

Data Lakehouse to support the development of AI models for predicting patient clinical response to targeted and immuno-therapies

Elodine COQUELET¹, Marta SILVA⁴, Laura BALBI⁴, Jofre RIBA⁵, Javier ALFARO², Fabio Massimo ZANZOTTO³, Catia PESQUITA⁴, Rohit KUMAR⁵ and Christophe BATTAIL¹

¹ Université Grenoble Alpes, IRIG, Laboratoire Biosciences et Bioingénierie pour la Santé, UA 13 INSERM-CEA-UGA, 38000 Grenoble, France.

² International Center for Cancer Vaccine Science, University of Gdansk, Poland

³ University of Rome Tor Vergata, Italy

⁴ LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

⁵ Fundacio Eurecat, Spain

contact: christophe.battail@cea.fr; elodine.coquelet@gmail.com

Abstract

In the context of the European project KATY on precision medicine, we prototyped a **Data Lakehouse** by integrating research studies that generated molecular profiling data from cohorts of kidney tumor tissues taken from patients included in drug clinical trials. Indeed, there is currently a lack of a database dedicated to support the development of AI models to help doctors in choosing the best drug for each patient.

The Data Lakehouse architecture, which we have implemented with **open source Delta Lake** technology, brings together the best features of Data Lake and Data Warehouse.

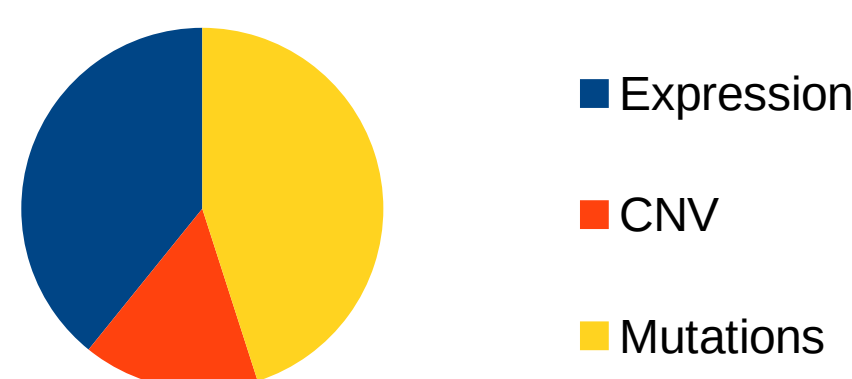
The Data Lakehouse will allow three types of access for the KATY consortium members:

- the implementation of **data analytics** approaches to query and visualize molecular and clinical data
- the targeted extraction of data for the **training and testing of AI models**
- feeding a **Knowledge Graph** to support the explainability of the predictive models using a priori biological and clinical knowledge.

Data Lakehouse architecture

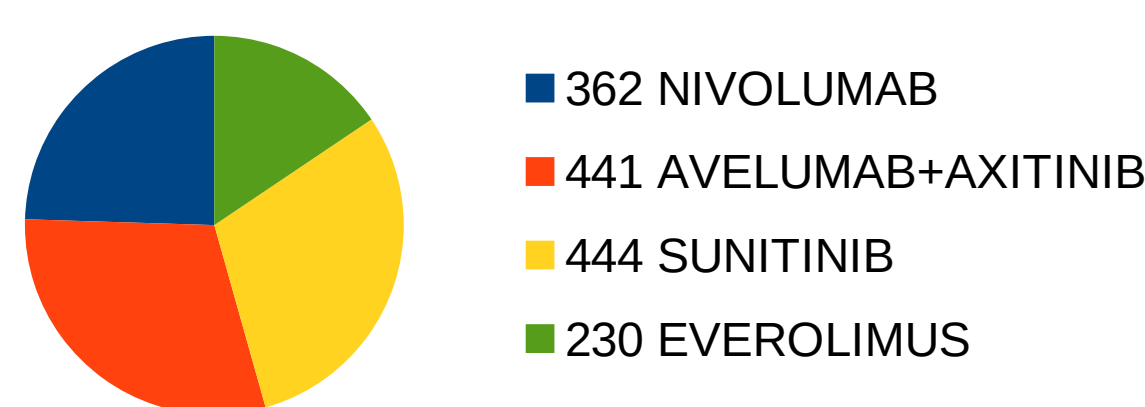
Repartition of experiments

2287 Samples

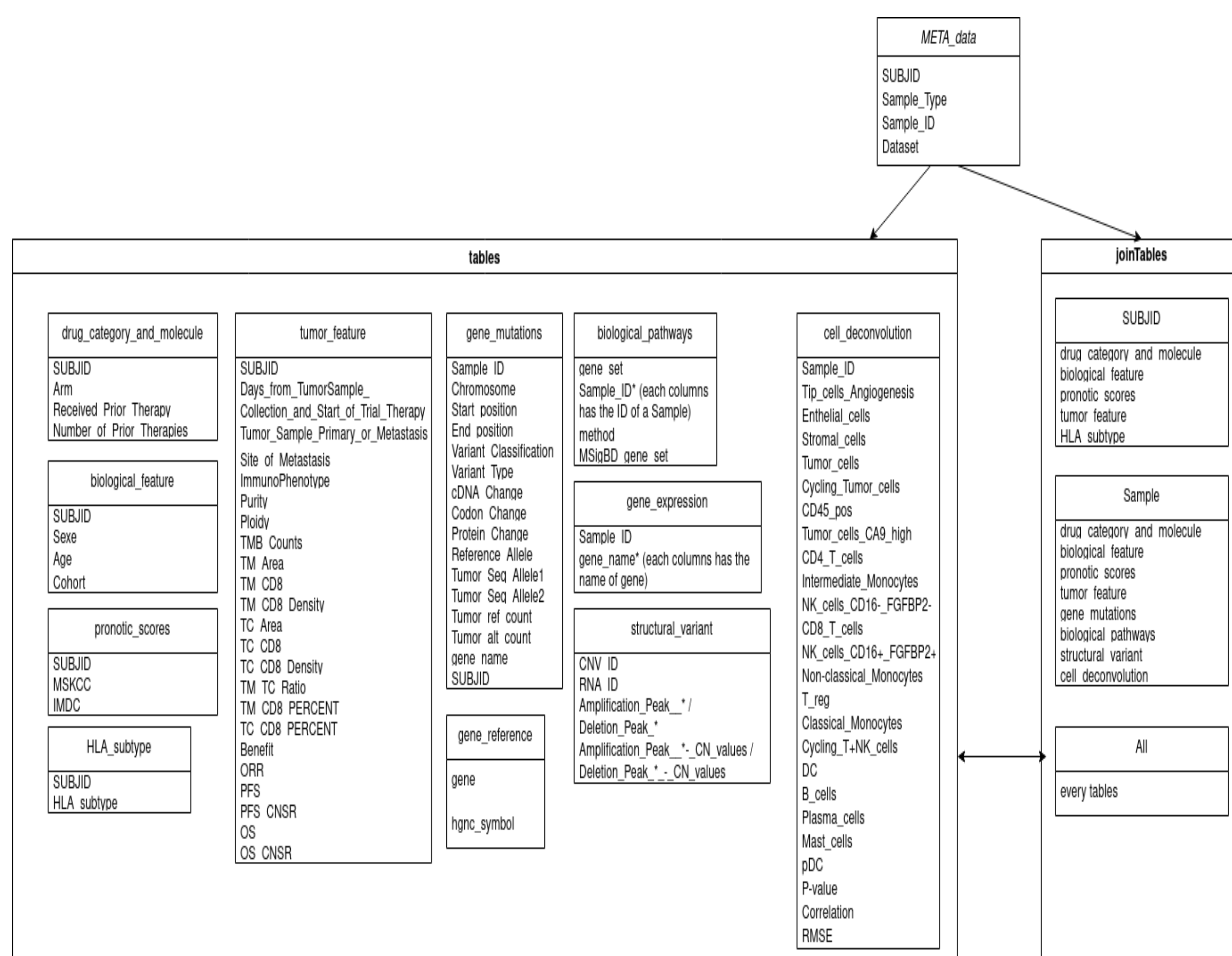


Repartition of Treatments

1478 Subjects



[3][4]



Data Lakehouse technology

A **Data Lakehouse** is a data model close to a Data Warehouse allowing to store structured and unstructured data, of different types and from different sources in a non-hierarchical way. Unlike a Data Warehouse, a Data Lakehouse contains the data needed to solve a **single problem**. Compared to the Data Lake, the Data lakehouse adds a **layer of metadata and governance** to secure and finely control access to data.

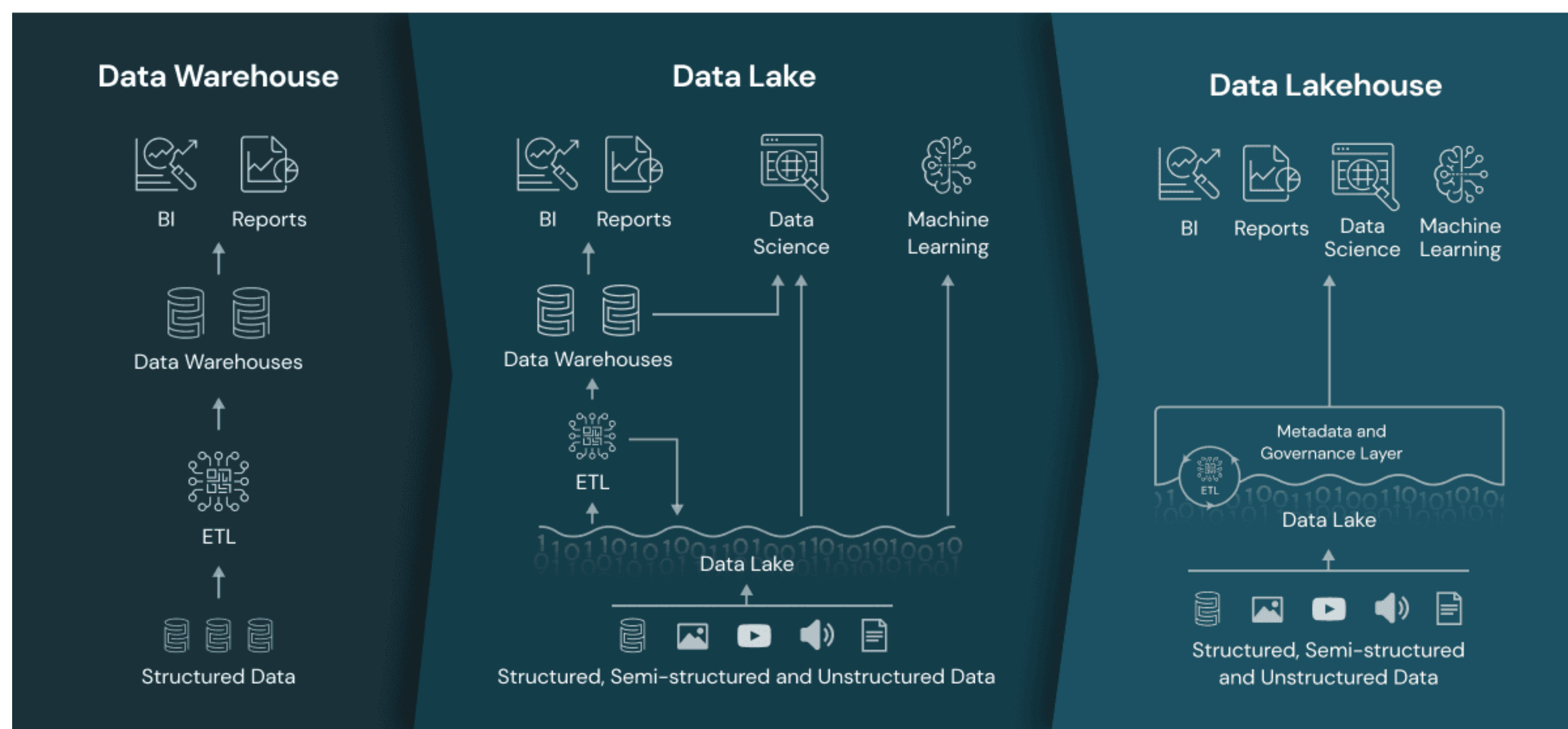
The **distributed infrastructure** of the Data Lakehouse, both in terms of storage and calculation, is suitable for managing large data.

The Data Lakehouse was developed with the **open source Delta Lake** technology built on an **Apache Spark** layer. The security and data governance aspects will use the **Apache Atlas and Ranger** software libraries.

Design strategies have been put in place to **minimize latency of the queries** and to **facilitate the update of the database with new datasets**.

A work in progress of **semantic ontologies** harmonizes and controls the vocabulary used in the database.

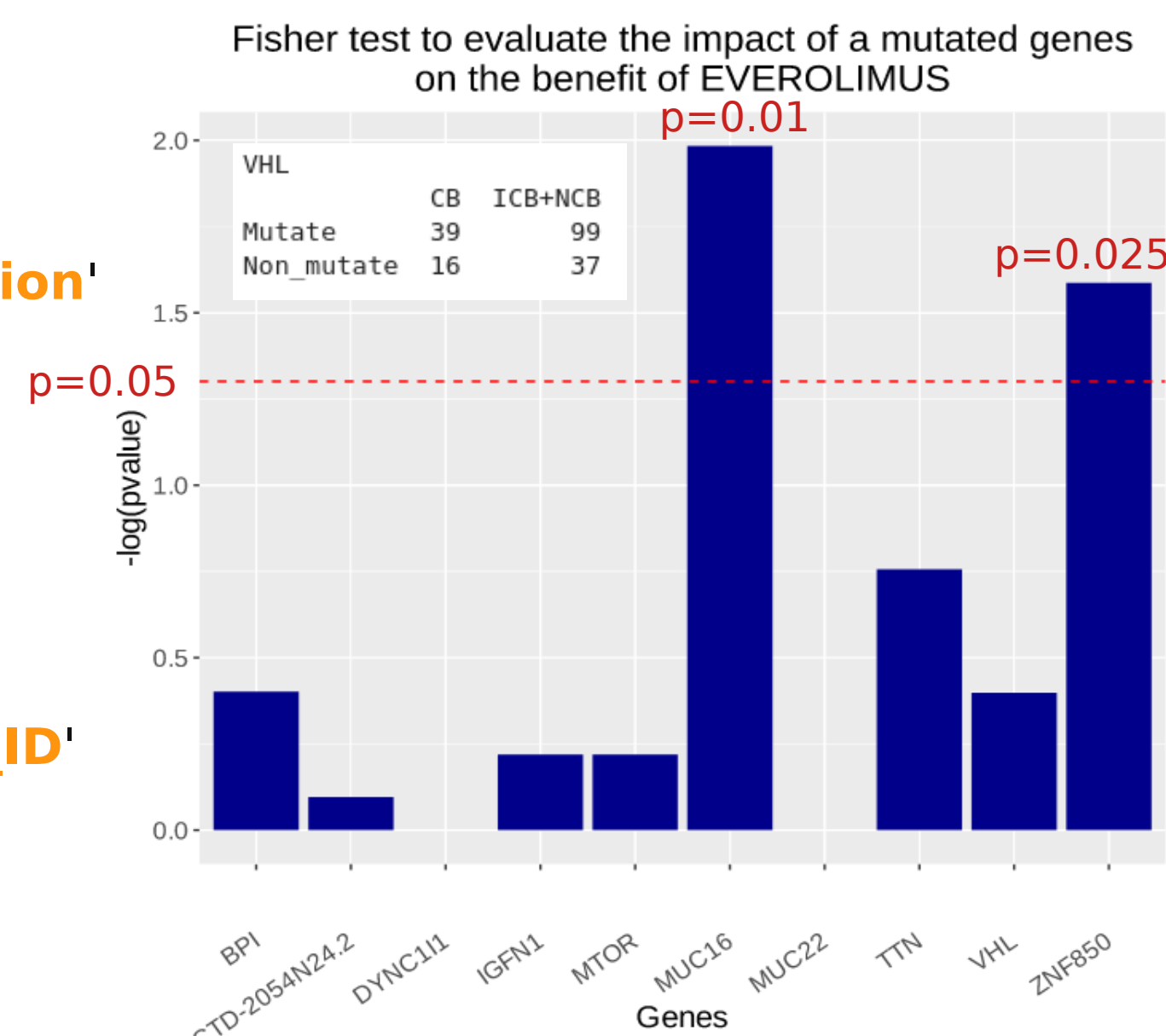
[1][2]



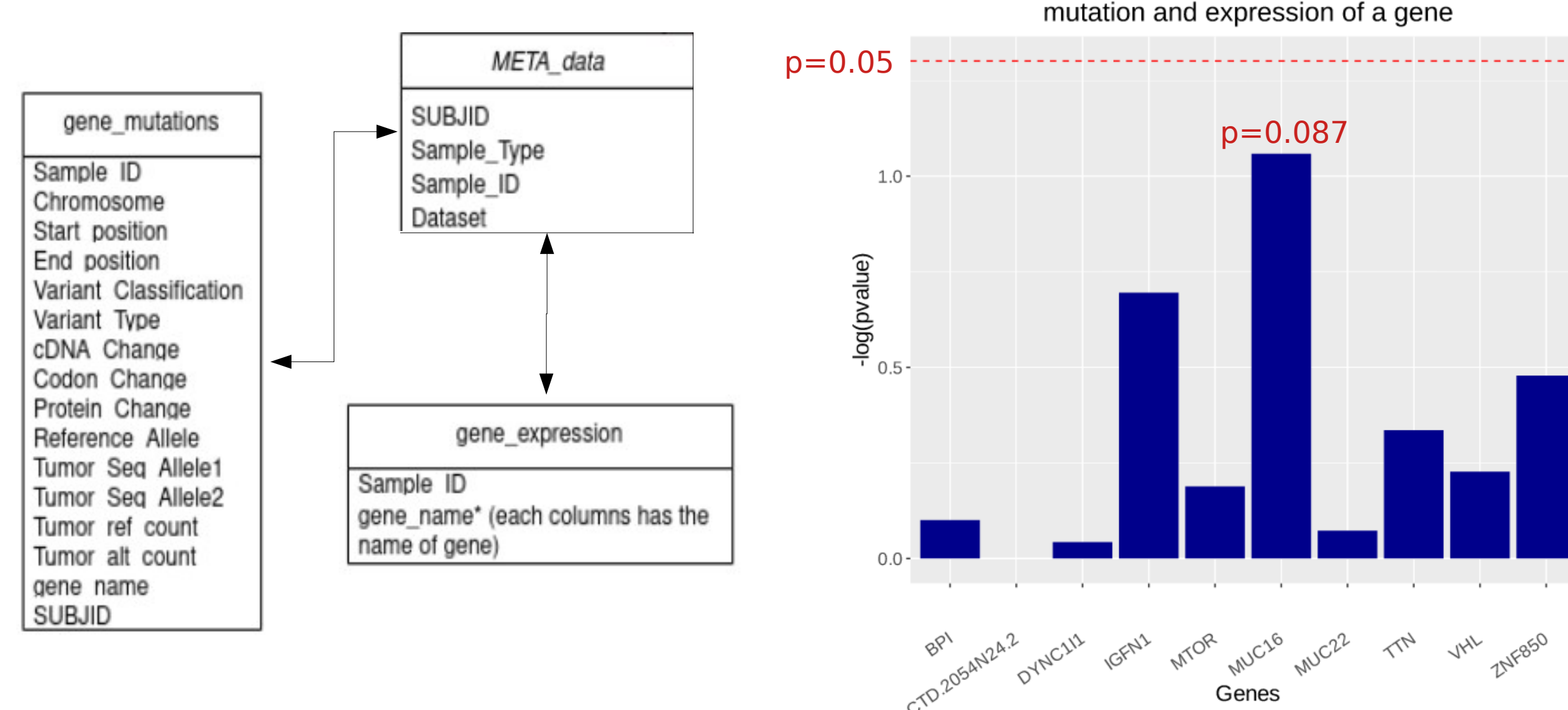
Data Queries

Use case 1: Investigate whether genes frequently mutated in kidney cancer are associated with patient response to treatments

- 1) **SELECT** gene_name, count(gene_name) **FROM** DELTA.`gene_mutations` **WHERE** (Variant_Type == 'SNP' **AND** Variant_Classification == 'Missense_Mutation') **GROUP BY** gene_name **ORDER BY** count(gene_name) **DESC LIMIT 10**
- 2) **SELECT** table1.SUBJID, table1.Sample_ID, table2.benefit **FROM** DELTA.`META_data` table1 **JOIN** DELTA.`joinTables/SUBJID` table2 **ON** table1.SUBJID == table2.SUBJID **WHERE** table1.Sample_Type == 'MAF_Tumor_ID' **AND** table2.Arm == 'EVEROLIMUS' **AND** table1.Dataset == 'BRAUN_2020'
- 3) For each mutation, count samples according to mutation status.



Use case 2: Investigate whether, for frequently mutated genes in kidney cancer, the mutation status is associated with change in gene expression



Conclusions

We prototyped of a **Data Lakehouse**, integrating patient molecular and clinical data, to support the development of AI models for predicting patient response to targeted and immuno-therapies.

The queries we implemented as part of the first two use cases allowed us to **validate our data structure**.

In perspective, we plan to add a layer of **security and data governance** and to **integrate additional data collections**.

References

- [1] Databricks <https://www.databricks.com/glossary/data-lakehouse>
- [2] Che, H., & Duan, Y. (2020). On the Logical Design of a Prototypical Data Lake System for Biological Resources. *Frontiers in bioengineering and biotechnology*, 8, 553904.
- [3] Braun, D. A., et al. (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nature medicine*, 26(6), 909-918.
- [4] Motzer, R. J., et al. (2020). Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. *Nature medicine*, 26(11), 1733-1741.