Linked Data Platform for Plant Breeding & Genomics

Gurnoor Singh [1], Arnold Kuzniar [2], Carlos Martinez Ortiz [2], Richard GF Visser [1], Richard Finkers [1]

[1] Plant Breeding Research, Wageningen University & Research, The Netherlands

[2] Netherlands eScience Center, Amsterdam, The Netherlands

Contact email: a.kuzniar@esciencecenter.nl, richard.finkers@wur.nl

Background

Genetics research is focusing more and more on mining fully sequenced genomes and their annotations to identify the causal genes associated with specific traits (phenotypes) of interest. However, a complex trait is typically associated with multiple quantitative trait loci (QTLs), each with hundreds of genes positively or negatively affecting the desired trait(s). To help breeders to mine candidate genes, we developed an analytics platform that provides semantically integrated geno- and pheno-typic data on *Solanaceae* species [1]. This platform combines both unstructured data from plant science literature and structured data from publicly available biological databases using a Linked Data approach and follows the FAIR Data principles [2].

Methodology

QTLs were extracted from tables of (open access) articles using our recently developed tool, QTLTableMiner++ [3], while the genome/proteome annotations were obtained from the Sol Genomics Network (SGN), UniProt, and Ensemble Plants databases. These data were transformed into and integrated as RDF graphs including cross-references to many other relevant databases such as Gramene, Plant Reactome, InterPro and KEGG (KO). The resulting linked datasets are available for queries and downstream analyses through a web interface or programmatically through SPARQL and RESTful services (APIs).

Example queries

I. Compare annotated features of the tomato reference genome (*Solanum lycopersicum* str. Heinz 1706) according to the SGN and Ensembl Plants databases.

SGN

feature_name	feature_id	number
exon	SO:0000147	160001
CDS	SO:0000316	157233
intron	SO:0000188	125276
protein_coding_gene	SO:0001217	34725
protein_coding_primary_transcript	SO:0000120	34725
genetic_marker	SO:0001645	30718
three_prime_UTR	SO:0000205	15343
five_prime_UTR	SO:0000204	13548
chromosome	SO:0000340	13
genome	SO:0001026	1

SO: Sequence Ontology

Ensemble Plants

feature_name	feature_id	number
exon	SO:0000147	162535
protein_coding_primary_transcript	SO:0000120	34725
protein_coding_gene	SO:0001217	33785
miRNA	SO:0000276	3153
miRNA_gene	SO:0001265	3153
tRNA_gene	SO:0001272	908
snoRNA	SO:0000275	390
snoRNA_gene	SO:0001267	390
snRNA_gene	SO:0001268	255
snRNA	SO:0000274	255
rRNA	SO:0000252	94
rRNA_gene	SO:0001637	94
pseudogenic_tRNA	SO:0000778	76
chromosome	SO:0000340	13
RNA	SO:0000356	2

II. Comparative genome analysis of the cultivated tomato (*S. lycopersicum*) and its wild relative (*S. pennellii*) according to the SGN database.

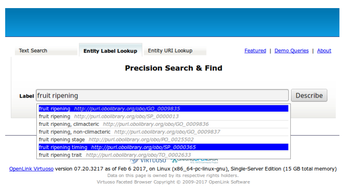
feature_name	feature_id	number
exon	SO:0000147	160001
CDS	SO:0000316	157233
intron	SO:0000188	125276
protein_coding_gene	SO:0001217	34725
protein_coding_primary_transcript	SO:0000120	34725
genetic_marker	SO:0001645	30718
three_prime_UTR	SO:0000205	15343
five_prime_UTR	SO:0000204	13548
chromosome	SO:0000340	13
genome	SO:0001026	1

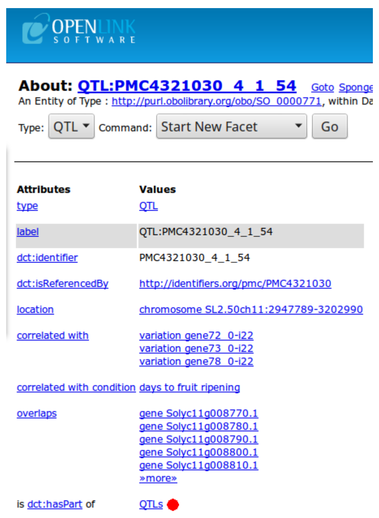
SO: Sequence Ontology

feature_name	feature_id	number
exon	SO:0000147	160001
CDS	SO:0000316	157233
intron	SO:0000188	125276
protein_coding_gene	SO:0001217	34725
protein_coding_primary_transcript	SO:0000120	34725
genetic_marker	SO:0001645	30718
three_prime_UTR	SO:0000205	15343
five_prime_UTR	SO:0000204	13548
chromosome	SO:0000340	13
genome	SO:0001026	1

III. Query gene-trait associations using a Virtuoso's facet (web-based) browser

Trait ⟷ QTL ⟷ Gene





Fruit Ripening

GO:0009835 (biological process)

SP:0000013

GO: Gene Ontology
SP: Solanacae Phenotypic Ontology

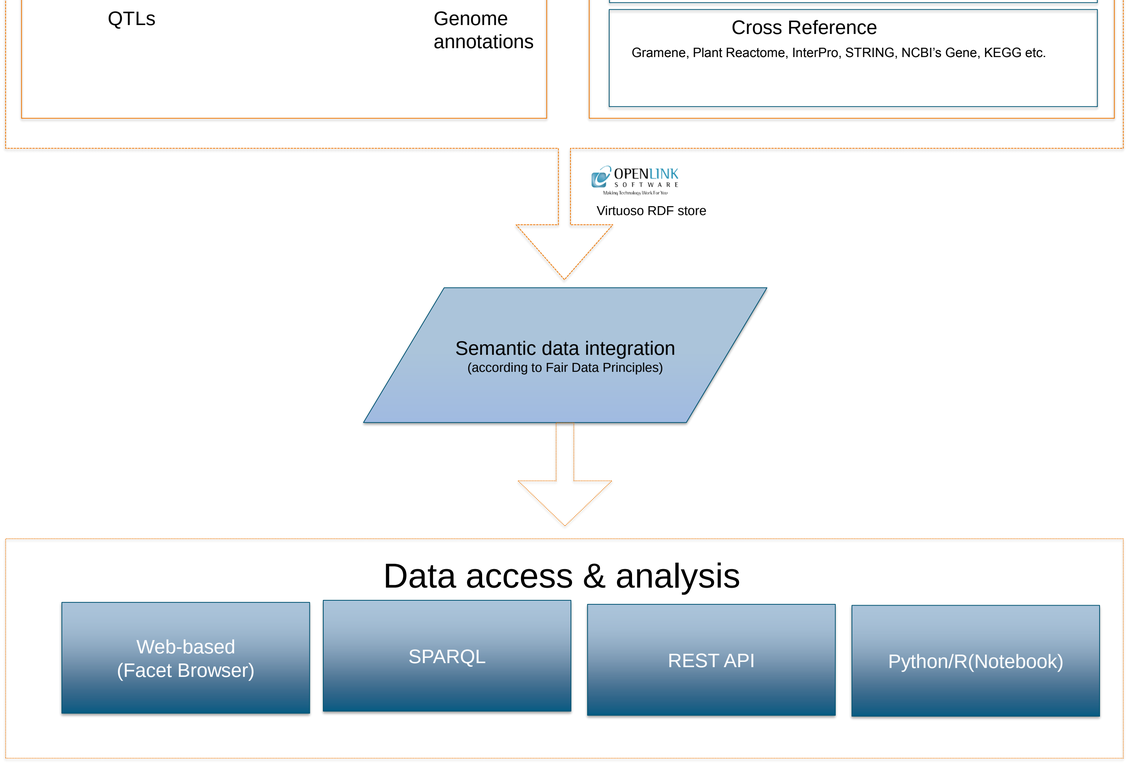


Figure 1. Data generation and ingestion pipeline

Conclusion

Our linked data platform provides semantically integrated plant-specific geno- and pheno-typic data to aid breeders prioritize over candidate genes associated with the trait of interest. We are working on extending this platform with algorithms to score and rank the genes according to the evidence in multiple data sources.

References

1. Kuzniar, A. et al. (2015). candYgene: enabling precision breeding through FAIR Data, doi: http://doi.org/10.5281/zenodo.30554
2. Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. http://doi.org/10.1038/sdata.2016.18
3. Singh, G. et al. (2018). QTLTableMiner++: semantic mining of QTL tables in scientific articles, doi: https://doi.org/10.1186/s12859-018-2165-7
4. Kuzniar, A. (2017) SIGA.py: a command-line tool to generate semantically interoperable annotations from GFF files. http://doi.org/10.5281/zenodo.1076438