

Solanacea-centric Linked Data Platform for Plant Breeding & Genomics

Scientific Symposium: FAIR Data Science for Green Life Sciences

netherlands

eScience center

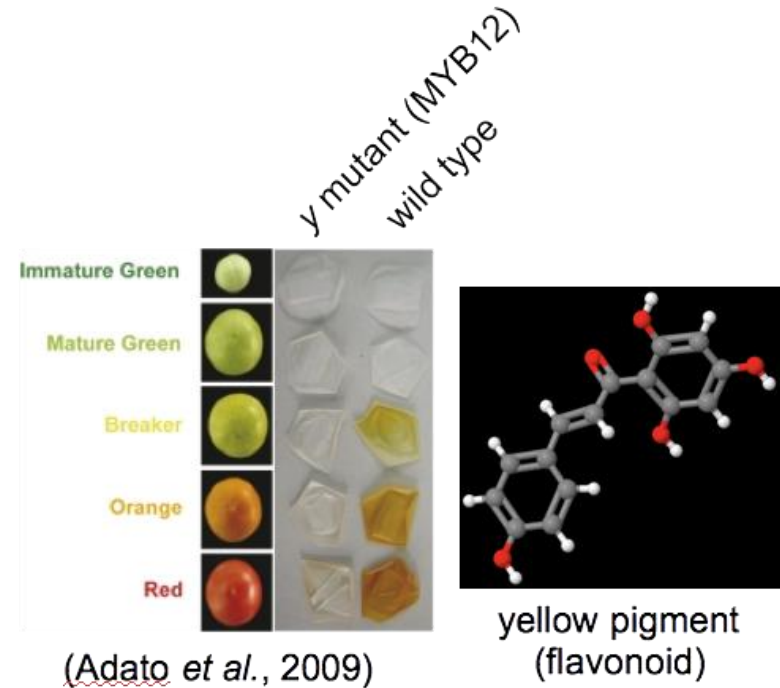


by SURF & NWO

Arnold Kuzniar

12/12/2018

Linking genes to traits

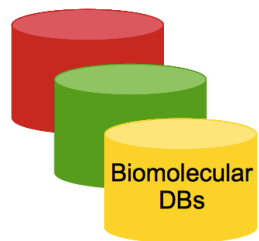


Methodology

Structured

Data

Unstructured

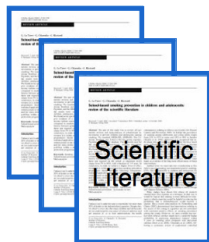


Retrieve
Genome Annotations



[SIGA.py](#) (RDFizer)

[xenon-cli](#) (distributed/parallel comp.)



Text/Table
Mining

Quantitative
Trait Loci

[QTLTableMiner++](#)

Semantic
Integration

RDF(S)

OWL

R2RML

SPARQL

W3C®



python™



FAIR
DATA



Linked Data

Analytics/
Ranking

Candidate
Gene(s)

[pbg-ld](#) (SPARQL API)

[grlc](#) (REST API)

[FAIR Data Point](#) (REST API [metadata])



<https://github.com/candYgene>



Linked Data approach



Sir Tim Berners-Lee (W3C)

- Use (persistent) URIs/IRIs to identify things
- Use HTTP-resolvable URIs
- Use RDF graph model (triples)
- Link to other resources



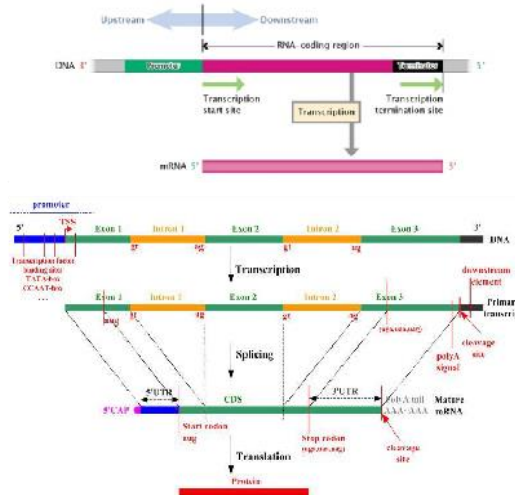
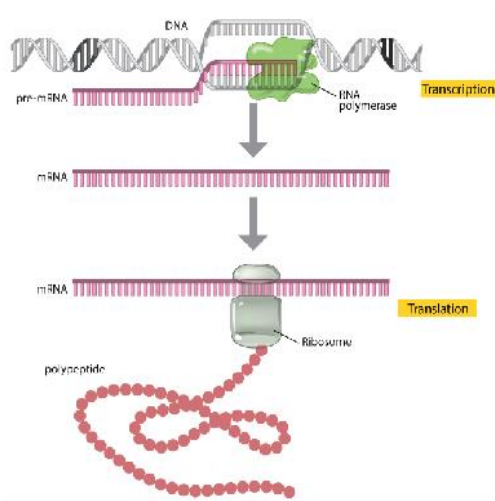
RDF turtle format

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .  
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .  
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .  
@prefix obo: <http://purl.obolibrary.org/obo/> .
```

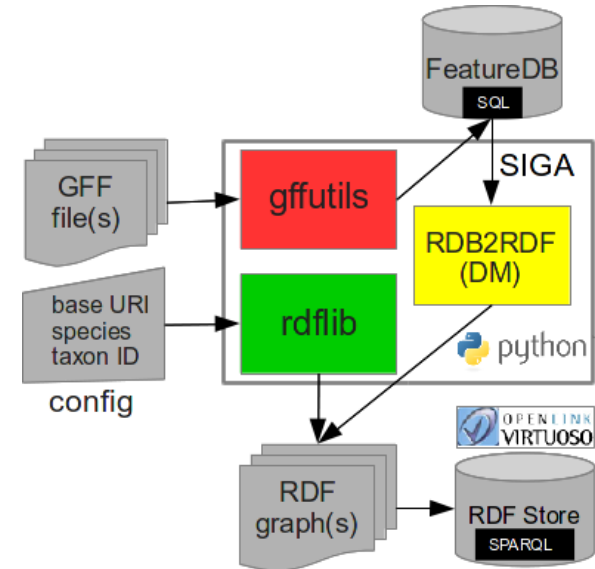
Solyc00g005000.2 is a gene

```
<http://solgenomics.net/genome/Solanum_lycopersicum/gene/Solyc00g005000.2> rdf:type obo:S0_0000704 ;  
rdfs:label "gene Solyc00g005000.2"^^xsd:string ;
```

Genome annotations

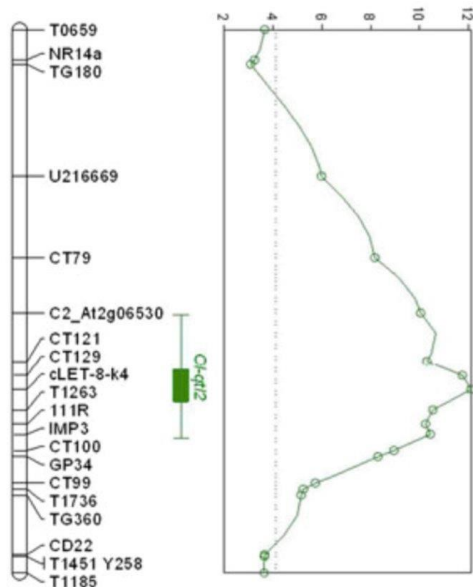


Use controlled vocabularies & ontologies
(e.g. GenBank Feature Table, SO[FA], FALDO, DCMI)



Software architecture

Mining QTLs from literature



e.g., QTL mapped on chr12
(Faino *et al.*, 2012)

REGRESSION APPROACH
An association mapping approach to identify favourable alleles for tomato fruit quality breeding

list of trait tables

Table 3 Phenotypic variation of traits analyzed in the whole collection

Trait	CV	Min	Max	Mean	CV%
Acetic Acid (mg 100g ⁻¹ FW)	0.06	2.04	32.25	33.24	17.33
β-carotene (μg g ⁻¹ FW)	0.75	0.53	1.76	1.39	10.02

Table 2 Multiple regression analyses between phenotypic traits and population structure

Trait	Regression results
Acetic Acid	0.120
β-carotene	0.079

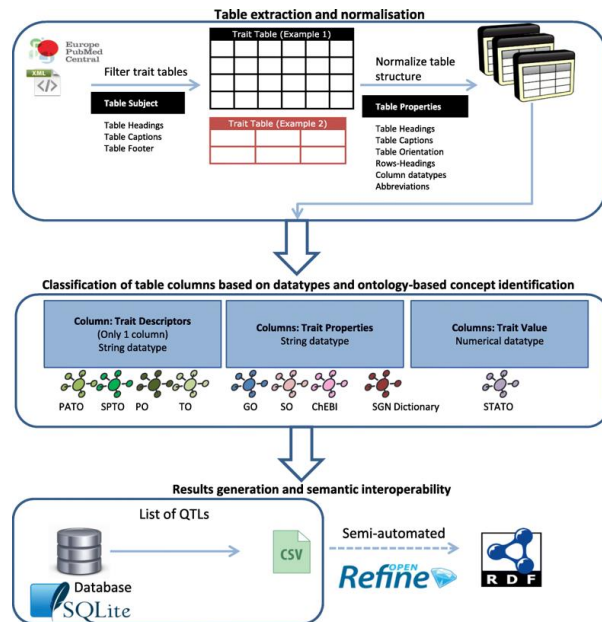
Table 3 Association statistics of markers significantly associated with seven traits by Mixed Linear Model (MLM) with two different RAUP thresholds (5% and 10%)

Marker	Marker name	SNP ID	Gene	Ch	BP (kb)	R ² (RAUP)	R ² (5%)	R ² (10%)
AUA	2983	NR14a_1_2006	Soly12G110802.1	3	5788039	2.76E-04	0.140	3.38E-04
	7988	NR14a_1_3077	Soly12G12302.1	3	5788994	2.76E-04	0.140	3.38E-04
	1245	NR14a_1_3255	Soly12G12455.1	3	6176003	3.68E-04	0.179	4.38E-04
tg (B-C)	3022	NR14a_1_1760	Soly12G1760.1	1	7454483	3.62E-04	0.138	4.38E-04
	3023	NR14a_1_1702	Soly12G1702.1	1	7454899	2.48E-04	0.098	
	3028	NR14a_1_1705	Soly12G1705.1	1	7455489	4.09E-04	0.132	2.48E-04

Accession	Trait ID	Transcriptomic Dataset	Gene/Protein Accession	Ref ID
Acetic Acid	CT08_22652	Chromosome: 3 CV%: 17.33 R ² : 0.16 Mean: 33.59 Min: 22.40 Max: 51.23 Regression Results P-value: 0.170 Regression Results R ² : 0.001 Marker Index: 2383 Site bp: 57066678	Gene: Soly12G12302.1 SolCapID: solcap_map_1_2006	PMCA256912
β-carotene	TQ_0002095	Chromosome: 1 CV%: 60.02 R ² : 0.75 Marker Index: 2022 Mean: 1.79 Min: 0.11 Regression Results P-value: 0.050 Regression Results R ² : 0.073 Site bp: 74514883	Gene: Soly12G1760.1 SolCapID: solcap_map_1_1760	PMCA256912

URL ID	Track URL	Genomic Coordinates
PMCA256912 - Table3_01	http://port.eurohistory.org/track/CT08_22652	R2: 504073, 4381133, 43811306
PMCA256912 - Table3_02	http://port.eurohistory.org/track/TQ_0002095	R2: 504073, 4381133, 43811306

QTLs "buried" in tables



QTL TableMiner++
(Singh *et al.* 2018)

Data sources



Domestic tomato (*Solanum lycopersicum*)



- gene models (SGN & Ensembl) + proteins (UniProt)
- genetic markers (SGN & SolCAP)
- QTLs (Europe PMC)



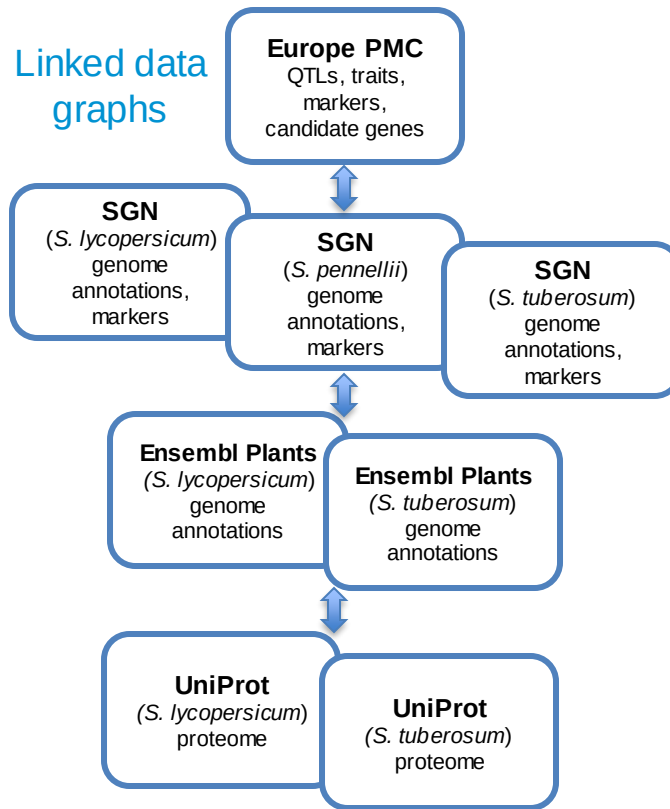
Wild tomato (*S. pennellii*)

- gene models (SGN)
- genetic markers (SGN)
- QTLs (Europe PMC)



Potato (*S. tuberosum phureja*)

- gene models (PGSC & Ensembl) + proteins (UniProt)
- genetic markers (SolCAP & DARt)
- QTLs (Europe PMC)



FAIR research output

The screenshot shows the 'candYgene' software entry. It includes a description: 'Software developed as part of the NLeSC & WUR project: Prediction of candidate genes for traits using interoperable genome annotations'. Below the description, there are tabs for 'Repositories', 'People', 'Teams', 'Projects', and 'Settings'. Under 'Repositories', three pinned repositories are listed: 'pbg-ld' (Linked Data Platform for Plant Breeding & Genomics), 'QTM' (Forked from PBR/QTM, OTL TableMiner++ tool for mining tables in scientific articles), and 'siga' (Semantically Interoperable Genome Annotations).

The screenshot shows the 'Research Software Directory' homepage. It features a search bar with the placeholder text 'Start typing here to search for software'. Below the search bar, there are tags for 'QTLTableMiner++', 'QTM', 'pbg-ld', and 'pb'. The 'QTLTableMiner++' tag is highlighted, showing a description: 'A tool to extract Quantitative Trait Locus data from tables in genetics literature and make this data more FAIR.' The 'QTM' tag is also visible, with a description: 'Provides easy access to integrated plant-specific data on genes associated with traits of interest.'

The screenshot shows the Zenodo page for 'pbg-ld: Linked Data Platform for Plant Breeding & Genomics'. It includes a description: 'Linked Data Platform for Plant Breeding & Genomics'. The page also lists the authors: 'Kuzmar, Arnold', 'Singh, Gurnoor', and 'Contact person(s): Piskorski, Richard'.

[doi:10.5281/zenodo.1458169](https://doi.org/10.5281/zenodo.1458169)

The screenshot shows the Zenodo page for 'Quantitative Trait Loci in Solanaceae species'. It includes a description: 'Quantitative Trait Loci in Solanaceae species'. The page also lists the authors: 'Kuzmar, Arnold', 'Singh, Gurnoor', and 'Contact person(s): Piskorski, Richard'.

[doi:10.5281/zenodo.1215044](https://doi.org/10.5281/zenodo.1215044)

The screenshot shows the Zenodo page for 'QTLTableMiner++: a tool for mining tables in scientific articles'. It includes a description: 'QTLTableMiner++: a tool for mining tables in scientific articles'. The page also lists the authors: 'Singh, Gurnoor', 'Kuzmar, Arnold', and 'Contact person(s): Piskorski, Richard'.

[doi:10.5281/zenodo.1193640](https://doi.org/10.5281/zenodo.1193640)

The screenshot shows the Zenodo page for 'SIGA.py'. It includes a description: 'SIGA.py'. The page also lists the authors: 'Kuzmar, Arnold', 'Researcher(s): Bakshi, Rama', and 'A light-weight command-line tool to generate semantically interoperable genome annotations.'

[doi:10.5281/zenodo.1076438](https://doi.org/10.5281/zenodo.1076438)

The screenshot shows the Zenodo page for 'FAIR Data Point: a web service to provide machine-readable descriptions about datasets'. It includes a description: 'FAIR Data Point: a web service to provide machine-readable descriptions about datasets'. The page also lists the authors: 'Kuzmar, Arnold', 'Khalafat, Raghav', and 'Researcher(s): Gaval, Anand', 'Bonne de Silva Santos, Luc Oliveira'.

[doi:10.5281/zenodo.1083951](https://doi.org/10.5281/zenodo.1083951)

The screenshot shows the article 'QTLTableMiner++: semantic mining of QTL tables in scientific articles' in BMC Bioinformatics. It includes a description: 'QTLTableMiner++: semantic mining of QTL tables in scientific articles'. The page also lists the authors: 'Gurnoor Singh, Arnold Kuzmar, Erik M. van Mulligen, Anand Gaval, Christian W. Bachem, Richard G.F. Visser and Richard Finkers'.

The screenshot shows the 'Precision Search & Find' application. The 'Entity Label Lookup' tab is selected. The 'Label' input field contains the text 'fruit shape'. To the right of the input field is a 'Describe' button. Below the input field, a list of search results is displayed, each with a label and a URL. The results include 'fruit shape', 'fruit shape circular', 'fruit shape eccentric', 'fruit shape ellipsoid', 'fruit shape index', 'fruit shape index external 1', 'fruit shape index external 2', 'fruit shape rectangular', 'fruit shape triangle', 'fruit shape triangle 10%', 'fruit shape triangle 20%', 'fruit shape triangle 30%', and 'fruit shape triangle 5%'. The interface also shows 'OpenLink Virtuoso vrs' and '62 GB total memory'.



OPEN INK

SOFTWARE

About: **OTL:4321030_4_14**

Goto Sponge

NotDistinct

Perma

An Entity of Type : http://purl.obolibrary.org/obo/SO_0000771, within Data Space

Type: QTL ▾

Command: Start New Facet ▾

Go

Attributes	Values
type	QTL
label	QTL:4321030_4_14
dct:identifier	QTL:4321030_4_14
dct:isReferencedBy	http://identifiers.org/pmc/PMC4321030
location	chromosome 11:33041803-51521123
correlated with	variation gene165_0-123 variation gene210_0-123 variation gene220_0-123 variation gene221_0-123
correlated with condition	fruit shape fruit shape
overlaps	gene Solyc11g038340.1 gene Solyc11g038350.1 gene Solyc11g039350.1 gene Solyc11g039360.1 gene Solyc11g039370.1 »more»



OPEN LINK

SOFTWARE

About: **gene [Slyc11g038340.1](#)**

[Goto Sponge](#)
[NotDist](#)

An Entity of Type : http://purl.obolibrary.org/obo/SO_0001217, within [Data Spa](#)

Type:

protein_coding_gene

Command:

Start New Facet

Go

Alias: *Slyc11g038340*; Name: *Slyc11g038340.1*

Attributes	Values
type	protein_coding_gene
label	gene Slyc11g038340.1
comment	Alias: Slyc11g038340 ; Name: Slyc11g038340.1
sameAs	gene Slyc11g038340.1
dct:identifier	Slyc11g038340.1
location	chromosome 11:45259363-45259605
has_part	prim_transcript Slyc11g038340.1.1
transcribed_to	prim_transcript Slyc11g038340.1.1

is overlaps of

[QTL:3464107_4_25](#)

[QTL:4321030_4_14](#)

[QTL:4321030_4_7](#)



Gene

<https://github.com/candYgene/pba-ld>

Data access & analysis

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

```
Query Text
PREFIX rd[is: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX so: <http://purl.obolibrary.org/obo/so#>

SELECT
  str(?qtl_id) AS ?qtl_id
  str(?sgn_gene_id) AS ?sgn_gene_id
  str(?sgn_trans_id) AS ?sgn_trans_id
  str(?annot) AS ?annot
WHERE {
  GRAPH <http://europepmc.org/articles> {
    ?qtl a obo:SO_0000771 ;
    obo:RO_0003308 ?trait ;
    so:overlaps ?gene ;
    dcterms:identifier ?qtl_id .
    FILTER(?trait = obo:SP_0000366)
  }
  GRAPH <http://solgenomics.net/genome/Solanum_lycopersicum> {
    ?gene so:transcribed_to ?transcript ;
    dcterms:identifier ?sgn_gene_id .
    ?transcript rdfs:comment ?annot ;
    dcterms:identifier ?sgn_trans_id .
  }
}
```

SPARQL API

local

Count genomic features

Get gene annotations from SGN given a gene ID

Get genes contained in a QTL

Get QTLs in an article

Get QTLs associated with a trait ID

REST API

```
In [8]: df = pd.read_csv(StringIO(resp.text))
df

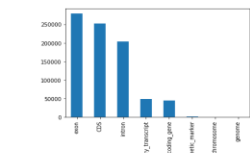
Out[8]:
```

	feature_name	feature_id	n
0	exon	(SO:000147)(http://purl.obolibrary.org/obo/SO_000147)	27874
1	CDS	(SO:000316)(http://purl.obolibrary.org/obo/SO_000316)	25290
2	intron	(SO:000188)(http://purl.obolibrary.org/obo/SO_000188)	204027
3	protein_coding_primary_transcript	(SO:000120)(http://purl.obolibrary.org/obo/SO_000120)	44823
4	protein_coding_gene	(SO:000121)(http://purl.obolibrary.org/obo/SO_000121)	44965
5	genetic_marker	(SO:0001645)(http://purl.obolibrary.org/obo/SO_0001645)	2225
6	chromosome	(SO:000340)(http://purl.obolibrary.org/obo/SO_000340)	13
7	genome	(SO:0001006)(http://purl.obolibrary.org/obo/SO_0001006)	1

We can then manipulate the DataFrame to visualize our results

```
In [9]: df = df.set_index('feature_name')
df['n'].plot(kind='bar')

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc250b95990>
```



Jupyter notebook

Conclusions & future work

- Use/develop open-source, standards-compliant, well-documented software
- Linked (Open) Data principles $\approx$ FAIR Data principles (in practice)
- FAIR compliance not only for data but also for software/workflows
- Integrating & "semantifying" (semi-)structured data requires significant time investment & domain expertise
- Improve data access & processing workflow
- Extend *pbg-ld* with algorithms for candidate gene prioritization
- Prepare new software & data releases
- Deploy *pbg-ld* software stack in production environment (HPC Cloud)



Acknowledgements

- Gurnoor Singh (WUR)
- Richard Finkers (WUR)
- Christian Bachem (WUR)
- Richard Visser (WUR)
- Erik van Mulligen (EMC)
- Anand Gavai
- Carlos Martinez-Ortiz (NLeSC)
- Lars Ridder (NLeSC)
- e-infra team (SURFsara)

