SGD - Phenotype data: Create Mappings collection, triples and publications

This activity is to create mappings collection documents for the SGD file :

phenotype_data.tab.txt

location source data file: /mnt/2TB_backup/datastraat/inputdata/sgd

The file is a tab separated file, with 1 entry per line. In total 134038 records in the file

Example:

MAL62 not in systematic sequence of S288C MAL62 S000029690 PMID: 22669197|SGD_REF: S000149697 classical genetics maltose overexpression Other fermentative growth: increased maltose fementation and leavening ability are enhanced

An entry contains the following contents, to be mapped to internal keys:

the combination **gene/sgdreference/phenotype/chemical** is unique key per document

	Source key	description	type (source)	example	Mappings collection Target key	example target
1	Feature name	gene name	mandatory	MAL62	gene	mal62
2	Feature type	type of gene	mandatory	not in systematic sequence of S288C	geneType	not in systematic sequence of S288C
3	Gene Name	gene name	optional	MAL62	geneName	mal62
4	SGDID	database ID	mandatory	S000029690	sgd_id	S000029690
5	Reference	pipe separated ref	mandatory; split into 2	PMID: 22669197|SGD	sgdReference	S000149697

		ids	keys in mappings	_REF: S000149697		
					pmidReference	22669197
6	experiment type	method used to detect phenotype	mandatory	classical genetics	experiment	classical genetics
7	mutant type	description of impact of mutation on the activity	mandatory	overexpression	mutation	overexpression
8	allele	allele description	optional	systematic mutation set	allele	systematic mutation set
9	strain	strain used in experiment	optional	S288C	strain	s288c
10	phenotype	observed feature and direction separated by “:”	mandatory; split into 2 keys in mappings	fermentative growth: increased	phenotype	fermentative growth
			optional		direction	increased
11	chemical	contextual chemical info	optional	maltose	chemical	maltose
12	condition	condition of experiment	optional		condition	
13	details		optional	maltose fementation and leavening ability are enhanced	details	maltose fementation and leavening ability are enhanced
14	reporter	protein or RNA used in experiment	optional		reporter	
15	date			date of file	date	

Triple generation

The source generates the 5 types of triples:

1. **<gene> - predicate - <phenotype>**

Measures:
experiment type: <experiment>
mutant type: <mutation>
allele: <allele>
condition: <condition>

Determine subject:

solr query- **pterm**: <gene> && taxonomies:"saccharomyces cerevisiae s288c"

If Solr returns no results, perform the following search:

Solr query: **pterm**: <geneName> && taxonomies:"saccharomyces cerevisiae s288c"

for example


```
{
    "id": "e233598f-c1ec-497b-8b36-e3ac5c15f2eb",
    "term": "yil165c (saccharomyces cerevisiae s288c)",
    "pterm": "yil165c (saccharomyces cerevisiae s288c)",
    "source": "entrezgene",
    "knowledgebase": "entrezgene",
    "semantictype": "28",
    "semanticcategory": "Genes & Molecular Sequences",
    "taxonomies": [
      "saccharomyces cerevisiae s288c"
    ],
    "uuid": "f8487be8-8b1a-45a3-99a1-d8012ca57ee9",
    "preferred": true,
    "_version_": 1514328750225883100
```


If Solr returns no UUID, move to the next document in the mapping collection.

Determine object:
solr query term: <phenotype>

If Solr returns no response, log the query and proceed to the following document in the mappings collection.

If more than 1 GI/UUID is returned, determine which GI contains entries with source: “SGD” with another synonym.

For example <phenotype> = overexpression ; the uuid is linked to a record (term"ypo:0000008”) from source: sgd. This is a match.

If more than 1 UUIDs/GI match, create a triple for each of the matching UUIDs/GI’s


```
{
    "id": "03a10191-e564-4010-a056-adcfc60e42e1",
    "term": "overexpression",
    "pterm": "overexpression",
    "source": "umls",
    "knowledgebase": "nci",
    "semantictype": "45",
    "semanticcategory": "Physiology",
    "uuid": "08f11c57-7772-420d-8e37-1d4b15955083",
    "preferred": false,
    "_version_": 1497363724014452700
  }
  {
    "id": "10170a74-1d5b-4686-88f4-1cfd620e1048",
    "term": "protein overexpression",
    "pterm": "protein overexpression",
    "source": "umls",
    "knowledgebase": "nci",
    "semantictype": "45",
    "semanticcategory": "Physiology",
    "uuid": "08f11c57-7772-420d-8e37-1d4b15955083",
    "preferred": true,
    "_version_": 1497363724297568300
  },
  {
    "id": "b7fc9ae3-dd30-4948-b791-9d13e5d645ed",
    "term": "ypo:0000008",
    "pterm": "ypo:0000008",
    "source": "sgd",
    "knowledgebase": "ascomycete phenotype ontology",
    "semantictype": "45",
    "semanticcategory": "Physiology",
    "uuid": "08f11c57-7772-420d-8e37-1d4b15955083",
    "preferred": false,
```

```
  "_version_": 1505505184067879000
},
```


Determine Predicate

Note that some ofthe predicates are new ; use the latest predicate tree file in /datastraatinput/predicates/september2015/

Based on the value of <direction>, a predicate is selected.

<gene> - variant_increases - <phenotype> | if <direction> = {increased, increased rate}

<gene> - variant_decreases - <phenotype> | if <direction> = {decreased, decreased rate, normal rate}

<gene> - variant_inhibits - <phenotype> | if <direction> = {absent, arrested}

<gene> - variant_results_in_abnormal- <phenotype> | if <direction>={abnormal, delayed, decreased duration, increased duration, premature}

<gene> - neg_variant_results_in_abnormal - <phenotype> | if <direction>= {normal}

If <direction> is not present / empty, the value of <phenotype> determines the predicate:

<gene> - gene_product_malfunction_associated_with - <phenotype> | if <phenotype>={viable, haploproficient, auxotrophy, cell cycle progression through the G2/M phase, petite-negative, petite, sterile}

2. <gene> - functionally_related_to - <chemical>

Determine subject:

solr query- **pterm**: <gene> && taxonomies:"saccharomyces cerevisiae s288c"

If Solr returns no results, perform the following search:

Solr query: **pterm**: <geneName> && taxonomies:"saccharomyces cerevisiae s288c"

For each of the Solr results, generate a triple.

Determine object

The <chemical> entry needs to be pre-processed / cleaned before querying because the value may contain a measure.

Any part of the string starting with value “[white space](“ needs to be deleted

examples

<chemical> value	<chemical solr query string>
sodium chloride (0.9 M)	sodium chloride
sodium chloride (0.9 M)	cycloheximide
nickel(2+) (3 mM NiCl2)	nickel(2+)
S-{2-[4-(dihydroxyarsino)phenylamino]-2-oxo ethyl}-glutathione	S-{2-[4-(dihydroxyarsino)phenylamino]-2-oxo ethyl}-glutathione

In addition, some <chemical> entries are ambiguous and need mapping to an ID, according to the following table. If the <chemical> has an entry in table below, use <mappedChemical>, ELSE use <chemical solr query string> in the Solr query.

<chemical solr query string>	<mappedChemical>
triglyceride	[chebi]17855
chloride	[chebi]17996
diglyceride	[chebi]18035
ceramide	[chebi]17761
L-tryptophan	[chebi]16828
amp	[chebi]16027

Solr query: term: <chemical solr query string>|<mappedChemical> && semanticcategory: “chemicals and drugs”
If Solr returns no results, skip this record and proceed to the next document in the mapping collection.
For each of the Solr results, generate a triple.

3. <chemical> - affects - <phenotype>

Resolve subject : see above <chemical>
Resolve object : see above <phenotype>

4. <gene> - predicate - <mutation>
Measures: <allele>

Determine subject:

solr query- **pterm**: <gene> && taxonomies:"saccharomyces cerevisiae s288c"

If Solr returns no results, perform the following search:

Solr query: **pterm**: <geneName> && taxonomies:"saccharomyces cerevisiae s288c"

Resolve object:

Only create a triple and perform solr query if

<mutation>= {overexpression, conditional, reduction of function, repressible, activation, gain of function, dominant negative, misexpression}

mutation <value>	solr query term	predicate
null	-	*donot create triple*
overexpression	APO:0000008	gene_product_has_abnormality
conditional	APO:0000014	gene_product_has_abnormality
reduction of function	APO:0000013	gene_product_hs_abnormality

unspecified	-	*donot create triple*
repressible	APO:0000012	gene_product_has_abnormality
activation	APO:0000009	gene_product_has_abnormality
gain of function	APO:0000010	gene_product_has_abnormality
dominant negative	APO:0000228	gene_product_has_abnormality
misexpression	APO:0000007	gene_product_has_abnormality

If Solr returns no results, skip this triple and proceed to the next document in the mapping collection.
For each of the Solr results, generate a triple.

5. <mutation> - predicate - <phenotype>

Measure: {<gene>}

Resolve subject (and predicate):

mutation <value>	solr query term	predicate
null	-	*donot create triple*
overexpression	APO:0000008	associated_with
conditional	APO:0000014	associated_with
reduction of function	APO:0000013	associated_with
unspecified	-	*donot create triple*
repressible	APO:0000012	associated_with
activation	APO:0000009	associated_with
gain of function	APO:0000010	associated_with
dominant negative	APO:0000228	associated_with
misexpression	APO:0000007	associated_with

Resolve <phenotype> : see above

Publication generation

If a mapping collection document contains a Pubmed reference (PMID), the Triple 1 (<gene> - predicate - <phenotype> will be added as reference to the existing pubmed publication record.

In addition, a separate publication is created containing all triples in a mapping record, referring to the source SGD:

Generate abstract:

title: <gene>-<phenotype>->chemical>
abstact text:

“Experiment:” <experiment>
“Condition:” <condition>
“Chemical:“ <chemical>
“Details” <details>
“Gene variation details:” <gene>, <mutation>, <allele>
“Reporter” : <reporter>
“Strain:”<strain>
“Phenotype:” <phenotype>, <direction>

“SGD Reference:” <sgdReference>
“Reference:” : <pmidReference>

metadata:
Publicationtype= curated database

Publicationtitle=<gene>-<phenotype>-<chemical>

Scientific value= 5

documentId= <sgdReference>/<gene>/<chemical>

Source: Saccharomyces Genome Database (SGD)

Publicationdate= <date>

url: http://www.yeastgenome.org/reference/<**sgdReference**>/overview

example:
http://www.yeastgenome.org/reference/S000180028/overview

Source Mongo collection	RD	RT
SGD	0	66