

VESPUCCI: the integrated gene expression database for grapevine

Marco Moretto

marco.moretto@fmach.it

Fondazione Edmund Mach
Research and Innovation Centre
Computational Biology Unit

7th February 2020



FONDAZIONE
EDMUND
MACH



Acknowledgments

- Paolo Sonego
- Stefania Pilati
- Giulia Malacarne
- Laura Costantini
- Claudio Donati
- Claudio Moser
- Kristof Engelen

Why VESPUCCI?

Thousands of transcriptome data sets in public databases

The screenshot shows the GEO homepage with a search bar and a list of search results. The results include metrics like Datasets: 4348, Series: 77711, Platforms: 10660, and Samples: 202947.

Gene Expression Omnibus
GEO is a public functional genomics data repository supporting MIAMIE-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Getting Started:
Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, How to Download Data.

Tools:
Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, GEO BLAST, Programming Access, FTP Site.

Information for Submitters:
Login to Submit, Submission Guidelines, Update Guidelines.

Browse Content:
Repository Browser, Datasets: 4348, Series: 77711, Platforms: 10660, Samples: 202947.

ArrayExpress
Home | Browse | Submit | Help | About ArrayExpress | Contact Us | Log In

Search Examples: C-MEXP-21, cancer, p53, Genomics | advanced search

ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

[Browse ArrayExpress](#)

Data Content

Updated today at 06:00

• 69504 experiments

• 219792 assays

• 44.50 TB of archived data

Latest News

6 December 2016 - RESTful API (version 3) for ArrayExpress data is released!

If you're looking for ways to access or mine ArrayExpress programmatically, then our API is for you. New features in version 3 include retrieving experiments based on sample attribute class (e.g., "organism part", "sex/strain") and getting annotation for each sample in an experiment (previously we only served information aggregated across samples in an experiment). You can now also retrieve the FTP download link for each sequencing raw data file (Fastq) via a sample, even though the files are actually hosted at the European Nucleotide Archive. As before, the API supports context-specific search and Experimental Factor Ontology driven search extension, which help to reduce "boycatch" in search results.

We hope you'll enjoy the new features. For any questions or comments, please email us at arrayexpress@ebi.ac.uk.

Links

Information about how to search ArrayExpress, understand search results, how to submit data and

Tools and Access

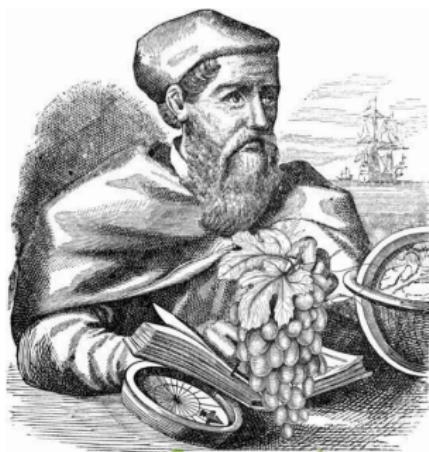
[Annotate](#) - web-based submission tool for ArrayExpress.

Related Projects

Discover up and down regulated genes in numerous experimental conditions in the ExpressionAtlas.

Why VESPUCCI?

Vitis
Expression
Studies
Platform
Using
COLOMBOS™
Compendia
Instances

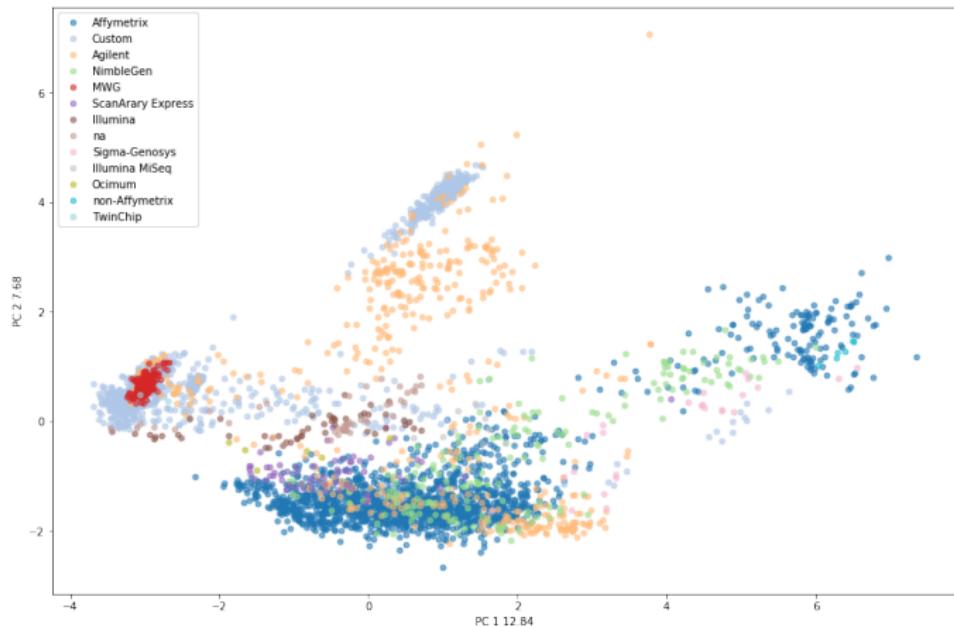


What is VESPUCCI? (In a nutshell)

- A database of **integrated** grapevine gene expression data
- A set of tools to **explore** such database

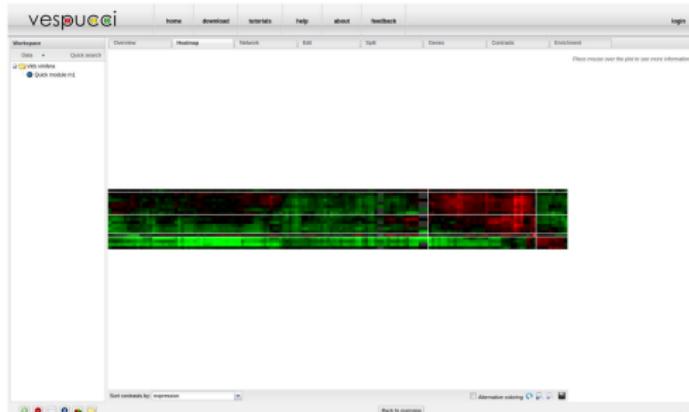
Why VESPUCCI?

Batch effect



Why VESPUCCI?

Ocean exploration analogy

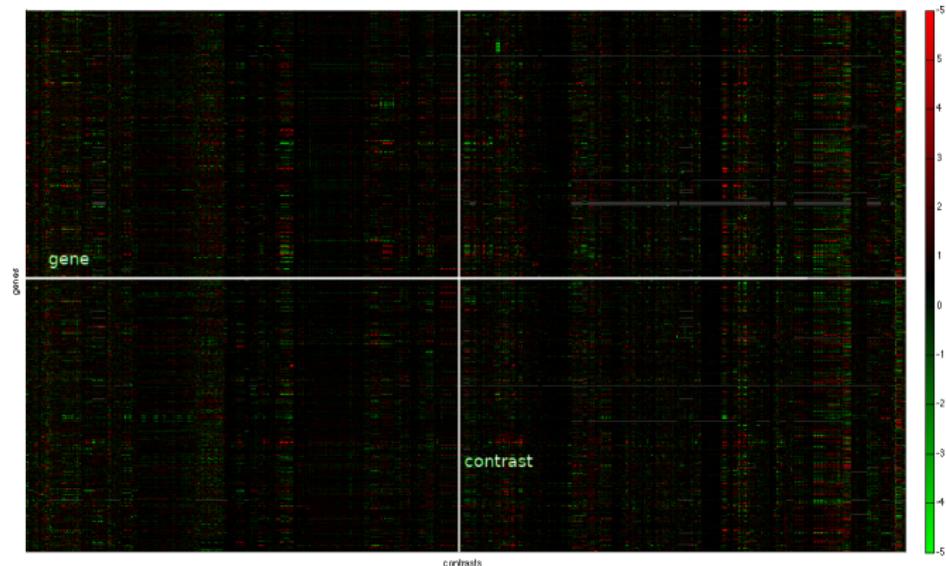


- Sea/Ocean
- Atlas/Map
- Path/Route
- Navigation tools

- All available gene expression data
- Gene-expression atlas (compendium)
- Gene-expression patterns
- VESPUCCI toolset

What is VESPUCCI?

Single coherent expression matrix in which each **row** represents a **gene** and each **column** represents a sample **contrast** (i.e. relative value given by the difference in log-scale between two conditions, a test and a reference).



What is VESPUCCI?

Module

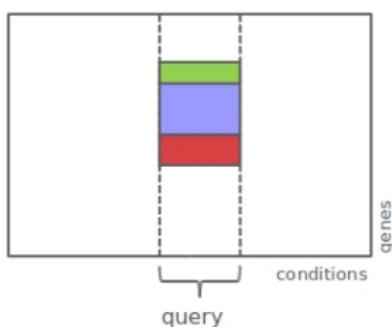
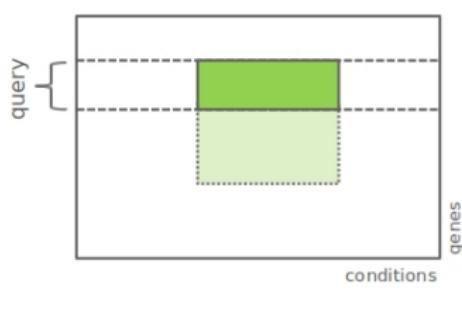
- basic concept (**unit**) in VESPUCCI
- **result** of every query
- is a **subset** of the whole gene expression matrix



Gene
centered



Sample/condition
centered



Updates

Technology

- Back-end (COMMAND>_)
- Normalizations
- Annotation system
- Front-end (COMPASS)

Data

- All transcriptomics experiments up until September 2019

How VESPUCCI is built

COMMAND>_

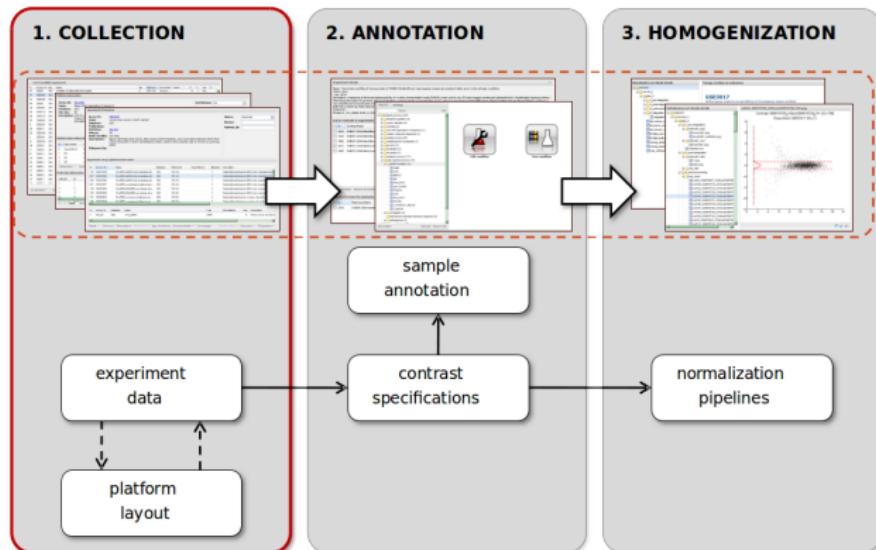
The screenshot shows the COMMAND software interface. On the left, there's a sidebar with tabs for 'Data collector', 'Options', and 'Address'. Below these are sections for 'Compendium: deneu_patio' and 'Parse Experiment GSE22540' (selected). Under 'Experiments', there's a tree view with 'Experiments' expanded, showing entries 1 through 13, each corresponding to a file named 'GSM1419678.ch1'. At the bottom of the sidebar are links for 'Experiment files' and 'Python editor'. The main area is titled 'Raw data for sample GSM1419678.ch1' and displays a table with columns 'ID', 'Bio Feature Reporter Name', and 'Value'. The table shows data for samples 78002 through 78033, with values ranging from 276.52 to 312.33.

ID	Bio Feature Reporter Name	Value
78002	276.52	316.67
78003	245.165	268.67
78004	231.919	270.11
78005	72.272	270.22
78006	83.187	258.22
78007	240.105	307.22
78008	88.250	319.67
78009	107.335	232.78
78010	128.196	255.67
78011	123.187	237.44
78012	143.301	375
78013	190.122	341.67
78014	240.68	313.11
78015	136.144	274.89
78016	85.271	234.89
78017	230.495	300.78
78018	320.129	341.78
78019	52.272	265
78020	190.328	285.44
78021	204.132	287.67
78022	63.57	290.56
78023	200.326	212.33

Publication Moretto *et al.* "First step toward gene expression data integration: transcriptomic data acquisition with COMMAND" *BMC Bioinformatics* 2019

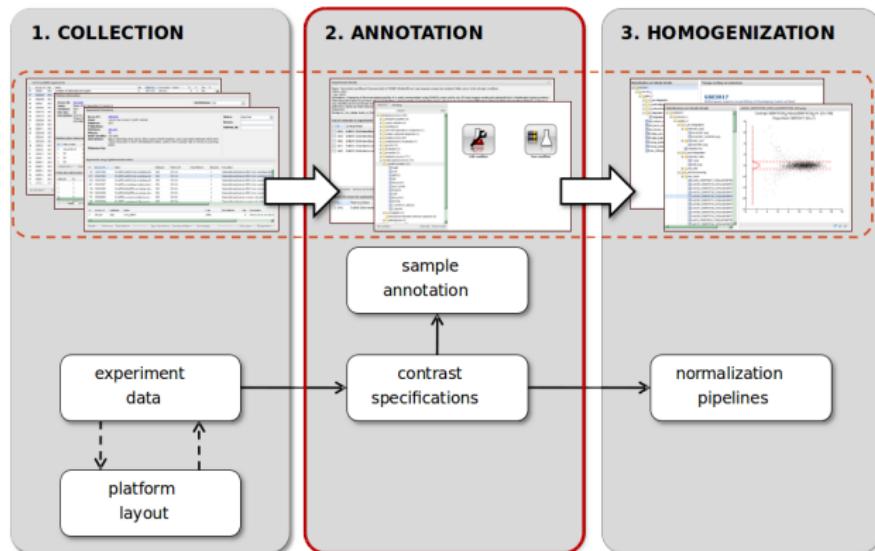
How VESPUCCI is built

COMMAND>_



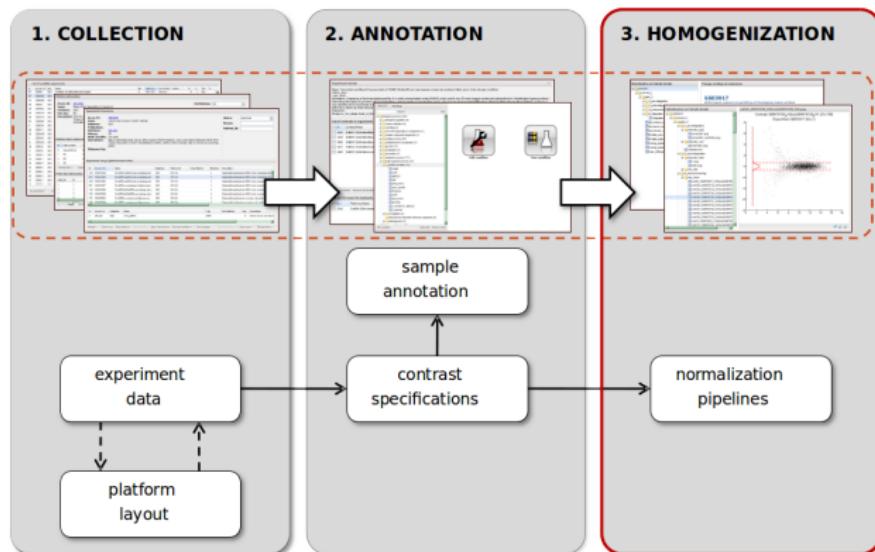
How VESPUCCI is built

COMMAND>_



How VESPUCCI is built

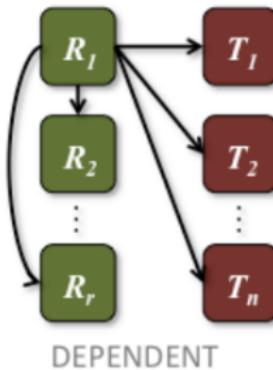
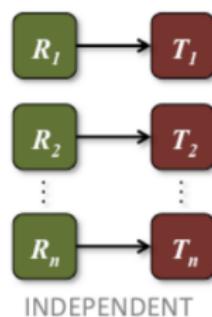
COMMAND>_



Normalization

The new technology supports different **normalization** strategies within the same compendium.

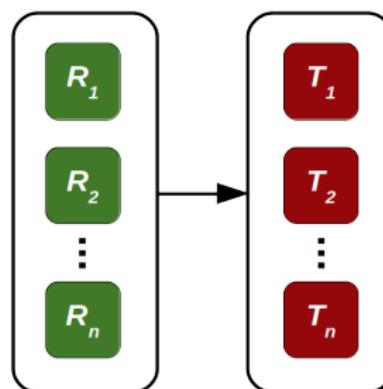
Legacy log-ratio



Normalization

The new technology supports different **normalization** strategies within the same compendium.

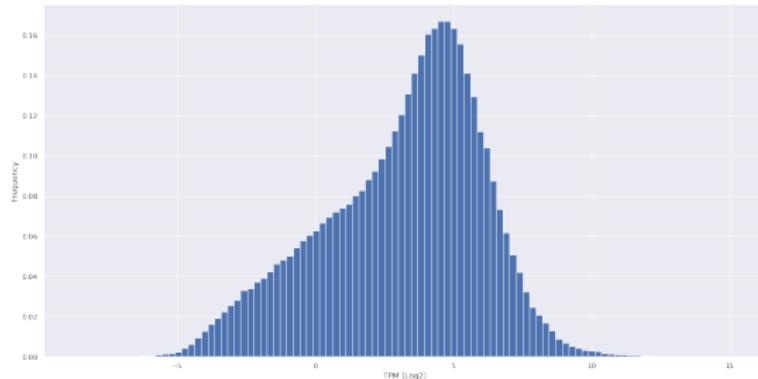
LIMMA



Normalization

The new technology supports different **normalization** strategies within the same compendium.

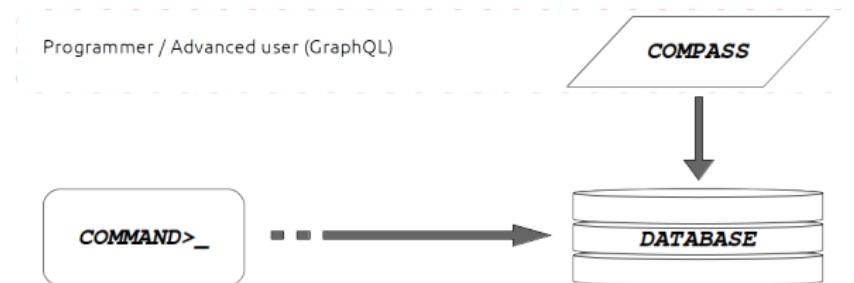
TPM - RNA-seq only



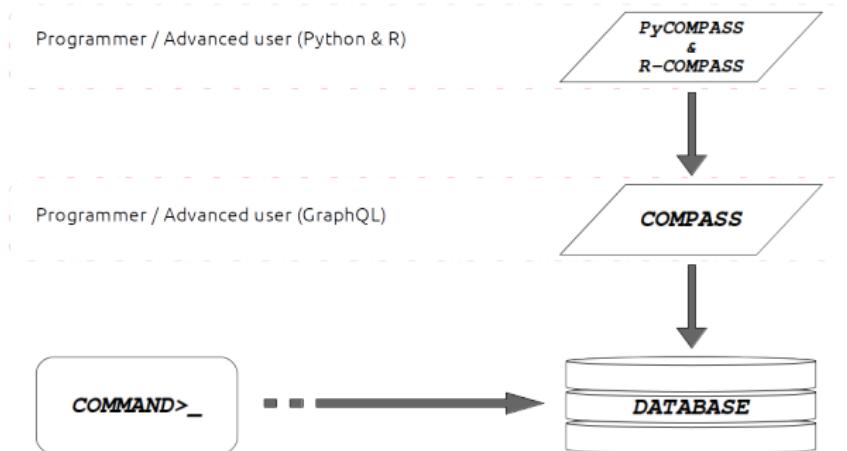
Accessing the data - the front-end



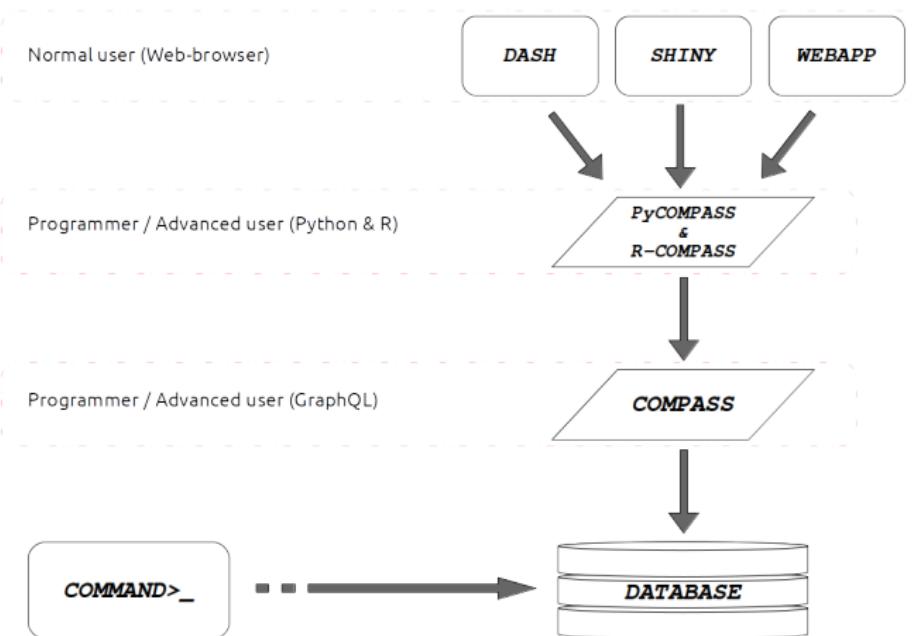
Accessing the data - the front-end



Accessing the data - the front-end



Accessing the data - the front-end



Accessing the data - the front-end: COMPASS the GraphQL interface

POST ▾ http://127.0.0.1:8000/graphql

Send **200 OK** TIME 475 ms SIZE 28.9 KB

GraphQL Auth Query Header Docs

schema not yet fetched

Query Variables **1**

```
1 r {
2   samples{compendium:"vitis_vinifera", experiment_ExperimentAccessId:"GGE3820"} {
3     edges {
4       node {
5         sampleName,
6         description
7       }
8     }
9   }
10 }
```

Preview ▾ Header **Header** Cookie Timeline

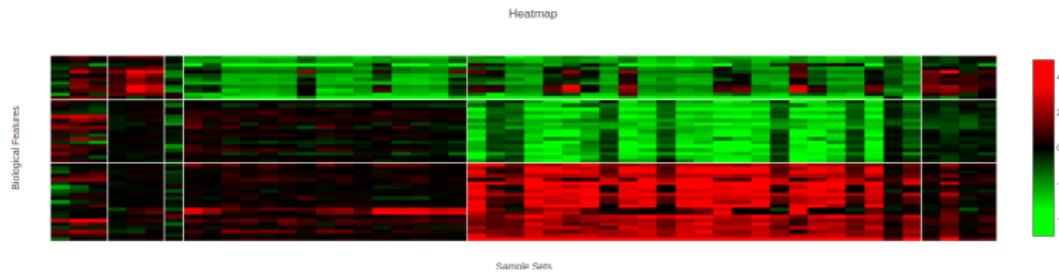
```
1 r {
2   "data": {
3     "samples": [
4       {
5         "edges": [
6           {
7             "node": {
8               "sampleName": "GGM27217.ch1",
9               "description": "Wild type (strain PY79)sporulation pert mutant Hour 6.5 exp. 20strains PY79 and PE454 were grown in parallel in hydrolyzed casein medium at 37°C to an OD600 of 0.6. Sporulation was induced by resuspension in Sterilin-Mendelstam medium at 37°C Six hour after rehydration a 100 microliter of each culture was collected for RNA isolation and immediately mixed with an equal volume of methanol at 4°C. RNA was isolated with a Hot Phenol/Trizol procedure. cDNA was prepared from 50 ug tot RNA and fluorescently labeled using SuperScript Reverse Transcriptase and Cy5-dUTP for the RNA sample from strain PY79 and Cy3-dUTP for the RNA sample from strain PE454. Protocols for RNA isolation, cDNA preparation and labeling, hybridization to the microarrays and washing procedure can be found at http://mba.harvard.edu/gkoch This is the third repetition of the experiment at this time point."
10           }
11         }
12       }
13     ]
14     "node": {
15       "sampleName": "GGM27217.ch2",
16       "description": "germ mutant (strain PE454)sporulation pert mutant Hour 6.5 exp. 20strains PY79 and PE454 were grown in parallel in hydrolyzed casein medium at 37°C to an OD600 of 0.6. Sporulation was induced by resuspension in Sterilin-Mendelstam medium at 37°C Six hour and thirty minutes after resuspension, 25 ml of each culture were collected for RNA isolation and immediately mixed with an equal volume of methanol at 4°C. RNA was isolated with a Hot Phenol/Trizol procedure. cDNA was prepared from 50 ug tot RNA and fluorescently labeled using superscript Reverse Transcriptase and Cy5-dUTP for the RNA sample from strain PY79 and Cy3-dUTP for the RNA sample from strain PE454. Protocols for RNA isolation, cDNA preparation and labeling, hybridization to the microarrays and washing procedure can be found at http://mba.harvard.edu/gkoch This is the third repetition of the experiment at this time point."
17     }
18   }
19   "node": {
20     "sampleName": "GGM27218.ch1",
21     "description": "wild type (strain PE454)sporulation pert mutant Hour 6.5 exp. 20strains PY79 and PE454 were grown in parallel in hydrolyzed casein medium at 37°C to an OD600 of 0.6. Sporulation was
22 }
```

Prettify GraphQL

Accessing the data - the front-end: pyCOMPASS the Python interface

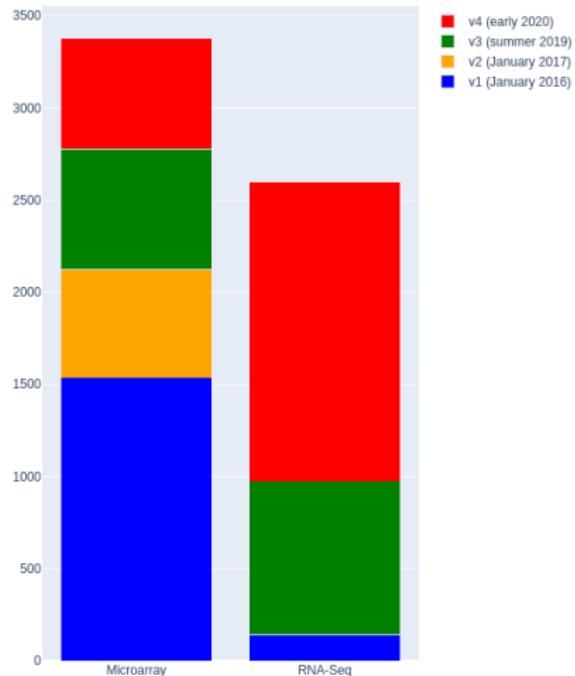
```
[6]: from pycompass import Connect, Compendium, Module, Sample, Platform, Experiment, Ontology, SampleSet, BiologicalFeature, Module, Plot
from IPython.core.display import display, HTML

conn = Connect('http://fempc0734:8000/graphql')
compendia = conn.get_compendia()
ss = SampleSet.using(compendia[0]).get(filter={'first': 50})
mod1 = Module.using(compendia[0]).create(samplesets=ss)
html = Plot(mod1).plot_heatmap(alternativeColoring=False, min=-5, max=5)
display(HTML(html))
```



Data

VESPUCCI growth over time (number of samples)



Gene and Sample annotation

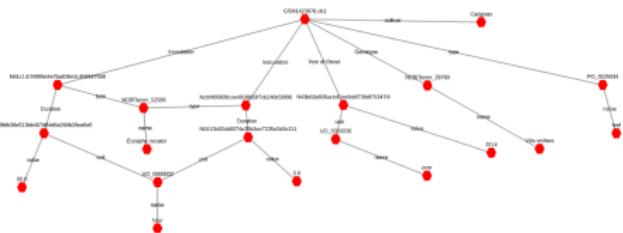
- From **home-made vocabulary** to standard **Ontologies**
- From “*bag of words*” to **RDF** triples

Subject	Predicate	Object
GSM1423076	cultivar	Carignan
GSM1423076	Genotype	NCBITaxon_29760
NCBITaxon_29760	name	Vitis vinifera
...

Gene and Sample annotation

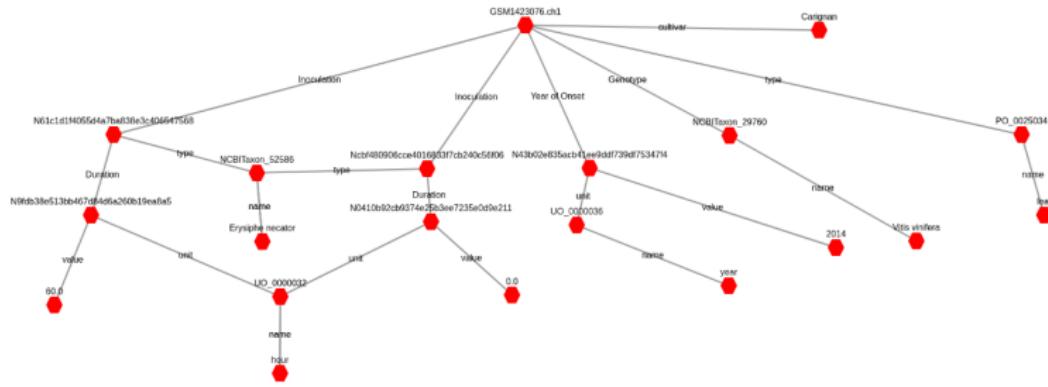
- From **home-made vocabulary** to standard **Ontologies**
- From “*bag of words*” to **RDF triples**

Subject	Predicate	Object
GSM1423076	cultivar	Carignan
GSM1423076	Genotype	NCBITaxon_29760
NCBITaxon_29760	name	Vitis vinifera
...



Gene and Sample annotation

- From **home-made vocabulary** to standard **Ontologies**
- From “*bag of words*” to **RDF triples**
- SPARQL
- Reasoning over Ontologies



Gene and Sample annotation

SPARQL Uniprot example

```
SELECT ?protein ?annotation
WHERE
{
    ?protein a up:Protein .
    ?protein up:organism ?organism .
    ?organism rdfs:subClassOf taxon:29760 .
    ?protein up:annotation ?annotation .
    ?annotation a up:Transmembrane_Annotation .
}
```

Gene and Sample annotation

Reasoning

A semantic reasoner, reasoning engine, rules engine, or simply a reasoner, is a piece of software able to infer logical consequences from a set of asserted facts or axioms. (Wikipedia)

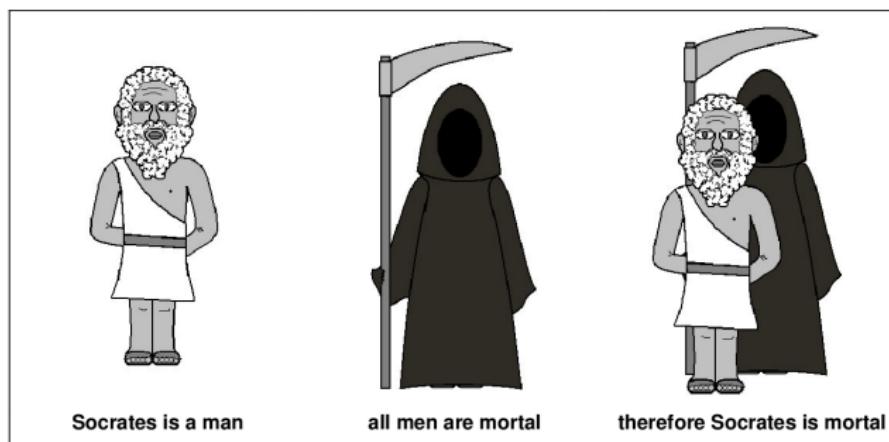


Figure 1 – deductive reasoning as propagated by Aristotle

Availability

GraphQL

<http://methada2020.uv.es:5555/graphql>

pyCOMPASS

<https://pypi.org/project/pyCOMPASS/>

Documentation

<https://command.readthedocs.io/>

<https://compass-.readthedocs.io/>

<https://pycompass.readthedocs.io/>

<https://vespucci.readthedocs.io/>