PIG QTL database : QTL annotations - Mappings, Thesaurus, Concept and Triple/Publication generation

Data Integration Activities

The following activities have to take place:

1. Create Mapping collection
2. Add terms/synonyms to thesaurus
3. Create Concepts, with measures
4. Create Triples, with measures
5. Create Publications, with measures

The input file name is dynamically created upon downloading. It can be found in AnimalQTLdb file in file server /backupdisk2/ inputdata.

Dataset / catalog level metadata

Dataset title: Pig QTLdb

Dataset description: Pig Quantitative Trait Locus (QTL) Database (Pig QTLdb), which is part of the Animal QTLdb, contains pig QTL and association data curated from published data. The database is designed to facilitate the process for users to compare, confirm, and locate the most plausible location for genes responsible for quantitative traits important to pig production.

Download URL: http://www.animalgenome.org/cgi-bin/QTLdb/SS/download?file=gbpSS_10.2

Note: This is an indirect link to the actual GFF file as the actual file changes its name based on session

Release/version: 30

Release issue date: 29-08-2016

Download date/time: 26-09-2016

distribution format: GFF3, https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

Data Record Metadata

The file is a standard GFF3 file ; a standard file format for storing genomic features in a text file. GFF stands for Generic Feature Format. GFF files are plain text, 9 column, tab-delimited files.

Column 9 "attributes": key=value pairs separated by ';', mandatory keys:

Example entry

Chr.X Animal QTLdb Health_QTL 113078996 113962590 . . . QTL_ID=31798;Name=PRRS viral load;Abbrev=PRRSVL;PUBMED_ID=24592976;trait_ID=856;trait=PRRS viral load;breed=duroc,large white,pietrain,yorkshire,Landrace;FlankMarkers=rs80851248,rs81305609;VTO_name=response to viral infection trait;Map_Type=Genome;Model=Mendelian;Test_Base=-;Significance=Significant;Variance=1.24

Create a mapping collection based on key=QTL_ID (column 9-1)

	key	M/O	Meaning	Syntax	example
1	seqid	M	Identifier, defines the coordinate system	single value - string : REMOVE: “CHR.” from the value	“Chr.X” -> “X”
2	source	M	Data base id / publication	single value - string	Animal QTLdb
3	type	O	QTL- main category. No formal ontology used	single value - string	Health_QTL
4	start	M	Starting base coordinate of the feature	Integer (1 based)	113078996
5	end	M	End base coordinate of the feature	Integer (1 based)	113962590
6	score	O	Pvalues ; missing value	Floating point number	.

			denoted by “.”		
7	strand	O	Direction	+ - ? .	.
8	phase	O	Indicates number of bases to be removed for beginning of feature	0,1,2	.
9	Attributes			Semicolon separated list of key-value pairs	
9-1	QTL_ID	M	Unique identifier of the feature in database	integer	31798
9-2	Name	M	Display name of the feature	string	PRRS viral load
9-3	Abbrev	O	Display name abbreviation	string	PRRSVL
9-4	PUBMED_ID	O	PMID for publication reference	integer	24592976
9-5	trait_ID	O	Id of trait (organism function)	integer	856
9.6	trait	O	Name of trait	string	PRRS viral load
9.7	breed	O	Specific Organism strains	List of strings, “,” separated	duroc,large white,pietrain,yorkshire,Landrace
9.8	FlankMarkers	O	Identifiers of genomic markers	List of strings, “‘“ separated	rs80851248,rs81305609
9.9	VTO_name	O	Trait Mapping to ontology	string	response to viral infection
9.10	CMO_na	O	Trait Mapping	string	hemoglobin

	me		to ontology		concentration
9.11	Map_Type	O	Method detail	string	Genome
9.12	Model	O	Method detail	string	Mendelian
9.13	Test-base	O	Method detail	string	Chromosome-wise
9.14	Significance	O	Outcome	string	Significant
9.15	P-value	O	Outcome measurement	Floating point	<0.05
9.16	F-Stat	O	Outcome measurement	Floating point	6.82
9.17	Variance	O	Outcome measurement	Floating point	1.24
9.18	Dominance_Effect	O	Outcome Measurement	Floating point	-1854
9.19	Additive_Effect	O	Outcome measurement	Floating point	-264
9.20	Bayes-value	O	Outcome measurement	Floating point	
9.21	Likelihood-ratio	O	Outcome measurement	Floating point	
9.22	LOD-score	O	Outcome measurement	Floating point	
9.23	gene_id	O	Gene	string	
Constant	genome_build	M	Concept measure	string	SS_10.2
Constant	date	M	29-08-2016	File date	29-08-2016
Constant	version	M	Source version	“Version: 30”	Version: 30

Constant	tax	M	Solr thesaurus category	“sus scrofa”	sus scrofa

Thesaurus and Concept generation

For this activity, the mapping collection provides:

- New terms / new uuids - insertion in Solr and Concept Collection
- Synonyms - insertion in Solr
- Defintions - insertion in Concept collection

The following 2 concepts and measures are created in the thesaurus and concept collection:

Type	Pref.Name : key in source	Synonym: Key in Source	Definition: key	Measure: key in source
qtl	<source>:<QTL_ID> (<ABBREV>)	<source>:<QTL_ID>	<type>, <Name>	Chromosome:<seqid>
				Start: <start>
				End: <end>
				Taxonomy: <tax>
chromosome	“Chromosome ”<seqid> “(<tax>)”	-	-	-

LOGIC overview

1. identify if a new uuid needs to be created
2. determine preferred term, semantic type
3. create new concept / uuid to Solr
4. Create concept in Concept collection
    - determine preferred term
    - Determine synonym

- determine definition

1. Identify if new UUID needs to be created for “QTL”

1. Verify if the <source>:<QTL_ID> (<ABBREV>)already exists in solr.

Example: term: “Animal QTLdb:31798 (PRRSVL)” .

If the <id> exists for 1 UUID/GI, proceed to next term.

3. If the concept does not yet exist in Solr, create uuid and add to Solr

Add Preferred Term: <source>:<QTL_ID> (<ABBREV>)

id":
"**term**": "<source>:<QTL_ID> (<ABBREV>)",
"**source**": “<source>",
"**knowledgebase**": "PigQTLdb",
"**semantictype**": 206
semanticcategory:”Genes and Molecular Sequences”
"**gi**": "generate",
“taxonomies”:sus scrofa
"**preferred**": "T",
"**_version_**":

Add synonym: <source>:<QTL_ID>

id":
"**term**": "<source>:<QTL_ID>",
"**source**": “<source>",
"**knowledgebase**": "PigQTLdb",
"**semantictype**": 206
semanticcategory:”Genes and Molecular Sequences”
"**gi**": "generate",
“taxonomies”:sus scrofa
"**preferred**": "F",

"**_version_**":

2. Identify if new UUID needs to be created for “Chromosome”

1. Verify if the term “chromosome ”<seqid> “(<tax>)” already exists in solr.

Example: “chromosome 1 (sus scrofa)”

If the “Chromosome ”<seqid> “(<tax>)” exists for 1 UUID/GI, proceed to next mapping document.

4. If the concept does not yet exist in Solr, create uuid and add to Solr

   Add Preferred Term: “chromosome ”<seqid> “(<tax>)”

id":
"**term**": ""chromosome ”<seqid> “(<tax>)",
"**source**": "<source>",
"**knowledgebase**": "chromosome",
"**semantictype**": 26
semanticcategory:”Anatomy”
"**gi**": "generate",
"**taxonomies**":<tax>
"**preferred**": "T",
"**_version_**":

Concept Generation in concept collection

Trigger for concept generation in Mongo collection

Source	Knowledgebase	Semantic Type	Semantic Group	Preferred T/F
QTL : Animal QTLdb	PigQTLdb	205	Genes and Molecular	T

			Sequences	
Chromosome: Animal QTLdb	chromosome	26	Anatomy	T

when Solr record contains source=AnimalQTL db && Preferred = T -> create a new Concept entry in Mongo.

For that gi, identify key to the mapping collection to get the definition:
The substring of the term of the Solr record (<<QTL_ID>>) is the <id> of the PigQTL mapping collection:

Identify Source-Mongo document and KEY

Solr Criteria	**Mongo Document Key**	**Concept attribute**	**Value type**	**Error handling**
knowledgebase_id = PigQTLdb	Substring of term : <QTL_ID>	<id>		do not create concept

Example: for the term : Animal QTLdb:31798 (PRRSVL) , the value 31798 is the unique key to the Mongo mapping collection to fetch the following data to create the concept record:

Mapping criteria QTL

Attribute	**Source key**	**Concept target key**
name	name	name
definition	<type> ,<name>	definition
semantictype		206
semanticcategory		
Measures:		
	<seqid>	Chromosome:
	<start>	Start:
	<end>	End:
	<genome_build>	Genome:
	<tax>	Taxonomy:

Mapping Criteria “Chromosome”

Attribute	Source key	Concept target key
name	name	name
semantictype		26
semanticcategory		Anatomy
	<genome_build>	Genome:

Access parameter

Source Mongo collection	RD	RT
PIGQTLdb	0	110

Triple Generation

There is 4 types of triple generated:

Triple 1: <source>:<QTL_ID> (<ABBREV>) - is associated with - VTO_name

Triple 2: <source>:<QTL_ID> (<ABBREV>) - is associated with - CMO_name

Triple 3: <source>:<QTL_ID> (<ABBREV>) - is part of - “<chromosome> (sus scrofa)”

Triple 4: gene_id - is part of - <source>:<QTL_ID> (<ABBREV>)

Triple 1:

For each document, create a triple:

Subject: <id> ;

solr query: term: <source>:<QTL_ID> (<ABBREV>) ; semantictype: 206

Predicate : is associated with

Object: Solr query: term: <VTO_name> && source: VTO

Create a triple for each uuid/GI

Triple 2:

For each document, create a triple:

Subject: <id> ;

solr query: term: <source>:<QTL_ID> (<ABBREV>) ; semantictype: 206

Predicate : is associated with

Object: Solr query: term: <CMO_name> && source: CMO

Create a triple for each uuid/GI

Triple 3:

For each document, create a triple:

Subject: <id> ;

solr query: term: <source>:<QTL_ID> (<ABBREV>) ; semantictype: 206

Predicate : part_of

Object: Solr query: “chromosome ”<seqid> “(<tax>)”

Triple 3:

For each document, create a triple:

Subject: <gene_id> ;

solr query: term: [entrezgene]<gene_id> && semantictype: 28

Predicate : part_of

Object: <id>
Solr query: term: <source>:<QTL_ID> (<ABBREV>) ; semantictype: 206

Create a triple for each uuid/GI

Measures (create for each of the triples)

key	value
Original trait description	<trait>
breed	<breed>
Associated Markers	<FlankMarkers>
Map type	<Map_Type>
Model	<Model>
Test base	<Test-base>
Significance	<Significance>
P-value	<P-value>
F-stat	<F-Stat>
Variance	<Variance>
Dominance effect	<Dominance_Effect>
Additive effect	<Additive_Effect>
Bayes value	<Bayes-value>
Likelihood Ratio	<Likelihood Ratio>
LOD score	<LOD-score>

Publication Generation

For all triples generated out of a mapping record (so per <id>) 1 publication is created with the following characteristics:

Measure: Publicationtype= QTL analysis

Scientific value = 3

Institution= Animal QTLdb

Publicationtitle= PigQTLdb/

Publication ID = PigQTLdb/QTL_ID

Publicationdate= <date>

Publicationsource:PigQTLdb

URL : http://www.animalgenome.org/cgi-bin/QTLdb/SS/qdetails?QTL_ID=<QTL_ID>

Triple Reference to Pubmed id:

Based on the <PUBMED_ID>, a reference to the pubmed abstract is created for each triple in the mapping document.

Source Mongo collection	RD	RT
PigQTLdb	0	110

Data Processing System