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

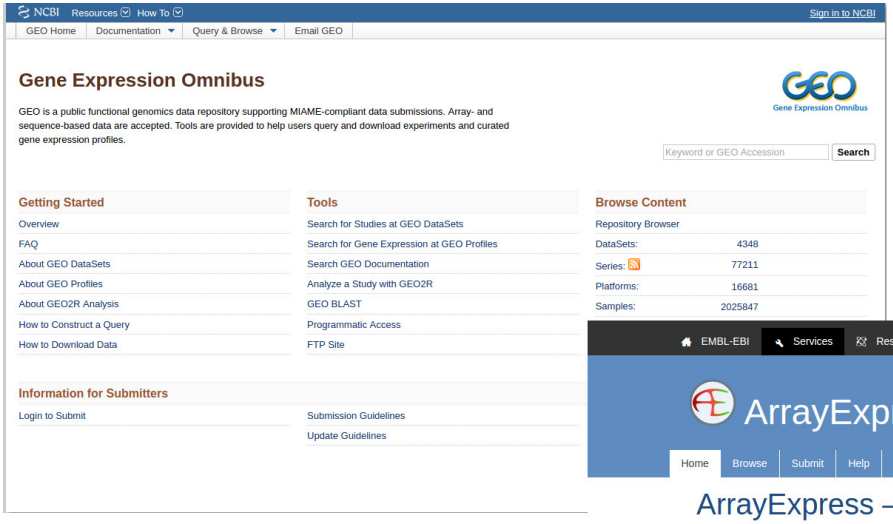
METHADA 2020
TRAINING SCHOOL IN TRANSCRIPTOMIC METADATA HANDLING & DATA ANALYSIS
COST Action CA17111 INTEGRAPE

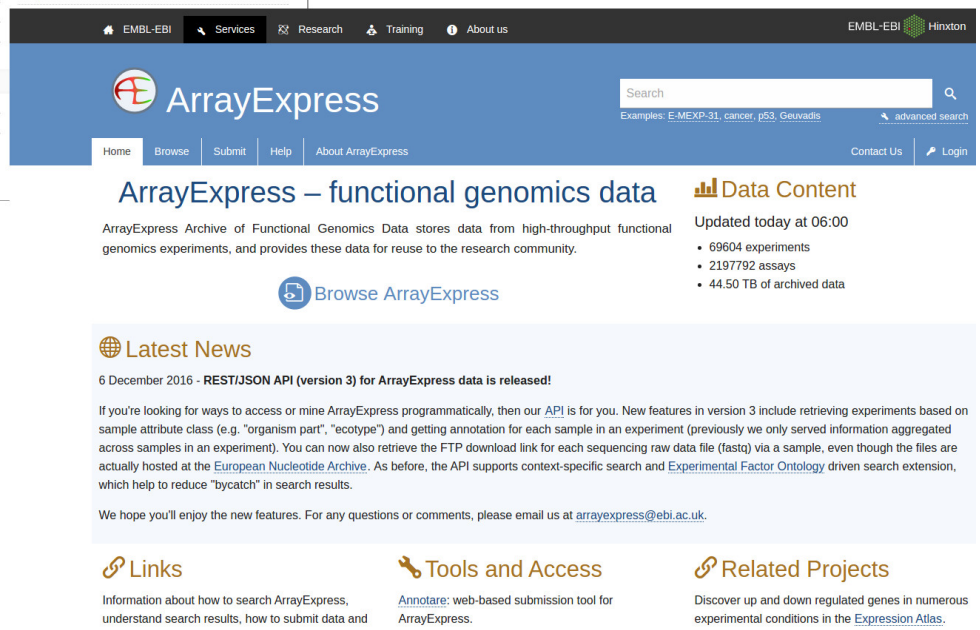
Acknowledgments

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

Why VESPUCCI?

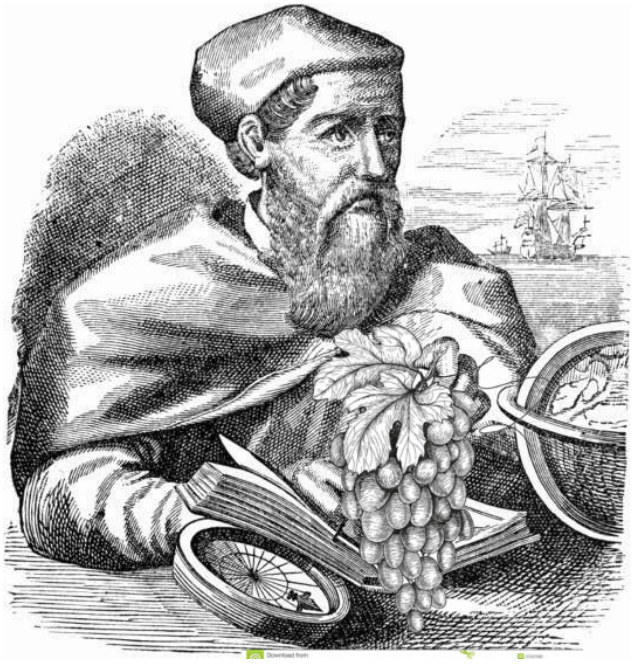
Thousands of transcriptome data sets in public databases





Why VESPUCCI?

Vitis
Expression
Studies
Platform
Using
COLOMBOS™
Compendia
Instances

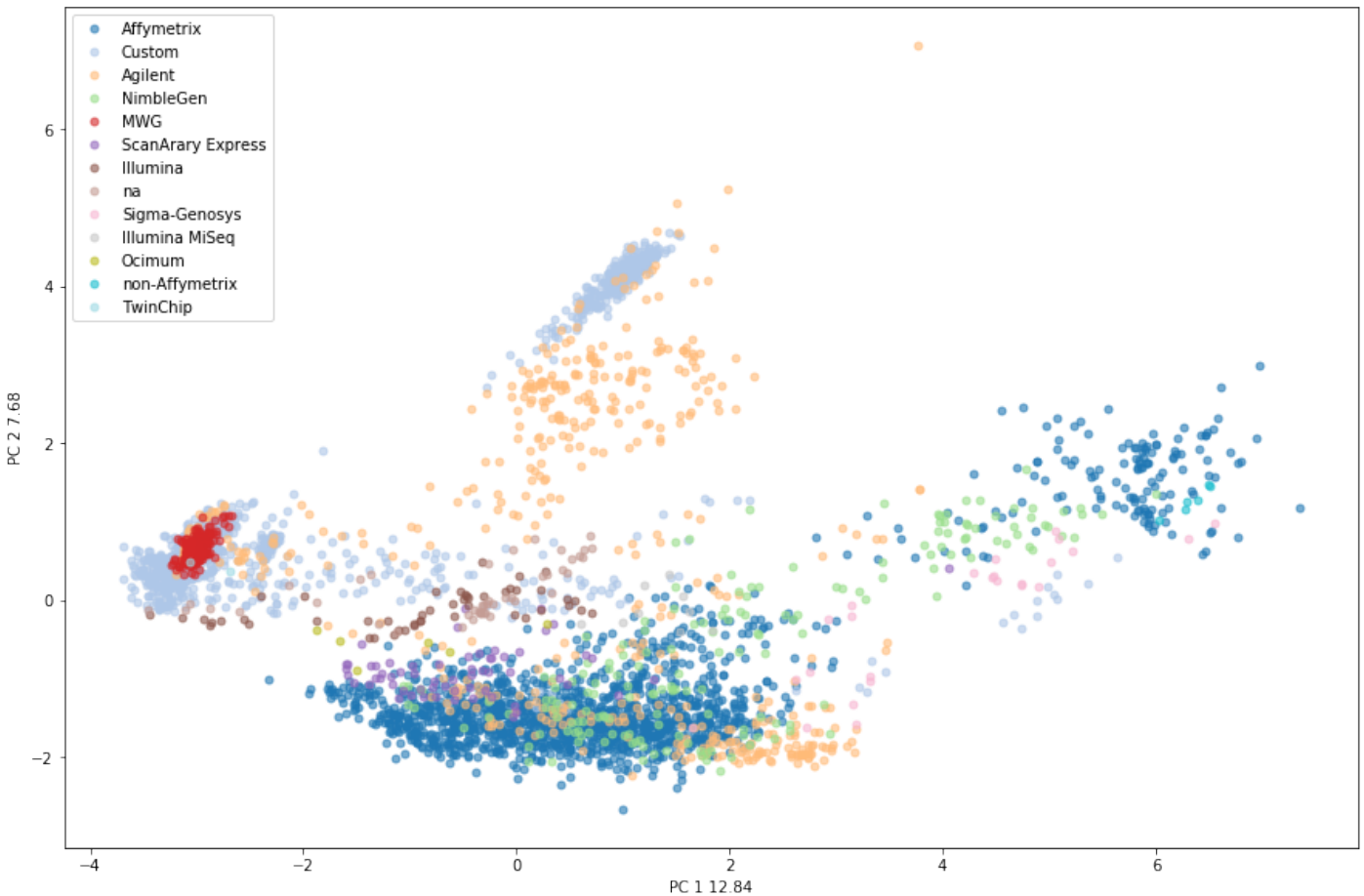


What is VESPUCCI? (In a nutshell)

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

Why VESPUCCI?

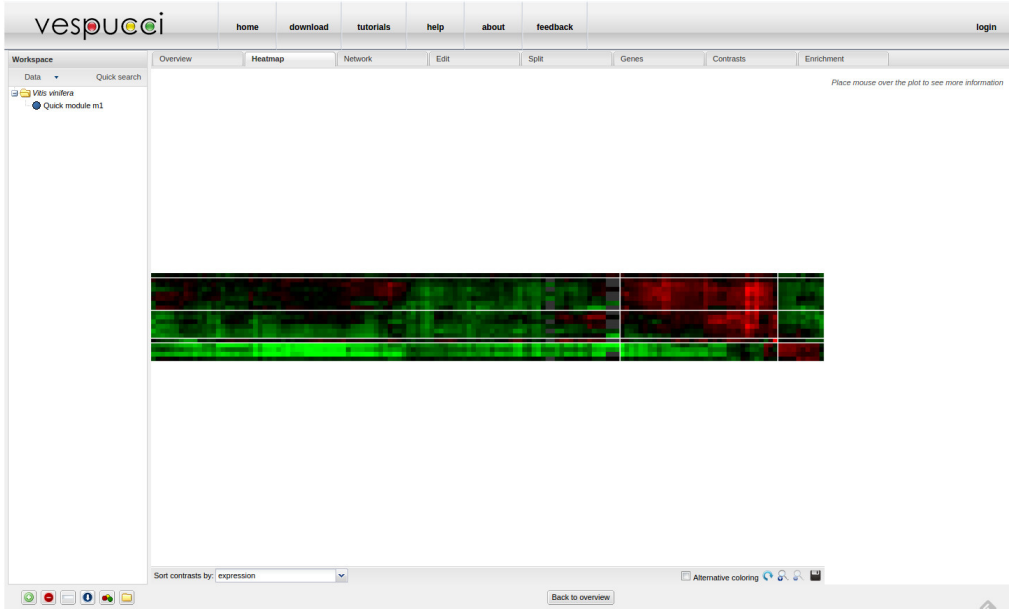
Batch effect



Why VESPUCCI?

Ocean exploration analogy

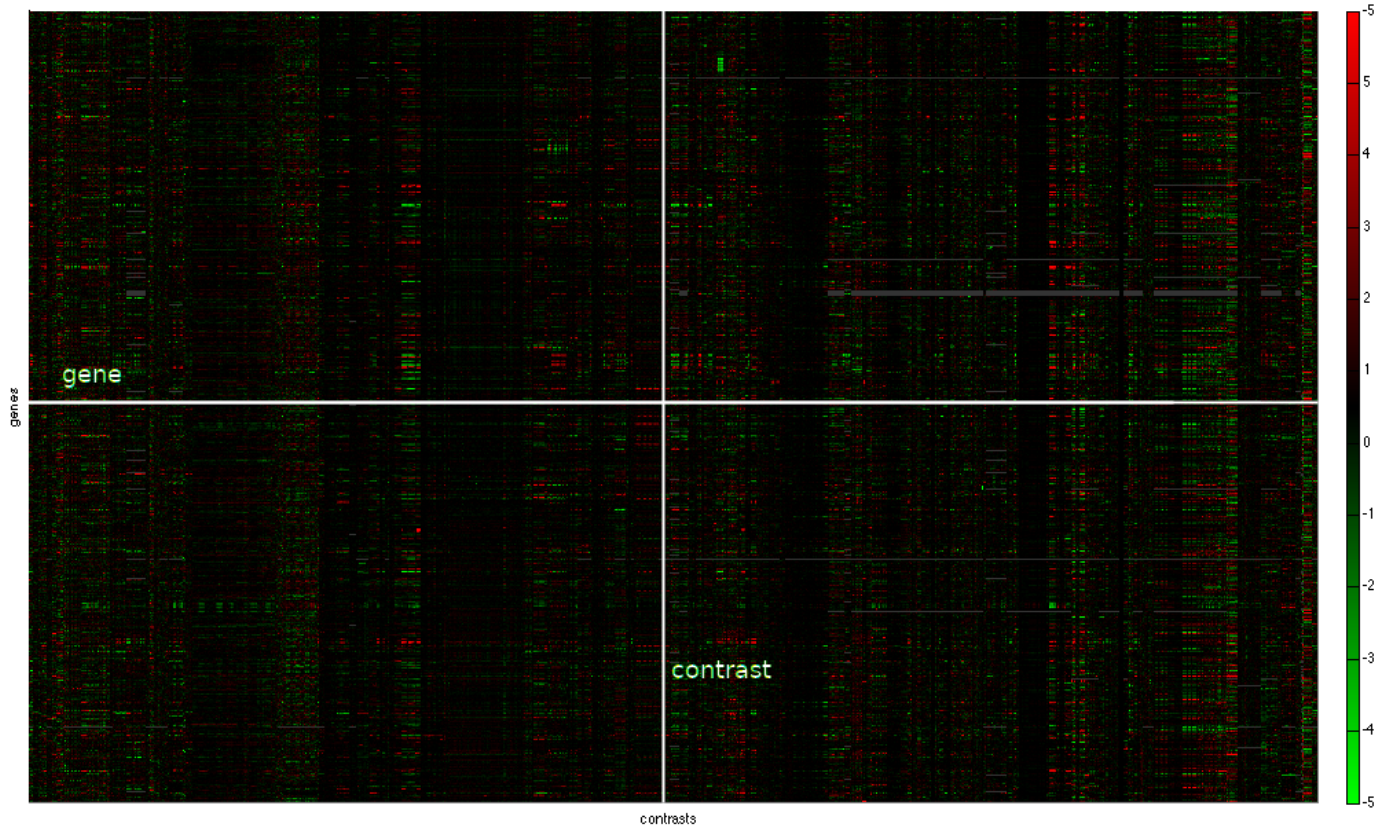
- Sea/Ocean
- Atlas/Map
- Path/Route
- Navigation tools



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

What is VESPUCCI?

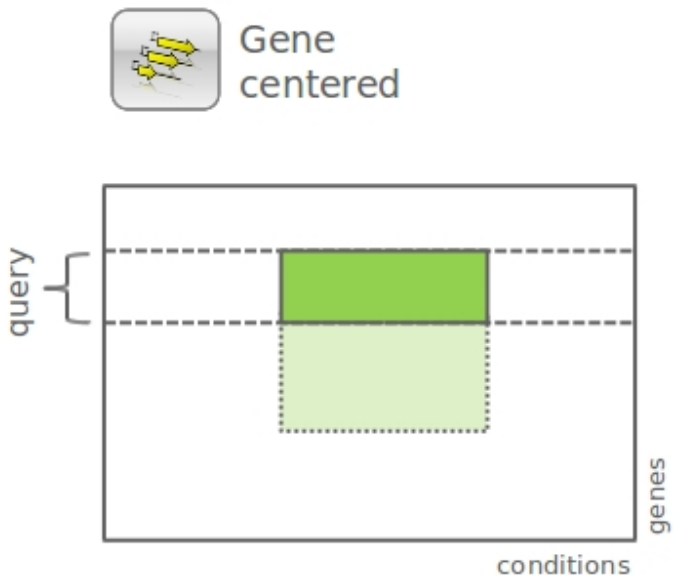
Single coherent expression matrix in which each **row** represents a **gene** and each **column** represents a sample **contrast** (i.e. relative value given by the difference in log-scale between two conditions, a test and a reference).

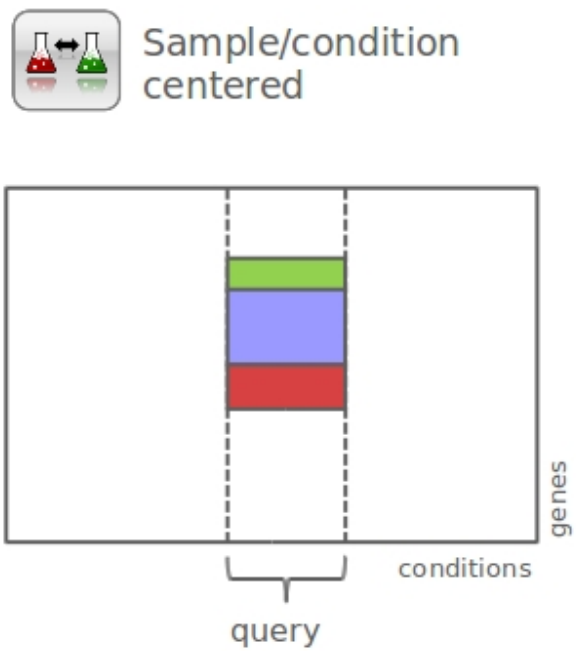


What is VESPUCCI?

Module

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





Updates

Technology

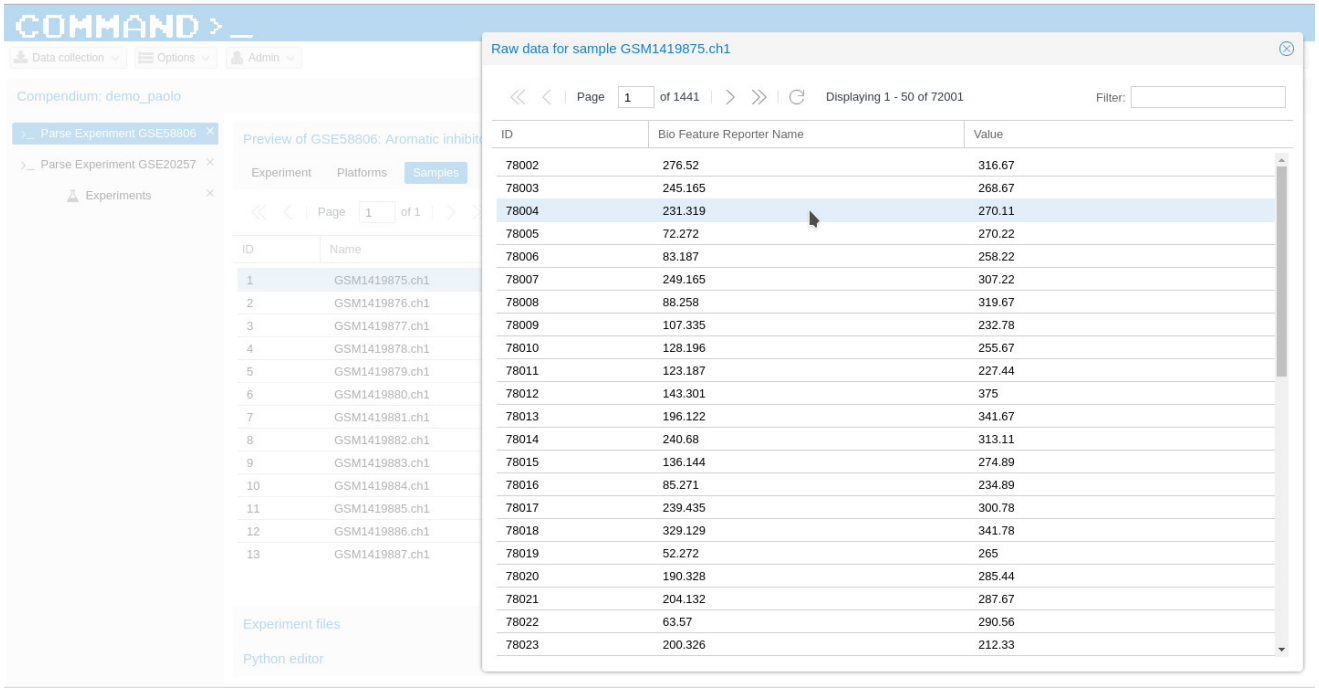
- Back-end (COMMAND>_)
- Normalizations
- Annotation system
- Front-end (COMPASS)

Data

- All transcriptomics experiments up until September 2019

How VESPUCCI is built

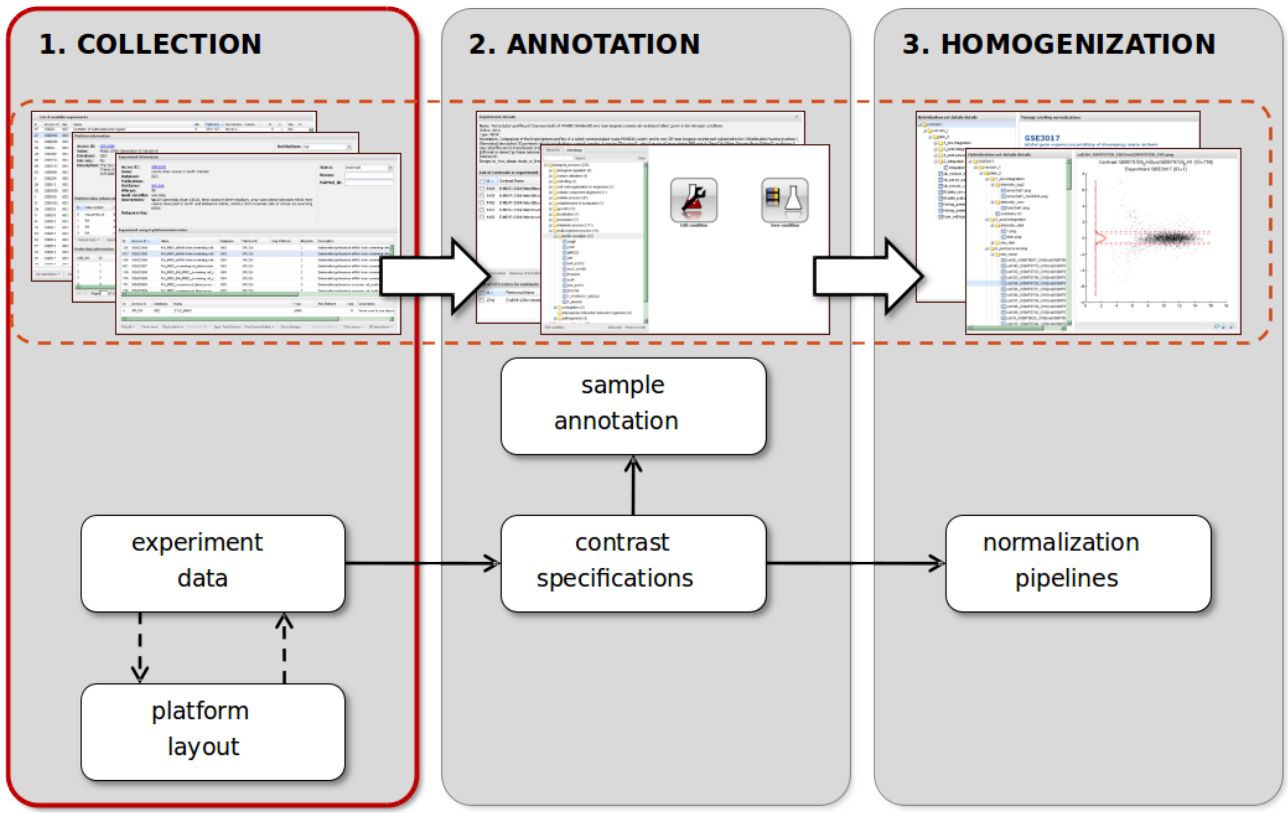
COMMAND>_



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

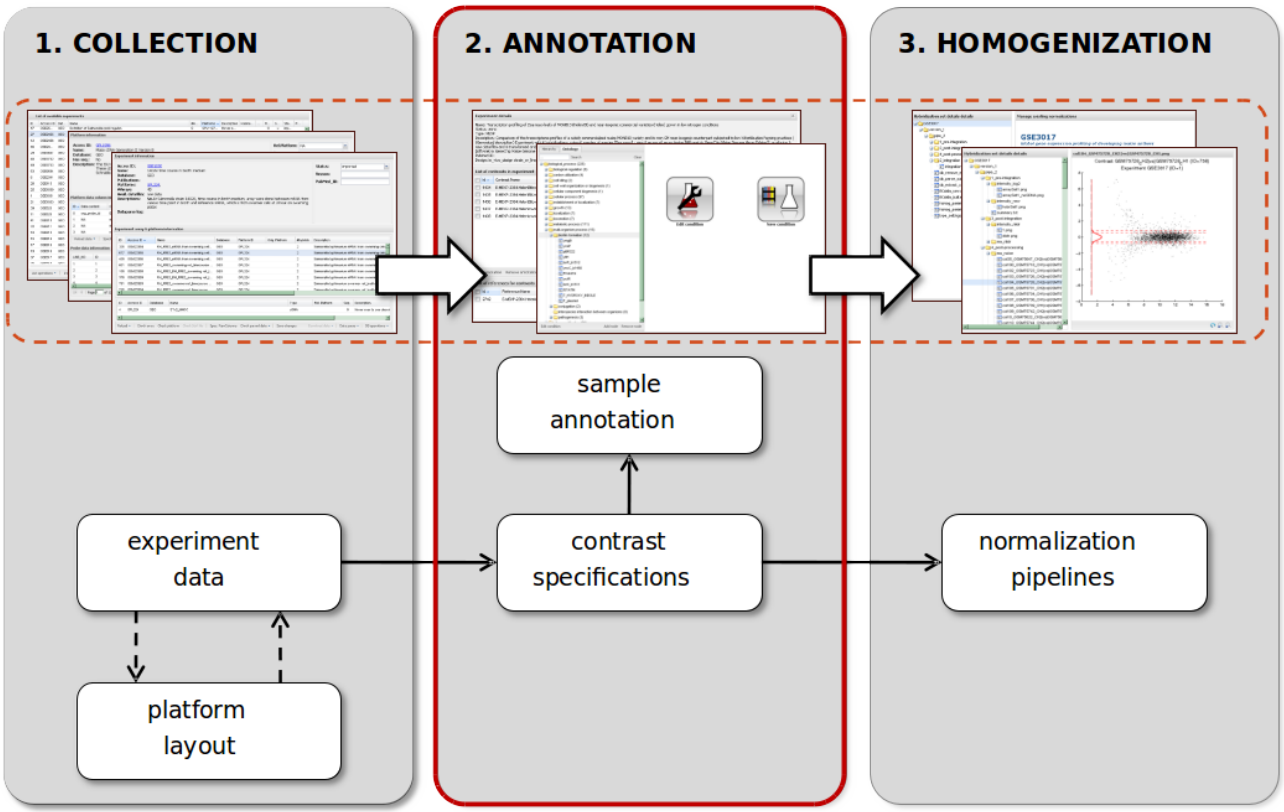
How VESPUCCI is built

COMMAND>_



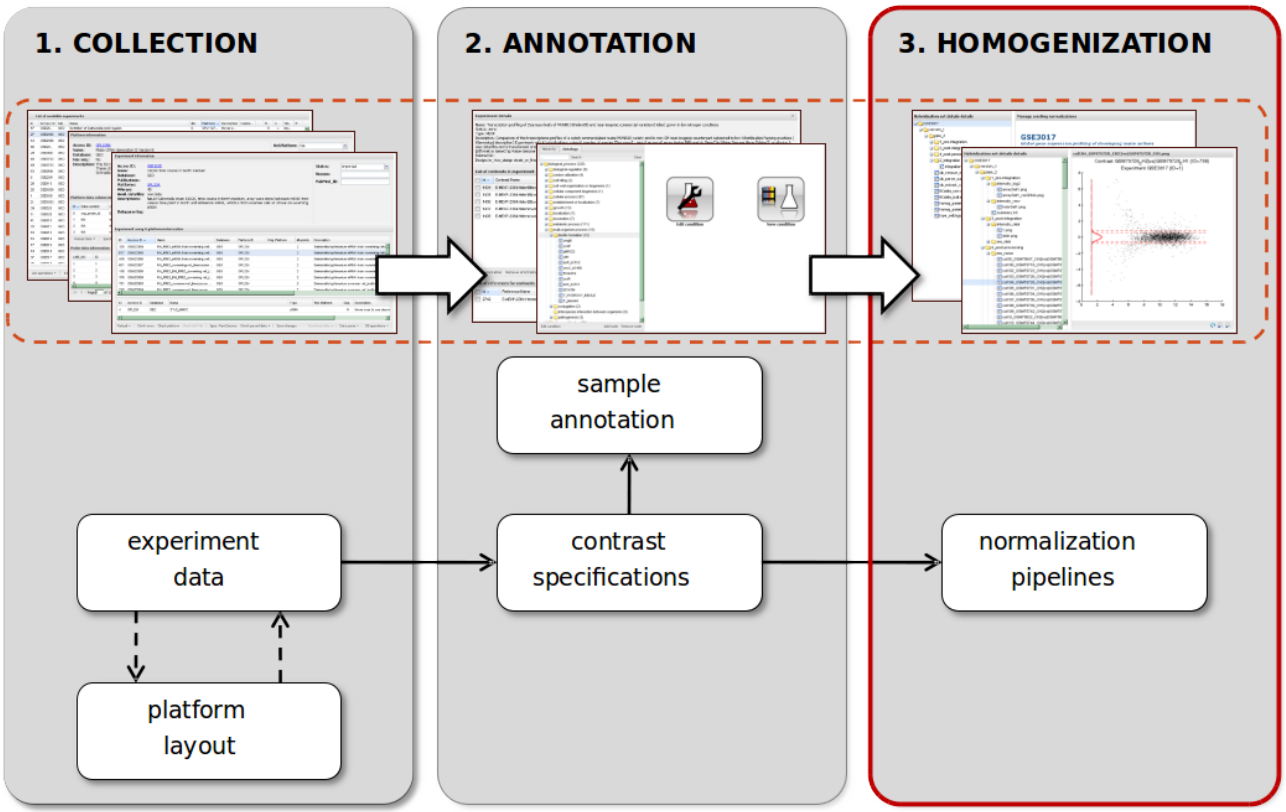
How VESPUCCI is built

COMMAND>_



How VESPUCCI is built

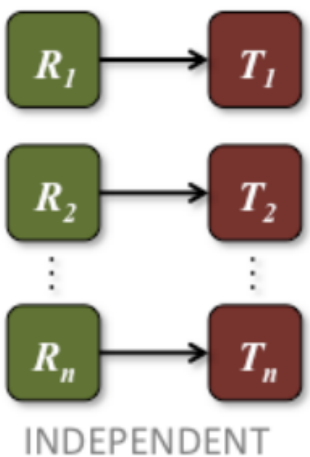
COMMAND>_

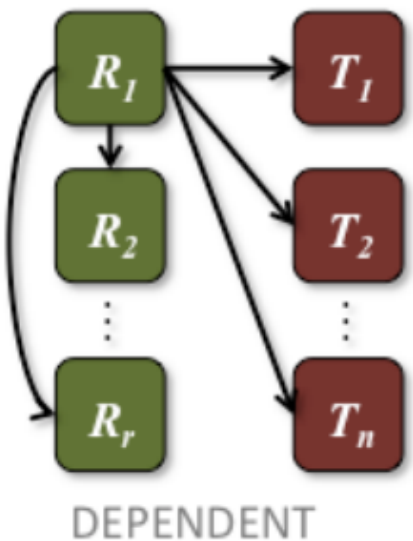


Normalization

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

Legacy log-ratio

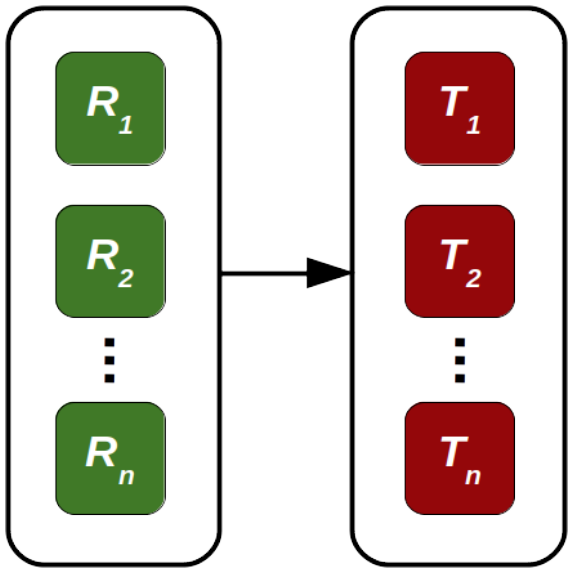




Normalization

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

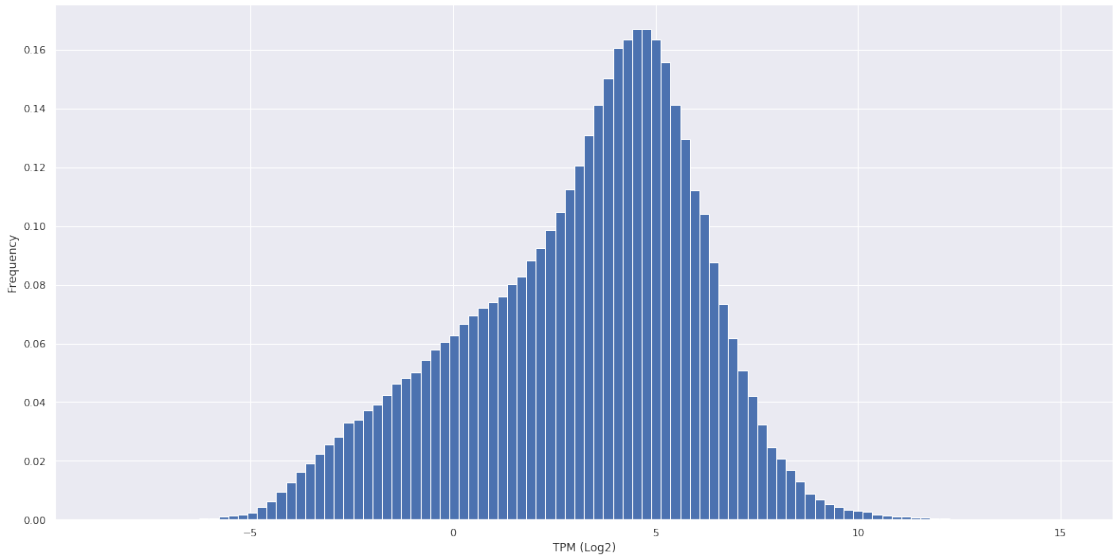
LIMMA



Normalization

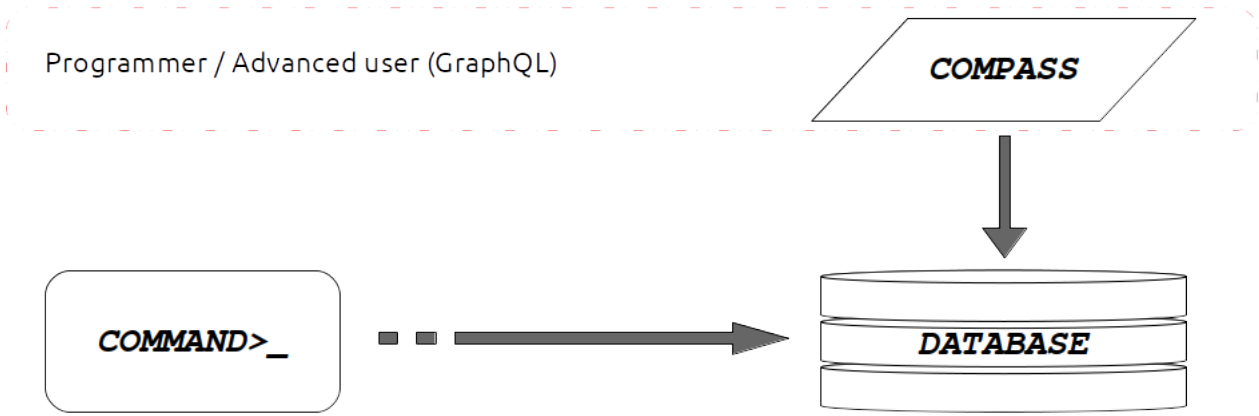
The new technology supports different **normalization** strategies within the same compendium.

TPM - RNA-seq only

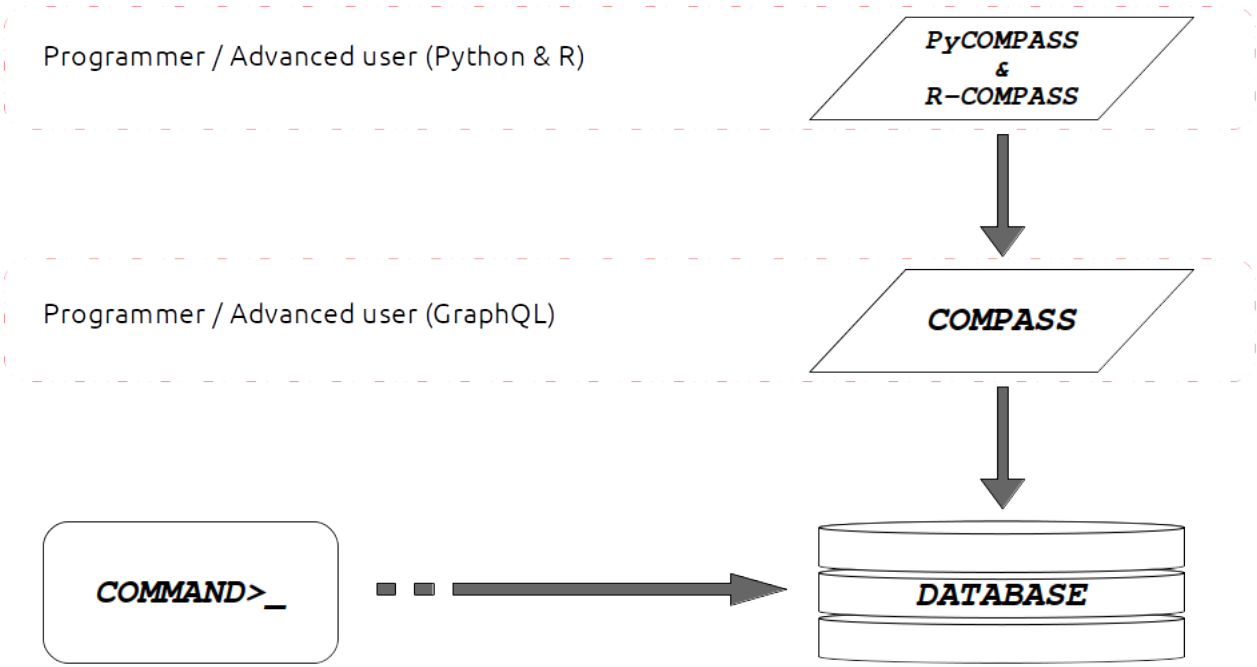


Accessing the data - the front-end

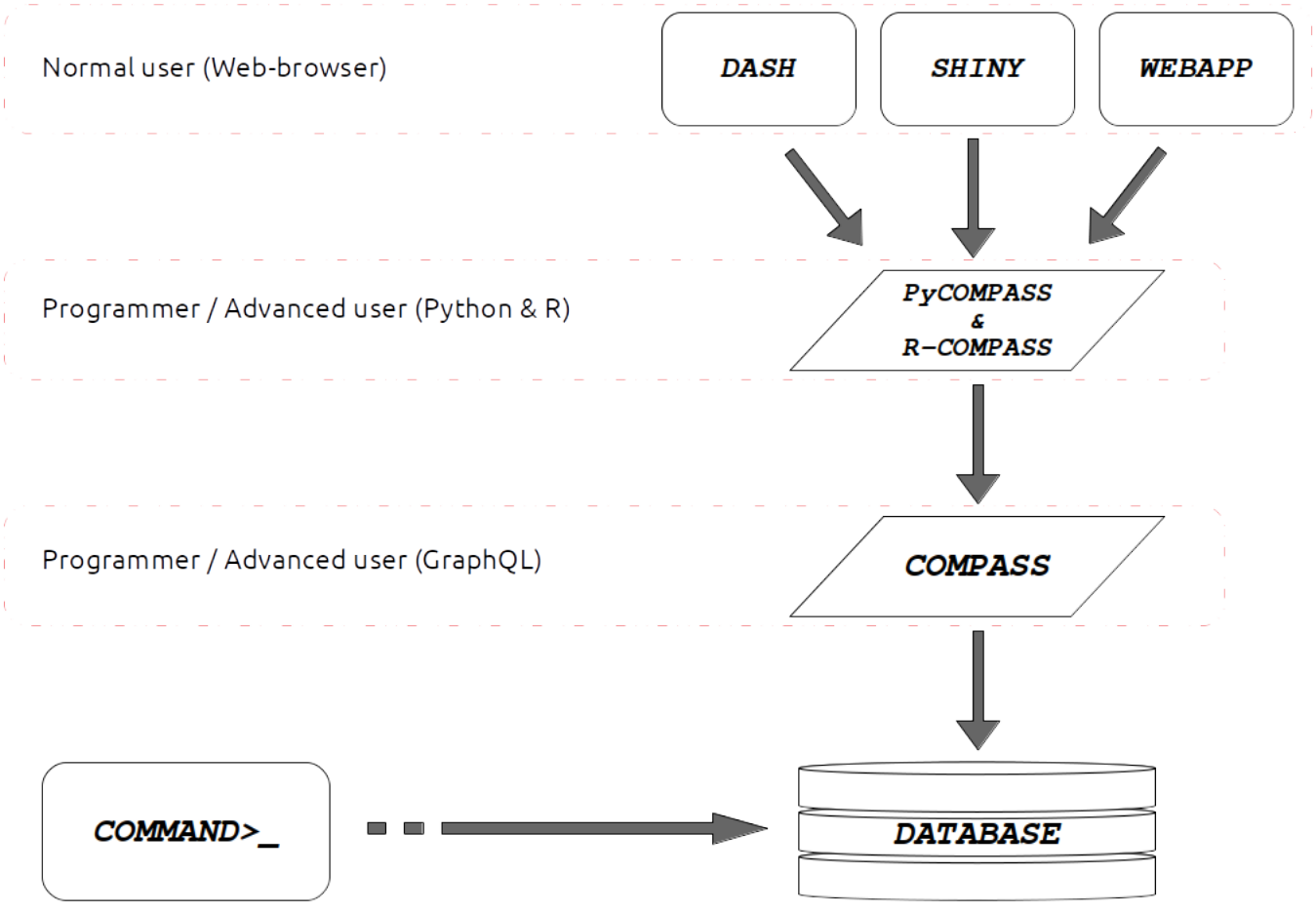
Accessing the data - the front-end



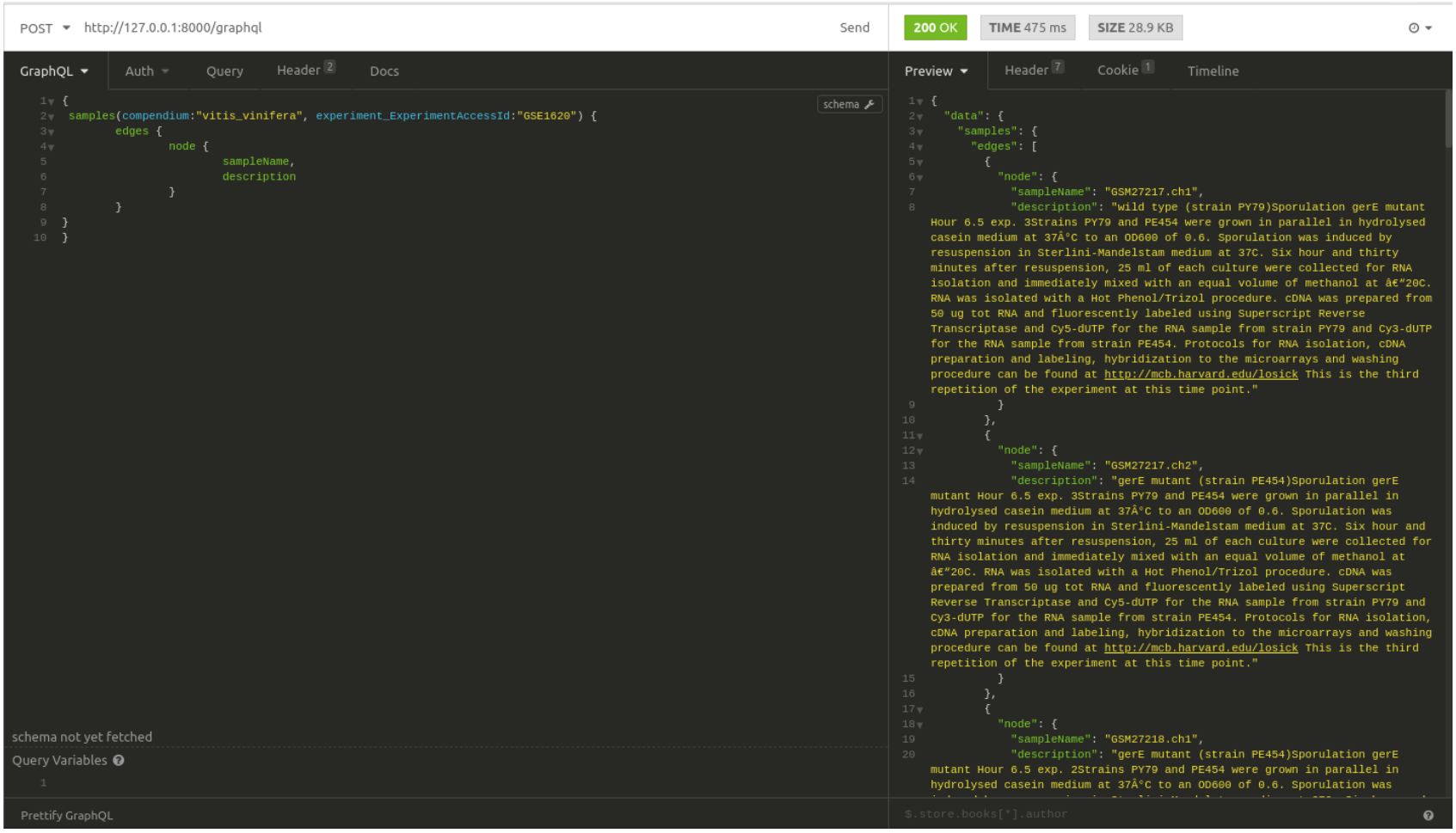
Accessing the data - the front-end



Accessing the data - the front-end



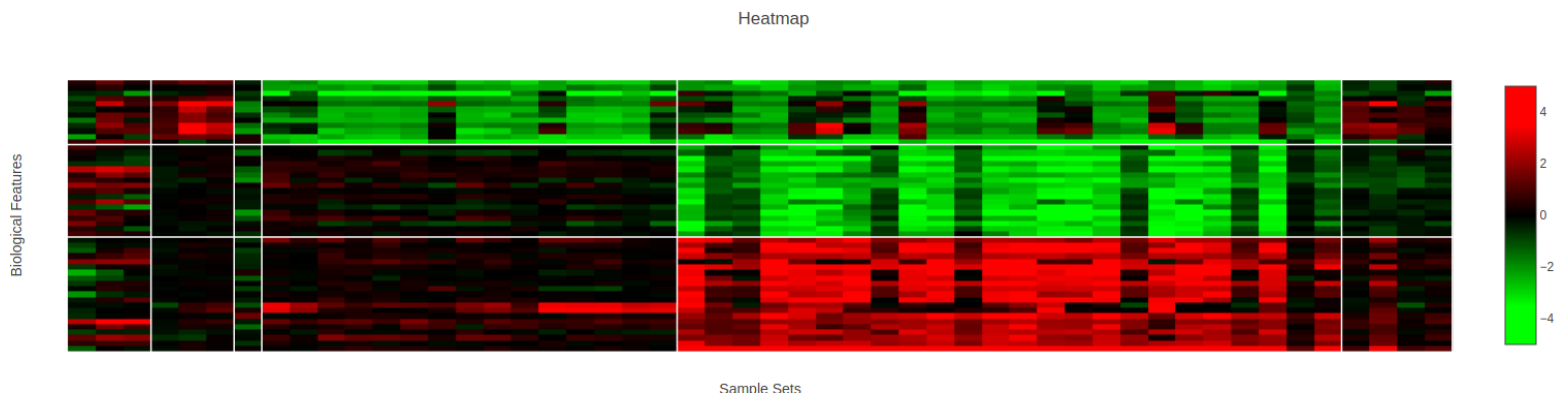
Accessing the data - the front-end: COMPASS the GraphQL interface



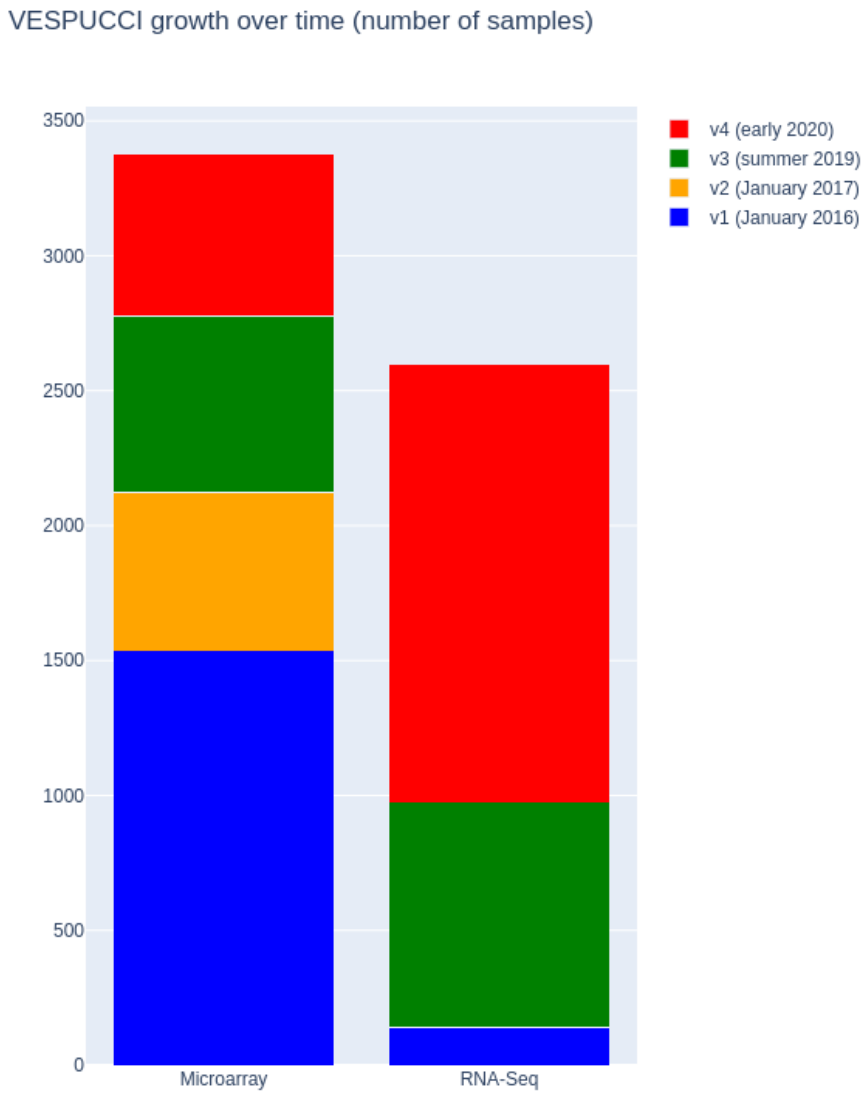
Accessing the data - the front-end: pyCOMPASS the Python interface

[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)
3500
3000
2500
2000
1500
1000
500
0
Microarray
RNA-Seq
v4 (early 2020)
v3 (summer 2019)
v2 (January 2017)
v1 (January 2016)

Gene and Sample annotation

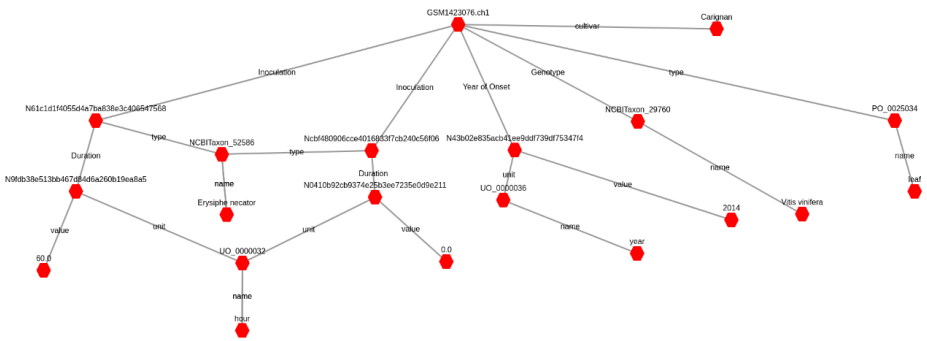
- From **home-made vocabulary** to standard **Ontologies**
- From *"bag of words"* to **RDF** triples

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

Gene and Sample annotation

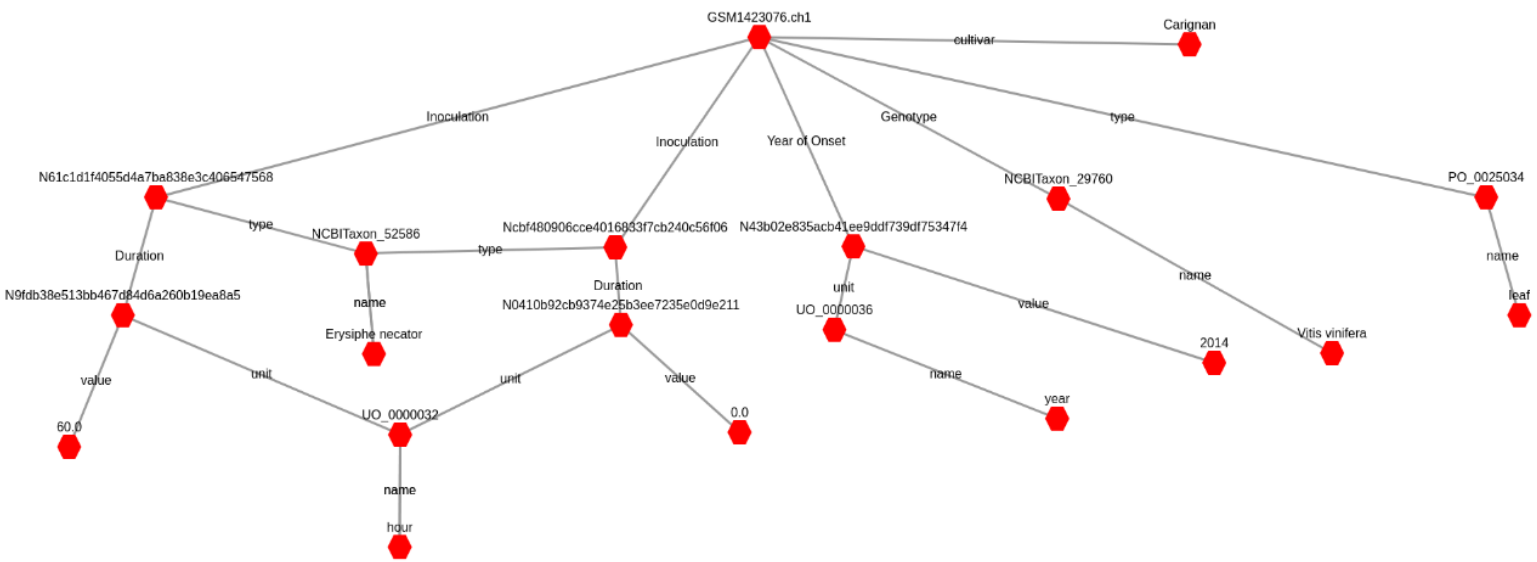
- From **home-made vocabulary** to standard **Ontologies**
- From *"bag of words"* to **RDF** triples

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)

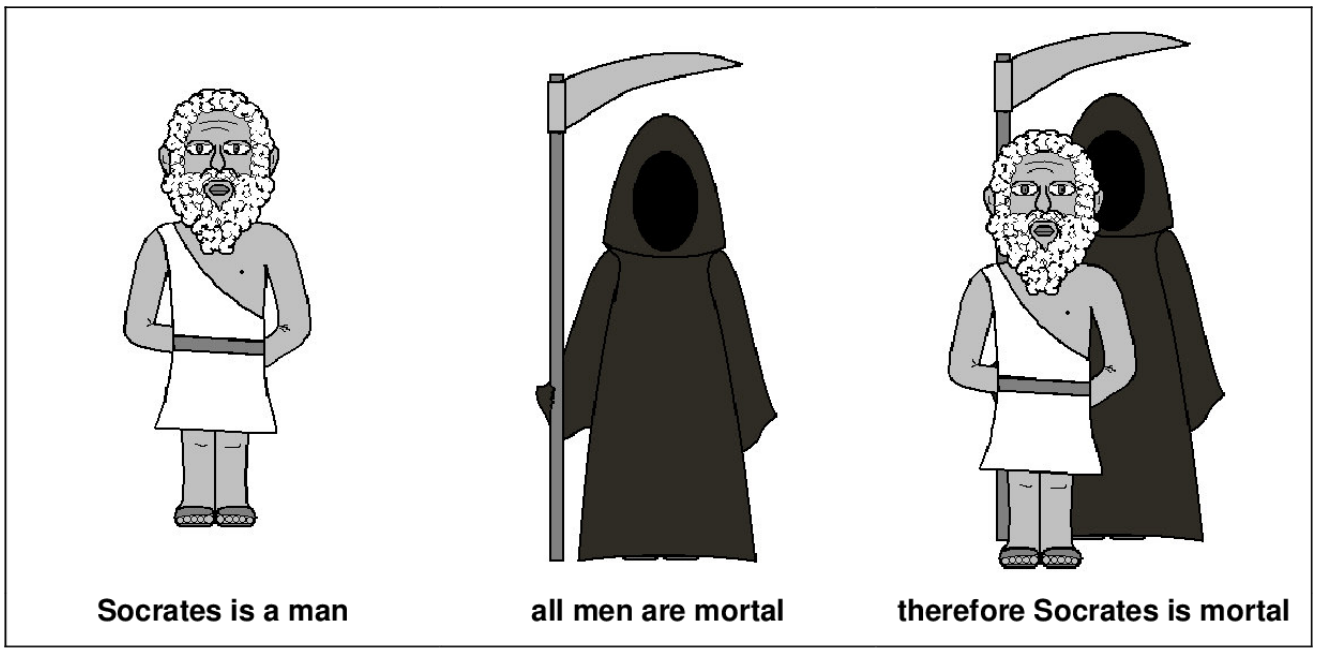


Figure 1 – deductive reasoning as propagated by Aristotle

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/`