FONDAZIONE
EDMUND
MACH

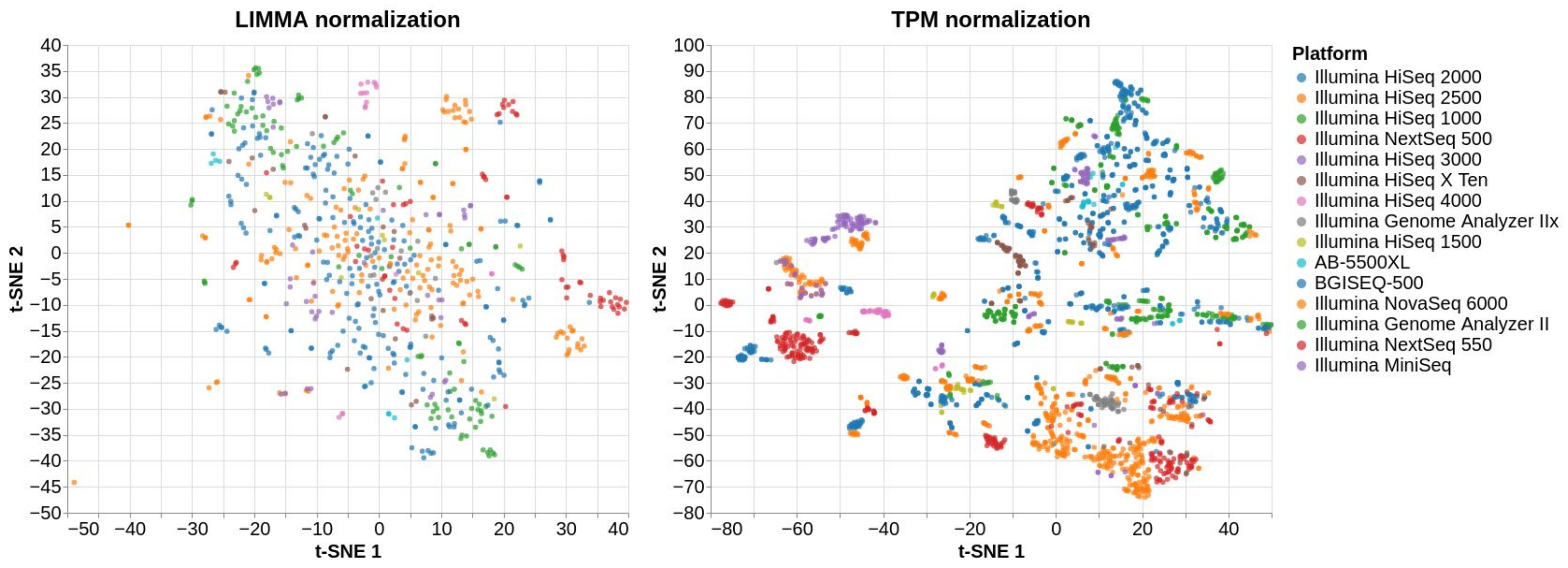
A COMPASS for VESPUCCI: a FAIR way to explore the grapevine transcriptomic landscape

Marco Moretto[1], Paolo Sonego[1], Stefania Pilati[2], José Tomás Matus[3], Laura Costantini[2], Giulia Malacarne[2] and Kristof Engelen[1]

1 Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach Via E. Mach 1, 38010 San Michele all'Adige, Italy
2 Department of Genomics and Biology of Fruit Crops, Research and Innovation Centre, Fondazione Edmund Mach Via E. Mach 1, 38010 San Michele all'Adige, Italy
3 Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Paterna, 46908, Valencia, Spain

Introduction

The problem of data integration concerns merging data coming from several sources while providing the user with a unique way to access and retrieve them. VESPUCCI, the integrated database of gene expression data for grapevine, has been updated to be FAIR-compliant, employing standards and created with open-source technologies. It includes all public grapevine gene expression experiments from both microarray and RNA-seq platforms.



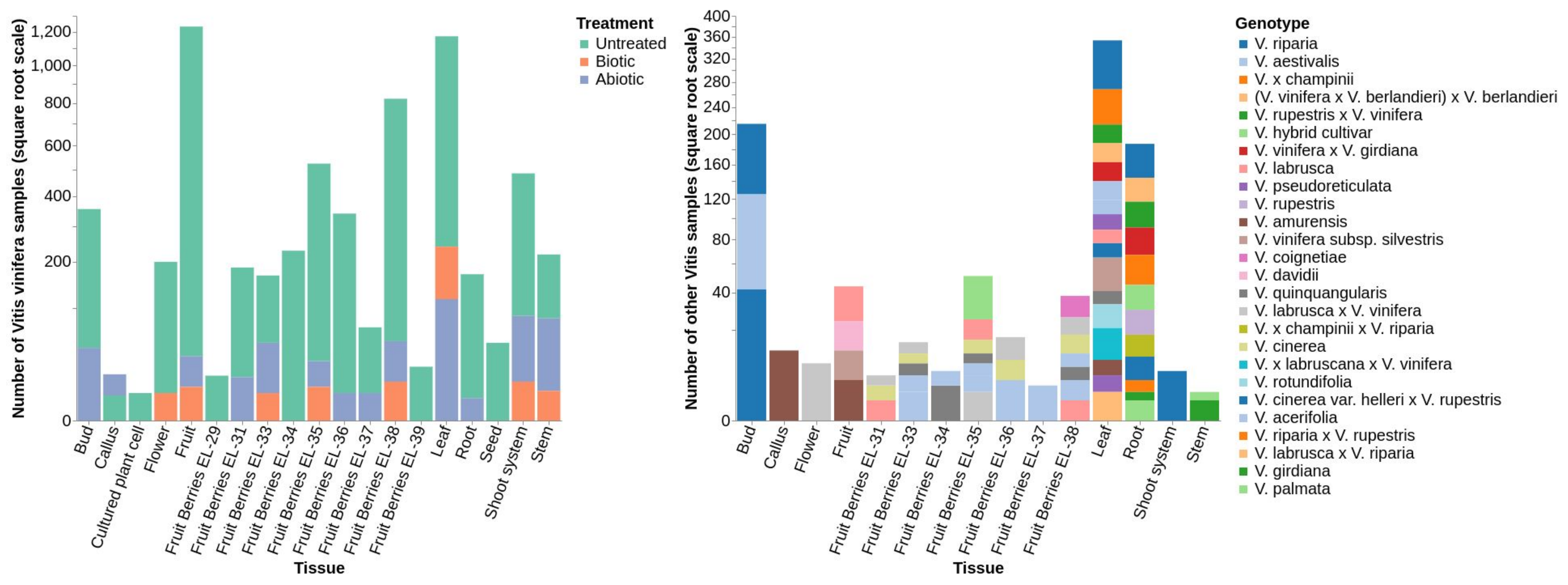
Data normalization and sample annotation

VESPUCCI v2 provides data normalized with two different approaches: logratios through the LIMMA package and TPM (Transcript Per Million) values, for RNA-seq only.

We adopted the RDF data model and the use of ontologies to formally describe sample conditions.

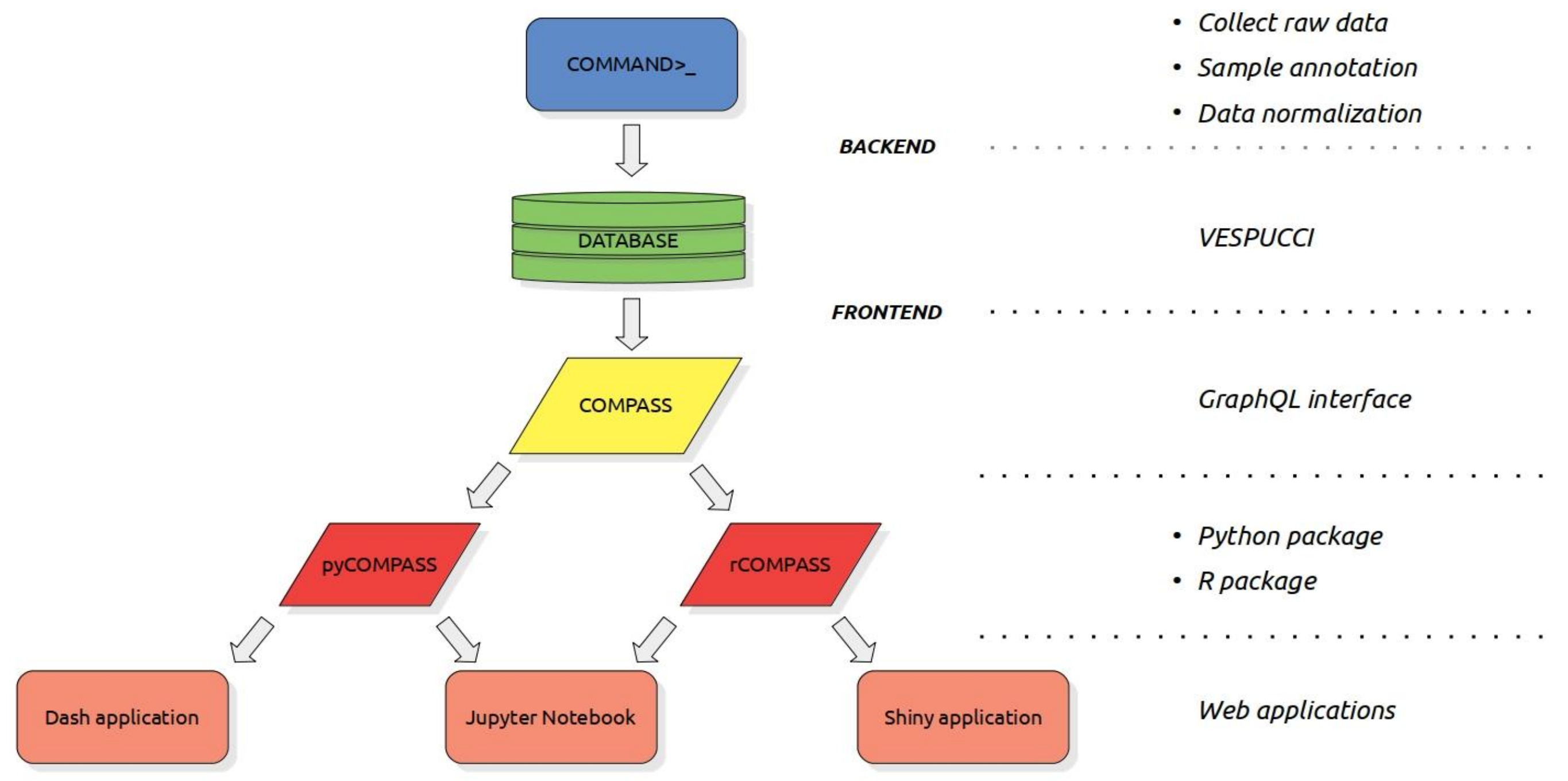
Accessing VESPUCCI using COMPASS

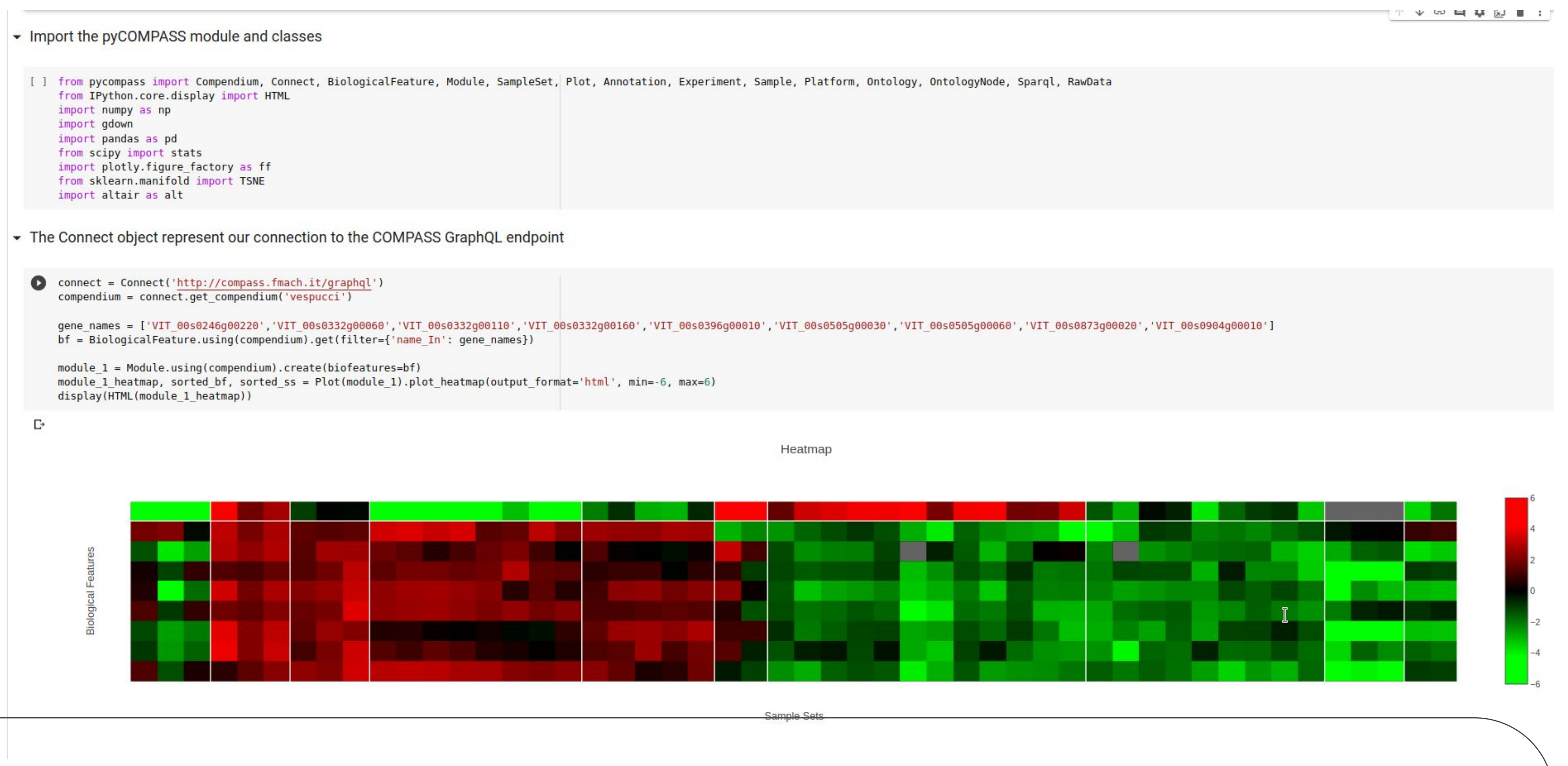
COMPASS (COMpendia Programmatic Access Support Software) is the frontend application and the main interface to access VESPUCCI. It is a GraphQL interface that allows querying every part of the VESPUCCI data model. pyCOMPASS and rCOMPASS are the two high-level packages (for Python and R respectively) built on top of the COMPASS interface that simplify the access and analysis of VESPUCCI's data. Last layer is composed of all the applications that rely on these packages such as GUI applications (like Dash or Shiny applications) as well as analysis workflows such as Jupyter Notebook and R markdown that further simplify the interactions with the database.
The multi-tier architecture separates the compendium (VESPUCCI) from the programmatic interface (COMPASS) making i embed the data in third party services or use them in any analysis workflow, enhancing reproducibility of results and interoperability of different resources.



Data content update

The VESPUCCI v2 compendium is a comprehensive database of nearly all transcriptomic experiments performed on grapevines during the last 15 years. It contains 3682 microarray and 3598 RNA-seq samples across 271 experiments collected until December 2020. The great majority of samples (47%) come from Vitis vinifera untreated fruit samples taken at different developmental stages. Non-vinifera species and hybrids samples represent 13% of the dataset while nearly 9% of the total are stress-related (2.5% being fruit) and 31% are Vitis vinifera untreated non-fruit samples




Import the pyCOMPASS module and classes

```
[ ] from pycompass import Compendium, Connect, BiologicalFeature, Module, SampleSet, Plot, Annotation, Experiment, Sample, Platform, Ontology, OntologyNode, Sparql, RawData
    from IPython.core.display import HTML
    import numpy as np
    import gdown
    import pandas as pd
    from scipy import stats
    import plotly.figure_factory as ff
    from sklearn.manifold import TSNE
    import altair as alt
```

The Connect object represent our connection to the COMPASS GraphQL endpoint

```
connect = Connect('http://compass.fmach.it/graphql')
compendium = connect.get_compendium('vespucci')

gene_names = ['VIT_00s0246g00220','VIT_00s0332g00060','VIT_00s0332g00110','VIT_00s0332g00160','VIT_00s0396g00010','VIT_00s0505g00030','VIT_00s0505g00060','VIT_00s0873g00020','VIT_00s0904g00010']
bf = BiologicalFeature.using(compendium).get(filter={'name_In': gene_names})

module_1 = Module.using(compendium).create(biofeatures=bf)
module_1_heatmap, sorted_bf, sorted_ss = Plot(module_1).plot_heatmap(output_format='html', min=-6, max=6)
display(HTML(module_1_heatmap))
```

Conclusion and future perspectives

VESPUCCI v2 attempts comply with the FAIR principales by providing a single point of access to the database via a web server GraphQL interface, a manually curated annotation of experimental conditions using ontologies and RDF as data models, and software packages for the two most widely used programming languages in data analysis, Python and R, to enhance the interoperability of VESPUCCI v2 with other resources and tools. uture versions of VESPUCCI will include, together with newly available experiments, SNPs marker information and new normalization techniques, as well as using the upcoming VCost.v4 gene annotation alongside the current versions

marco.moretto@fmach.it

http://compass.fmach.it