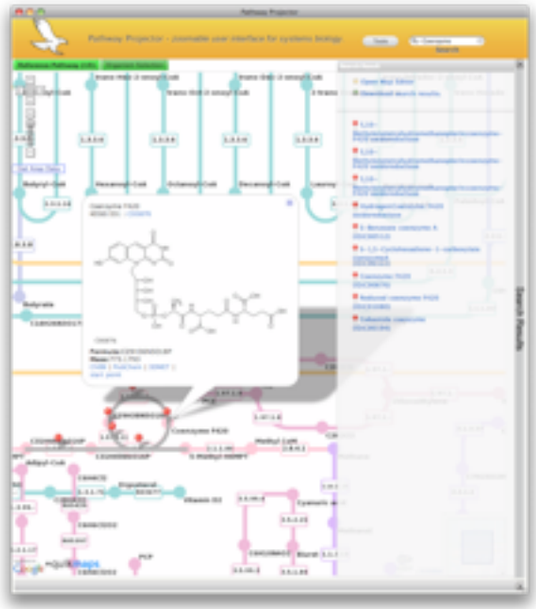
- Pathway database for various species
- Has metabolic pathways, signal pathways and so on
- Users can color, map object to the pathway and use analysis tools for the pathway
- you use it from your code by KEGG API



http://www.kegg.jp/kegg/kegg2.html

Pathway Projector

- Made in IAB, Keio univ. Now in press
  (this is where I spent my undergrad)
- Has global metabolic pathway map integrated from KEGG ,with google map interface
- User can color, map, draw objects on the map and use analysis tools for the pathway
- Please send us feedbacks if you have ANY suggestions or demands (please!)



http://www.g-language.org/PathwayProjector/

About perl and bioperl

maybe slides here would help you understand the codes,
but you can Ignore the slides to go on

bioperl

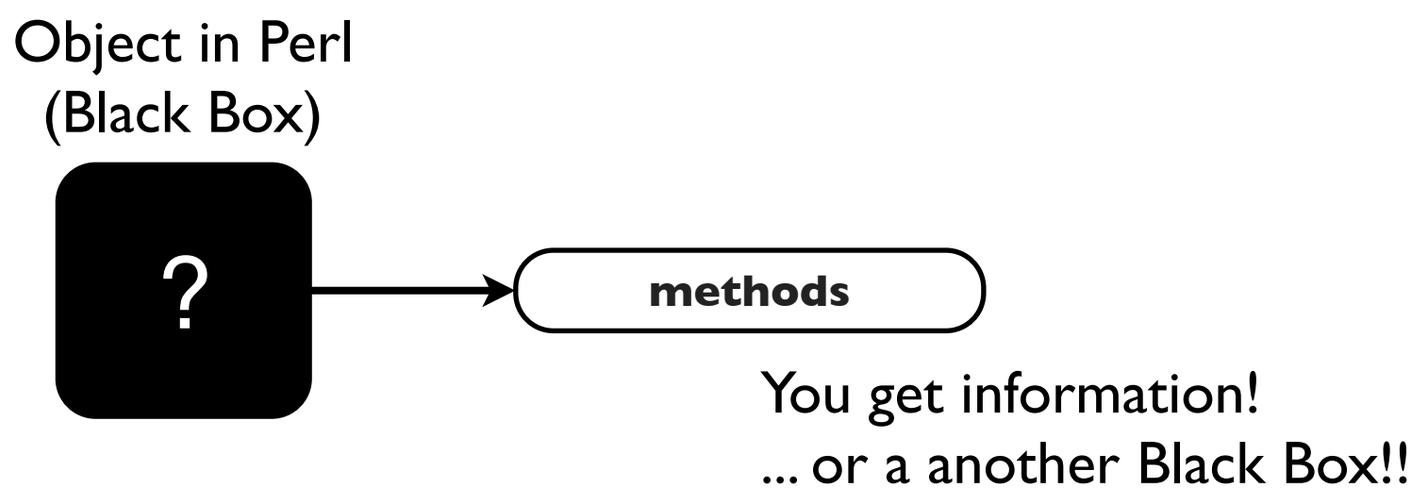
- bioperl is made with object-orientation

  ... with fastidious care

- So you will have to know what object-orientation is like to use bioperl

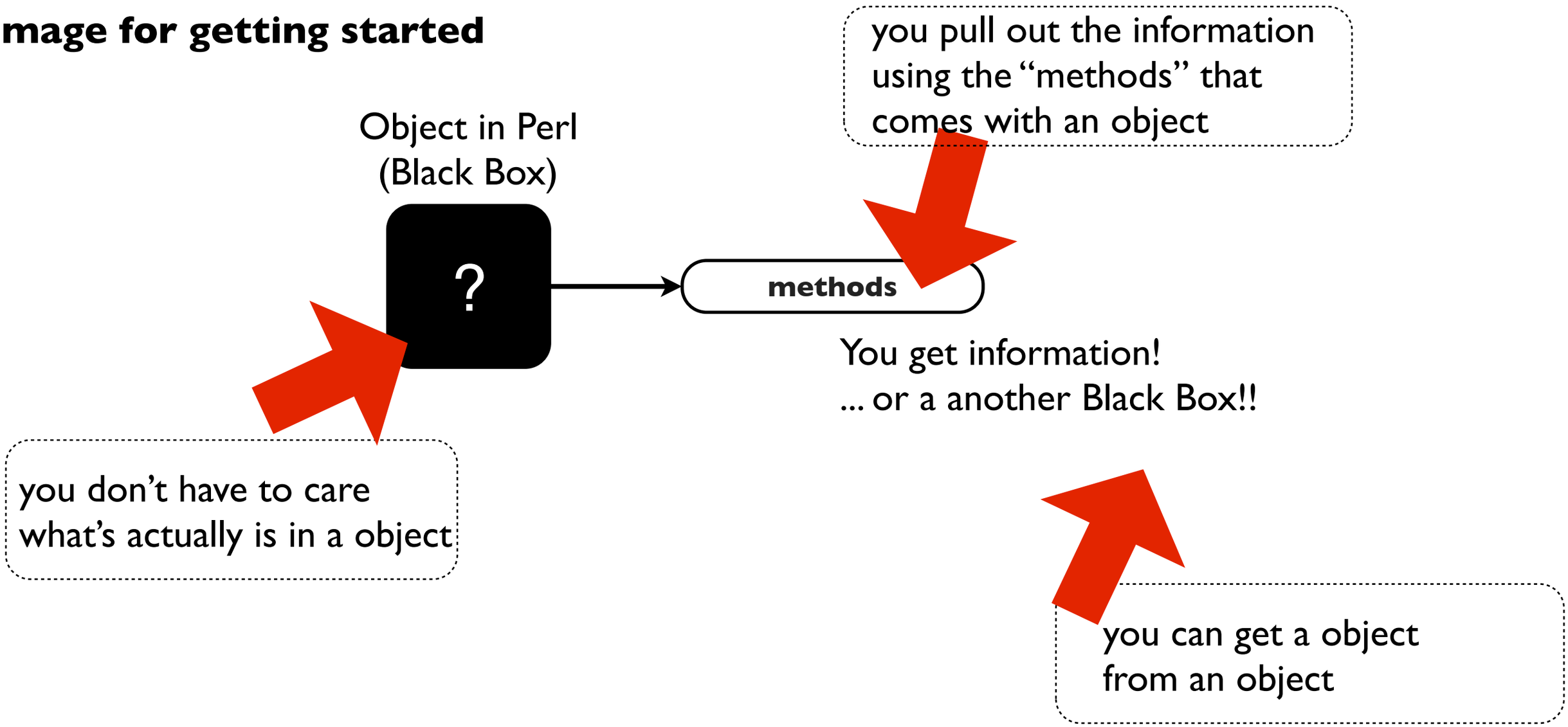
Super Easy guide to Object-orientation in perl

Image for getting started



Super Easy guide to Object-orientation in perl

Image for getting started



Example

$tom is the Person object

```
use Person;
my $tom = Person->new(-name => "Tom",-age=>"24",);

print $tom->name;
print $tom->age;
print $tom->age_in_seconds;
```

Example

$tom is the Person object

"use" lets you to use the OBJECT, "Person"

by "new" you create the OBJECT

some objects needs parameters to create

```
use Person;
my $tom = Person->new(-name => "Tom",-age=>"24",);

print $tom->name;
print $tom->age;
print $tom->age_in_seconds;
```

(probably) prints "Tom"

(probably) prints "24"

(probably) prints "756,864,000"

I wrote probably because it depends on how the methods are written

Example in bioperl

this script prints all the amino acid sequence in a genome
(this example is not accurate, but will make things easier to explain)

```
use Bio::SeqIO;
my $seqio_object = Bio::SeqIO->new(-file => "ecoli.gbk");
my $seq_object = $seqio_object->next_seq;

for my $feat_object ($seq_object->get_SeqFeatures) {
  print $feat_object->seq->translate->seq;
}
```

Examples in bioperl

Bio::SeqIO object

```
use Bio::SeqIO;
my $seqio_object = Bio::SeqIO->new(-file => "ecoli.gbk");
my $seq_object = $seqio_object->next_seq;
```

Bio::Seq object

```
for my $feat_object ($seq_object->get_SeqFeatures) {
  print $feat_object->seq->translate->seq;
}
```

Bio::SeqFeature::Generic object

Bio::PrimarySeq object (for nucleotide)

Bio::PrimarySeq object (for AA)

string of AA(not object)

you get the AA sequence printed out!!!

As a whole... bioperl!!!

I know this is super ultra hard for beginners to understand.
Sometimes, It is also driving me nuts.
But, bioperl is useful.

Here is the list of all the Classes in bioperl

http://search.cpan.org/~cjfields/BioPerl-1.6.1/

There is a useful copy & paste scripts around the internet.
In the bioperl wiki, there are helpful tutorials, HOWTOs, and Mailing lists where people can ask questions about bioperl.

http://www.bioperl.org/wiki/Main_Page

Implementation

How the program works

for each entry in Uniprot
    make the SGD and Uniprot ID relation hash table

Step1

for each SNPs
    for each CDS check if SNP is in the CDS region
        if so, check if SNP changes the AA
            if so, get UniProt id by using the hash table in step1
                check if mutation is in the feature tables
                get bunch of annotations
Print the result

Step2

Outputs

- Protein_List.txt
  tells you what protein had SNPs
  (GO terms, protein definition)
- Protein_details.txt
  tells you more info in to the domain (FT)

Protein_List.txt (maybe better looking with Excel)

```
YNL176C S000005120      YNR6_YEAST      sce:YNL176C     GO:0000324,GO:0016021   RecName: Full=Uncharacterized protein YNL176C;
PKH1    S000002898      PKH1_YEAST      sce:YDR490C     GO:0005829,GO:0005524,GO:0005515,GO:0004674,GO:0006897,GO:0000196,GO:0006468    RecName: Full=Serine/threonine-pro
YPR097W S000006301      YP097_YEAST     sce:YPR097W     GO:0019898,GO:0005739,GO:0035091,GO:0005515,GO:0007154  RecName: Full=PX domain-containing protein YPR097W;
UFD2    S000002349      UFD2_YEAST      sce:YDL190C     GO:0005737,GO:0005634,GO:0000151,GO:0005515,GO:0034450,GO:0016567,GO:0006950,GO:0006511 RecName: Full=Ubiquitin co
RAD4    S000000964      RAD4_YEAST      sce:YER162C     GO:0000111,GO:0000108,GO:0003684,GO:0005515,GO:0000715  RecName: Full=DNA repair protein RAD4;
PKC1    S000000201      KPC1_YEAST      sce:YBL105C     GO:0005737,GO:0005856,GO:0005634,GO:0001950,GO:0030427,GO:0005524,GO:0005509,GO:0019992,GO:0005515,GO:0004697,GO:0
POS5    S000006109      POS5_YEAST      sce:YPL188W     GO:0005759,GO:0005524,GO:0003951,GO:0042736,GO:0006741,GO:0006979       RecName: Full=NADH kinase POS5, mitochondr
ATP1    S000000195      ATPA_YEAST      sce:YBL099W     GO:0042645,GO:0005754,GO:0005524,GO:0046933,GO:0046961,GO:0015986       RecName: Full=ATP synthase subunit alpha,
YNL193W S000005137      YNT3_YEAST      sce:YNL193W             RecName: Full=Uncharacterized protein YNL193W;
OAF1    S000000048      OAF1_YEAST      sce:YAL051W     GO:0005634,GO:0005515,GO:0043565,GO:0016563,GO:0003700,GO:0008270,GO:0006631,GO:0016481,GO:0007031,GO:0045941,GO:0
MCH2    S000001704      MCH2_YEAST      sce:YKL221W     GO:0016021,GO:0015293,GO:0006810        RecName: Full=Probable transporter MCH2;
VPS15   S000000301      VPS15_YEAST     sce:YBR097W     GO:0031225,GO:0005768,GO:0000139,GO:0005739,GO:0034271,GO:0034272,GO:0005524,GO:0005515,GO:0004674,GO:0048017,GO:0
THI22   S000006325      THI22_YEAST     sce:YPR121W     GO:0005576,GO:0009228   RecName: Full=Thiamine biosynthesis protein THI22; Flags: Precursor;
SRO77   S000000202      SNI2_YEAST      sce:YBL106C     GO:0005886,GO:0006887,GO:0006893        RecName: Full=Protein SNI2; AltName: Full=Suppressor of RHO3 protein 77;
YMR027W S000004629      YMR7_YEAST      sce:YMR027W     GO:0005737,GO:0005634   RecName: Full=UPF0364 protein YMR027W;
RTC1    S000005498      YO128_YEAST     sce:YOL138C     GO:0000324,GO:0005515,GO:0008270        RecName: Full=Uncharacterized WD repeat-containing protein YOL138C;
NAB6    S000004585      NAB6_YEAST      sce:YML117W     GO:0005737,GO:0003723   RecName: Full=RNA-binding protein NAB6; AltName: Full=Nucleic acid-binding protein 6;
MKK1    S000005757      MKK1_YEAST      sce:YOR231W     GO:0005934,GO:0005524,GO:0004674,GO:0006468,GO:0007165  RecName: Full=MAP kinase kinase MKK1/SSP32; EC=2.7.12.2;
ORT1    S000005656      ORT1_YEAST      sce:YOR130C     GO:0016021,GO:0005743,GO:0005488,GO:0000064,GO:0006526,GO:0000066       RecName: Full=Mitochondrial ornithine carr
SST2    S000004444      SST2_YEAST      sce:YLR452C     GO:0005886,GO:0005096,GO:0005515,GO:0004871,GO:0000754,GO:0007242       RecName: Full=Protein SST2;
PAT1    S000000673      PAT1_YEAST      sce:YCR077C     GO:0000932,GO:0022627,GO:0005515,GO:0007049,GO:0051301,GO:0007059,GO:0033962,GO:0000290,GO:0006446      RecName: F
PDR3    S000000101      PDR3_YEAST      sce:YBL005W     GO:0005737,GO:0005634,GO:0003704,GO:0016563,GO:0003700,GO:0016564,GO:0008270,GO:0000122,GO:0045944,GO:0042493,GO:0
YET1    S000001548      YET1_YEAST      sce:YKL065C     GO:0005783,GO:0016021,GO:0005515,GO:0006886,GO:0016192  RecName: Full=Endoplasmic reticulum transmembrane protein
PSK1    S000000015      KAB7_YEAST      sce:YAL017W     GO:0005737,GO:0005524,GO:0042802,GO:0004674,GO:0004871,GO:0006078,GO:0019318,GO:0006468,GO:0007165      RecName: F
CTR9    S000005505      CTR9_YEAST      sce:YOL145C     GO:0016593,GO:0016944,GO:0045142,GO:0007059,GO:0016571,GO:0045449,GO:0006368    RecName: Full=RNA polymerase-assoc
DPB2    S000006379      DPB2_YEAST      sce:YPR175W     GO:0005737,GO:0008622,GO:0003677,GO:0003887,GO:0007049,GO:0006273,GO:0006272,GO:0006298,GO:0006289      RecName: F
PPX1    S000001244      PPX1_YEAST      sce:YHR201C     GO:0005829,GO:0005759,GO:0005886,GO:0004309,GO:0030145,GO:0006798       RecName: Full=Exopolyphosphatase; Short=Ex
KRE33   S000005076      YNN2_YEAST      sce:YNL132W     GO:0030686,GO:0005730,GO:0005524,GO:0042274     RecName: Full=UPF0202 protein YNL132W;
GLN1    S000006239      GLNA_YEAST      sce:YPR035W     GO:0005737,GO:0005524,GO:0004356,GO:0005515,GO:0006542  RecName: Full=Glutamine synthetase; Short=GS; EC=6.3.1.2;
YCR024C-B       S000028818      YC204_YEAST     sce:YCR024C-B   GO:0016021      RecName: Full=Uncharacterized protein YCR024C-B;
SLH1    S000003503      SLH1_YEAST      sce:YGR271W     GO:0005737,GO:0005524,GO:0008026,GO:0003676,GO:0005515,GO:0006417,GO:0009615    RecName: Full=Antiviral helicase S
YMR185W S000004797      YM48_YEAST      sce:YMR185W     GO:0005515      RecName: Full=Uncharacterized protein YMR185W;
MMT2    S000006145      MMT2_YEAST      sce:YPL224C     GO:0016021,GO:0005739,GO:0008324,GO:0005506,GO:0006879,GO:0006826       RecName: Full=Mitochondrial metal transpor
FLC2    S000000049      FLC2_YEAST      sce:YAL053W     GO:0005783,GO:0016021,GO:0015230,GO:0005515,GO:0015883,GO:0009272,GO:0006457,GO:0055085 RecName: Full=Flavin carri
YBR204C S000000408      YB54_YEAST      sce:YBR204C     GO:0005777,GO:0042802,GO:0017171,GO:0016042     RecName: Full=Putative peroxisomal lipase YBR204C; EC=3.1.1.-;
BRN1    S000000193      CND2_YEAST      sce:YBL097W     GO:0005737,GO:0000799,GO:0005515,GO:0051301,GO:0007076,GO:0070058       RecName: Full=Condensin complex subunit 2;
DYN3    S000004914      DYN3_YEAST              GO:0005737,GO:0005868,GO:0005881,GO:0003774,GO:0005515,GO:0030473       RecName: Full=Cytoplasmic dynein intermediate ligh
YIL083C S000001345      YII3_YEAST      sce:YIL083C     GO:0005737,GO:0005634,GO:0005515,GO:0015937     RecName: Full=Uncharacterized protein YIL083C;
MNN1    S000000803      MNN1_YEAST      sce:YER001W     GO:0005794,GO:0016021,GO:0000033,GO:0006491,GO:0006493  RecName: Full=Alpha-1,3-mannosyltransferase MNN1; EC=2.4.1
ALD5    S000000875      ALDH5_YEAST     sce:YER073W     GO:0005759,GO:0004029,GO:0033721,GO:0004030,GO:0005515,GO:0019413,GO:0055114    RecName: Full=Aldehyde dehydrogena
TIM18   S000005823      TIM18_YEAST     sce:YOR297C     GO:0016021,GO:0042721,GO:0020037,GO:0008565,GO:0006915,GO:0006612,GO:0006626,GO:0001101,GO:0046685,GO:0006970,GO:0
RPL39   S000003725      RL39_YEAST      sce:YJL189W     GO:0022625,GO:0003735,GO:0006412        RecName: Full=60S ribosomal protein L39; AltName: Full=L46; AltName: Full=
BAP2    S000000272      BAP2_YEAST      sce:YBR068C     GO:0016021,GO:0015171,GO:0005515,GO:0006865     RecName: Full=Leu/Val/Ile amino-acid permease; AltName: Full=Branc
TRZ1    S000001787      RNZ_YEAST       sce:YKR079C     GO:0005739,GO:0005634,GO:0042781,GO:0008270,GO:0034414  RecName: Full=Ribonuclease Z; Short=RNase Z; EC=3.1.26.11;
UBP13   S000000163      UBP13_YEAST     sce:YBL067C     GO:0004221,GO:0004843,GO:0006511        RecName: Full=Ubiquitin carboxyl-terminal hydrolase 13; EC=3.1.2.15; AltNa
SPC72   S000000045      YAE7_YEAST      sce:YAL047C     GO:0005824,GO:0005200,GO:0007020,GO:0031578,GO:0000070,GO:0000022,GO:0030473    RecName: Full=Uncharacterized prot
ERG7    S000001114      ERG7_YEAST      sce:YHR072W     GO:0005783,GO:0019898,GO:0012511,GO:0005886,GO:0042802,GO:0000250,GO:0006694    RecName: Full=Lanosterol synthase;
FTH1    S000000411      FTH1_YEAST      sce:YBR207W     GO:0000329,GO:0016021,GO:0042802,GO:0005506,GO:0005381,GO:0006897,GO:0006827    RecName: Full=Iron transporter FTH
```

Protein_details.txt

01	41240	GPB2	S000000052	A	G	GPB2_YEAST	661	I	V	CONFLICT:I -> V (in Ref. 1; BAD04042)
01	46833	FLC2	S000000049	G	A	FLC2_YEAST	312	D	N	TOPO_DOM:Lumenal (Potential)
01	48772	OAF1	S000000048	T	A	OAF1_YEAST	70	W	R	DNA_BIND:Zn(2)-C6 fungal-type
01	50327	OAF1	S000000048	C	A	OAF1_YEAST	588	T	K	
01	55954	SPC72	S000000045	A	T	YAE7_YEAST	302	I	N	
01	120442	PSK1	S000000015	C	G	KAB7_YEAST	73	Q	E	
01	167551	BUD14	S000000069	G	C	BUD14_YEAST	439	A	G	
01	179761	YAR023C	S000000074	C	A	YAJ3_YEAST	20	V	F	couldn't see details, becuase AA in SGD & UniProt is different
02	11379	SRO77	S000000202	A	C	SNI2_YEAST	834	V	G	REPEAT:WD 14
02	13102	SRO77	S000000202	C	A	SNI2_YEAST	260	A	S	REPEAT:WD 5
02	13477	SRO77	S000000202	G	A	SNI2_YEAST	135	P	S	REPEAT:WD 3
02	13492	SRO77	S000000202	A	T	SNI2_YEAST	130	F	I	REPEAT:WD 3
02	15101	PKC1	S000000201	T	G	KPC1_YEAST	866	I	L	DOMAIN:Protein kinase
02	15835	PKC1	S000000201	T	C	KPC1_YEAST	621	K	R	
02	38067	ATP1	S000000195	C	T	ATPA_YEAST	340	P	S	couldn't see details, becuase AA in SGD & UniProt is different
02	42377	BRN1	S000000193	C	G	CND2_YEAST	517	A	G	CONFLICT:A -> G (in Ref. 3; AAS56403)
02	55145	TEL1	S000000184	A	C	ATM_YEAST	1412	F	C	
02	95345	UBP13	S000000163	G	T	UBP13_YEAST	180	H	Q	
02	217890	PDR3	S000000101	C	A	PDR3_YEAST	140	Q	K	
02	375081	BAP2	S000000272	C	A	BAP2_YEAST	203	G	W	TOPO_DOM:Extracellular (Potential)
02	375272	BAP2	S000000272	T	A	BAP2_YEAST	139	E	V	TRANSMEM:Potential
02	437344	VPS15	S000000301	A	G	VPS15_YEAST	134	T	A	DOMAIN:Protein kinase, CONFLICT:T -> A (in Ref. 1; AAA35214)
02	439496	VPS15	S000000301	T	G	VPS15_YEAST	851	I	R	
02	633189	YBR204C	S000000408	T	A	YB54_YEAST	63	E	V	
02	635192	FTH1	S000000411	A	G	FTH1_YEAST	18	K	E	TRANSMEM:Potential, CONFLICT:K -> E (in Ref. 1; AAD53168)
02	635247	FTH1	S000000411	A	G	FTH1_YEAST	36	D	G	TOPO_DOM:Cytoplasmic (Potential), CONFLICT:D -> G (in Ref. 1; AAD53168)
03	162636	YCR024C-B	S000028818	T	G	YC204_YEAST	76	S	R	
03	250563	PAT1	S000000673	A	T	PAT1_YEAST	688	V	D	
03	275421	KIN82	S000000687	A	G	KIN82_YEAST	341	M	V	DOMAIN:Protein kinase
04	121289	UFD2	S000002349	G	A	UFD2_YEAST	102	S	L	CONFLICT:S -> L (in Ref. 1; AAC49024)
04	1296177	ERD1	S000002822	C	G	ERD1_YEAST	168	G	A	TOPO_DOM:Cytoplasmic (Potential), CONFLICT:G -> A (in Ref. 1; CAA36211)
04	1433703	PKH1	S000002898	A	T	PKH1_YEAST	187	F	I	DOMAIN:Protein kinase
04	1519663	YDR541C	S000002949	C	G	YD541_YEAST	341	E	Q	
05	154530	MNN1	S000000803	T	A	MNN1_YEAST	338	S	T	TOPO_DOM:Lumenal (Potential), CONFLICT:S -> T (in Ref. 1; AAA53676)
05	305258	ALD5	S000000875	G	A	ALDH5_YEAST	411	G	E	CONFLICT:G -> E (in Ref. 1; AAB01220)
05	502222	RAD4	S000000964	T	A	RAD4_YEAST	223	E	V	couldn't see details, becuase AA in SGD & UniProt is different
06	66443	YFL034W	S000001860	C	A	YFD4_YEAST	323	N	K	TOPO_DOM:Cytoplasmic (Potential)
07	607109	PEF1	S000003290	T	G	CPNSH_YEAST	324	Y	D	DOMAIN:EF-hand 3
07	622408	YGR067C	S000003299	A	T	YG2A_YEAST	795	*	K	
07	783805	ENP2	S000003377	A	C	NOL10_YEAST	678	E	D	
07	1031948	SLH1	S000003503	C	A	SLH1_YEAST	51	P	Q	
07	1033108	SLH1	S000003503	C	T	SLH1_YEAST	438	P	S	DOMAIN:Helicase ATP-binding 1
08	240687	ERG7	S000001114	G	A	ERG7_YEAST	530	D	N	couldn't see details, becuase AA in SGD & UniProt is different
08	499958	PPX1	S000001244	C	T	PPX1_YEAST	396	E	K	
09	203638	YIL083C	S000001345	A	C	YII3_YEAST	338	I	S	
09	318692	VID28	S000001279	T	A	VID28_YEAST	758	N	Y	
09	330114	PDR11	S000001275	C	A	PDR11_YEAST	776	G	V	TOPO_DOM:Cytoplasmic (Potential), DOMAIN:ABC transporter 2
10	76241	RPL39	S000003725	T	C	RL39_YEAST	104	Y	H	couldn't see details, becuase AA in SGD & UniProt is different
11	7131	MCH2	S000001704	C	G	MCH2_YEAST	342	R	G	TOPO_DOM:Extracellular (Potential)

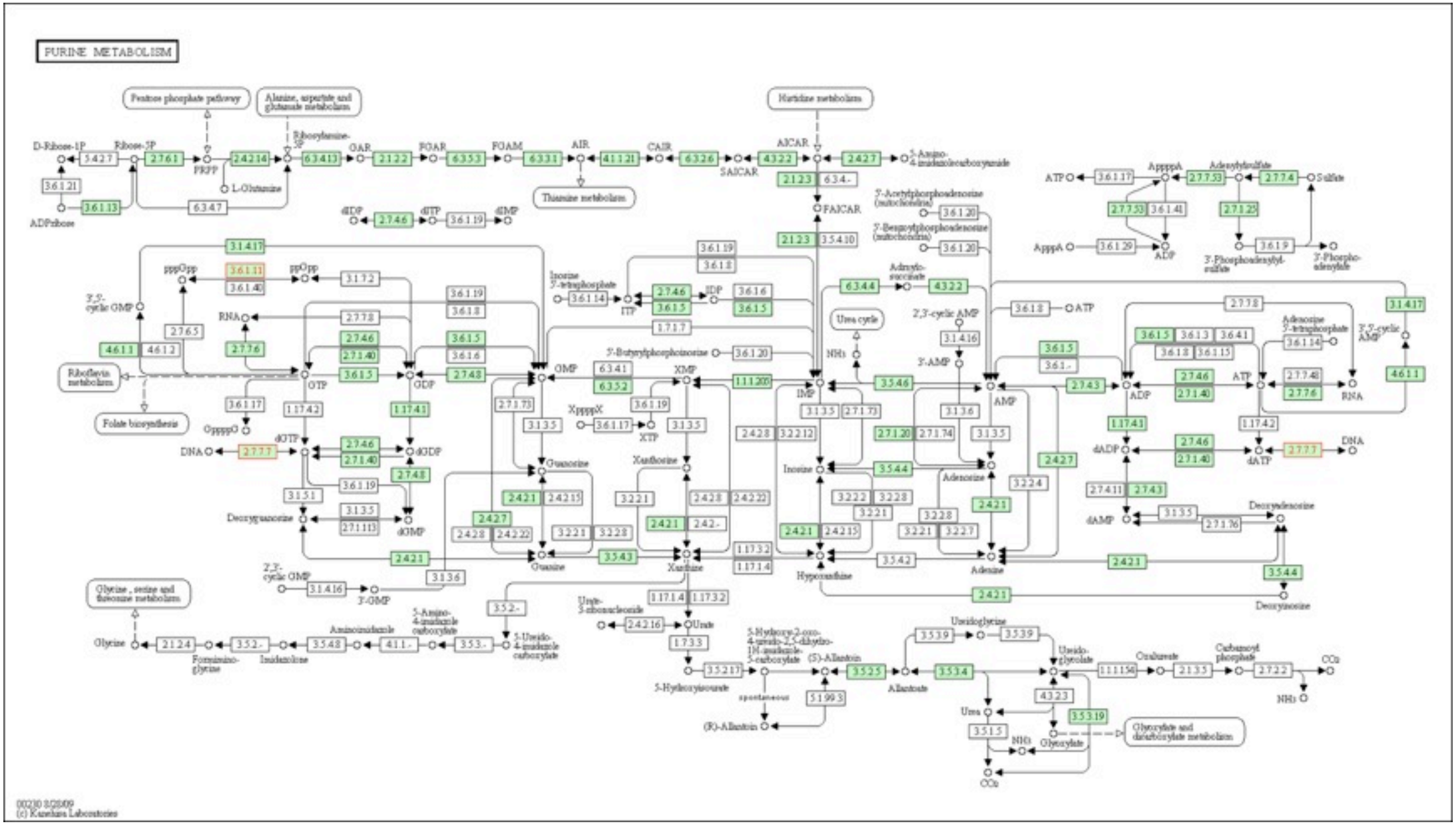
Additional scripts

- map_protein_to_kegg_in_png.pl
  map the result of Protein_List.txt to png images in KEGG

- map_protein_to_kegg_in_html.pl
  map the result of Protein_List.txt to html images in KEGG

- input_for_pp.pl
  create an input file for Pathway Projector from Protein_List

map_protein_to_kegg_in_html.pl
map_protein_to_kegg_in_png.pl



enzymes in green is the genes in yeast, frame with red is the mutated protein

Pathway Projector - Vimperator

Google リーダー (1000+) | YouTube - Phoenix - Lisztoma... | ftp://ftp.genome.jp/pub/kegg/... | Pathway Projector

Pathway Projector - *zoomable user interface for systems biology.*

Tools | ALD5

Reference Pathway | Organism Selection | **Saccharomyces cerevisiae S288C(9)**

Gene: *ALD5*
MIPS|NCBI-GI|NCBI-GeneID|PROSITE|Pfam|SGD|UniProt
Gene: *ALD4*
MIPS|NCBI-GI|NCBI-GeneID|PROSITE|Pfam|SGD|UniProt
Gene: *ALD6*
MIPS|NCBI-GI|NCBI-GeneID|PROSITE|Pfam|SGD|UniProt

ALD5 ALD4 ALD6

Orthology: K00128
Enzyme: 1.2.1.3
Expasy|MetaCyc|Brenda|IntEnz|PUMA2|IUBMB

Acetyl-CoA
Glutaryl-CoA
Homocitrate
Homoisocitrate
2-Oxoadipate
HMG-CoA
3-Oxopropanoate
Propionyl-CoA
Propionate
C6H11NO3
C6H11NO4
Ethylene oxide
Oxaloacetate
Citrate
Neuberg ester
Malate
Saccharopine
L-Lysine
Fumarate
Succinate
Succinyl-CoA
2-Oxoglutarate
Glyoxalate
10-Formyl-THF
Oxalosuccinate
2-Chloroethanal
POWERED BY Google+quikmaps

< http://ws.g-language.org/g4/main.cgi?diaAtabareaname=1&flag=1256927213 [+]

HOW to use

An actual example

Data Download

- Get the *Yeast* Solexa read in NCBI SRA

  from : http://www.ncbi.nlm.nih.gov/sites/entrez?db=sra&term=SRX003233&report=full

- Get the fungi swiss-prot data in UniProt

  from : ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/

- Get the Genome/FASTA data of yeast in SGD

  from : http://downloads.yeastgenome.org/sequence/NCBI_genome_source/

prepare Data for MAQ

- From the data you got from SGD, Make a big FASTA that contains all chromosomes
- FASTQ in SRA has little trouble with running MAQ, so we'll have to fix the format

I wrote a simple script "fix_format_SRA.pl" that does that. You only have to get rid of the space in the read_name in FASTQ file.

Post about this topic in seqanswers

http://seqanswers.com/forums/showthread.php?t=1488&highlight=short+read+archive

RUN the program!

- Do MAQ against the SRA data and SGD genome
- Run perl scripts

These are the things you have to do on shell

```
> ssh [ your user name] @cl2.msi.umn.edu
```

login to MSI workstation

```
> cd [ your working directory]
> mkdir lib
> cd lib
> cp [all the perl scripts I wrote]
> cd ../
```

creating the working directory

```
> mkdir -p data/SRA
> mkdir -p data/SGD
> mkdir -p data/UniProt

> cd data/SRA
> wget ftp://ftp.ncbi.nlm.nih.gov/sra/static/SRX003/SRX003233/SRR014437.fastq.gz
> gunzip SRR014437.fastq.gz
```

download the SRA

```
> cd ../UniProt
> wget ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/
uniprot_sprot_fungi.dat.gz
> gunzip uniprot_sprot_fungi.dat.gz
```

download the UniProt

```
> cd ../SGD
> wget --level=1 -A gbf,fsa -r -nd http://downloads.yeastgenome.org/sequence/NCBI_genome_source/
```

download the SGD

```
> cat chr01.fsa chr02.fsa chr03.fsa chr04.fsa chr05.fsa chr06.fsa chr07.fsa chr08.fsa chr09.fsa chr10.fsa
chr11.fsa chr12.fsa chr13.fsa chr14.fsa chr15.fsa chr16.fsa > all.fsa
> cd ../../lib
> perl5.10.0 fix_format_SRA.pl ../data/SRA/SRR014437.fastq > ../data/SRA/fixed_ SRR014437.fastq
```

prepare for MAQ

```
> module load bioinformatics
> maq.pl easyrun -d maqout data/SGD/all.fasta data/SRA/fixed_SRR014437.fastq
```

run MAQ

```
> perl5.10.0 maq_to_genbank_and_uniprot.pl --up=../data/Uniprot/uniprot_sprot_fungi.dat --sgd_dir=../
data/SGD --maq=../data/maqout/cns.final.snp
```

running the scripts

how to run the additional scripts

```
> cd [your working dir]
> mkdir -p data/KEGG
> cd data/KEGG
> wget ftp://ftp.genome.jp/pub/kegg/genes/organisms/sce/S.cerevisiae.ent
> wget ftp://ftp.genome.jp/pub/kegg/genes/organisms/sce/sce_sgd-sce.list
> cd ../../lib
> perl5.10.0 map_protein_to_kegg_in_png.pl Protein_List.txt
> perl5.10.0 map_protein_to_kegg_in_png.pl Protein_List.txt
> perl5.10.0 input_data_for_pp.pl Protein_List.txt
```

Very important notes

- `module load bioinformatics`
  be sure to do this before using the bioinformatics software

- `perl5.10.0`
  if you are working on the MSI workstation, use "perl5.10.0" instead of "perl" !
  there is a bug in bioperl version 1.4 with perl 5.8.8

  you can always check the version of bioperl by typing the following

  > perl -MBio::Root::Version -e 'print $Bio::Root::Version::VERSION,"\n"'

Discussion

- Needs more improvement in detecting SNP positions
  (Just running the MAQ in the default parameter)
- Maybe, there is a better software that does these kinds of things

Future Works

- Look at SNPs out of the CDS region, like TF binding sites. And look at their regulation network
- Look for database that has information of known SNPs
- use GO slim to get the overview of protein groups. You can do this online , or map2slim have to be installed in the MSI servers
- See the rate of mapped reads between the wild type and mutant to find gene duplication, deletion

Conclusion

- Creating new Mapping/Alignment software for nextgen sequencing is a tough work to do
- Although there are numbers of useful tools,
- And we don’t want to re-invent the wheels
- Scripts here will show you what kinds of information you can get by combing the existing software/database
- I don’t know if this script is actually useful, but I am happy if you could get any ideas from this.

Thank you

- Feel free to contact me at any time, for anything

Satoshi Tamaki

coela.st@gmail.com

NARA INSTITUTE of SCIENCE and TECHNOLOGY
Comparative Genomics Lab