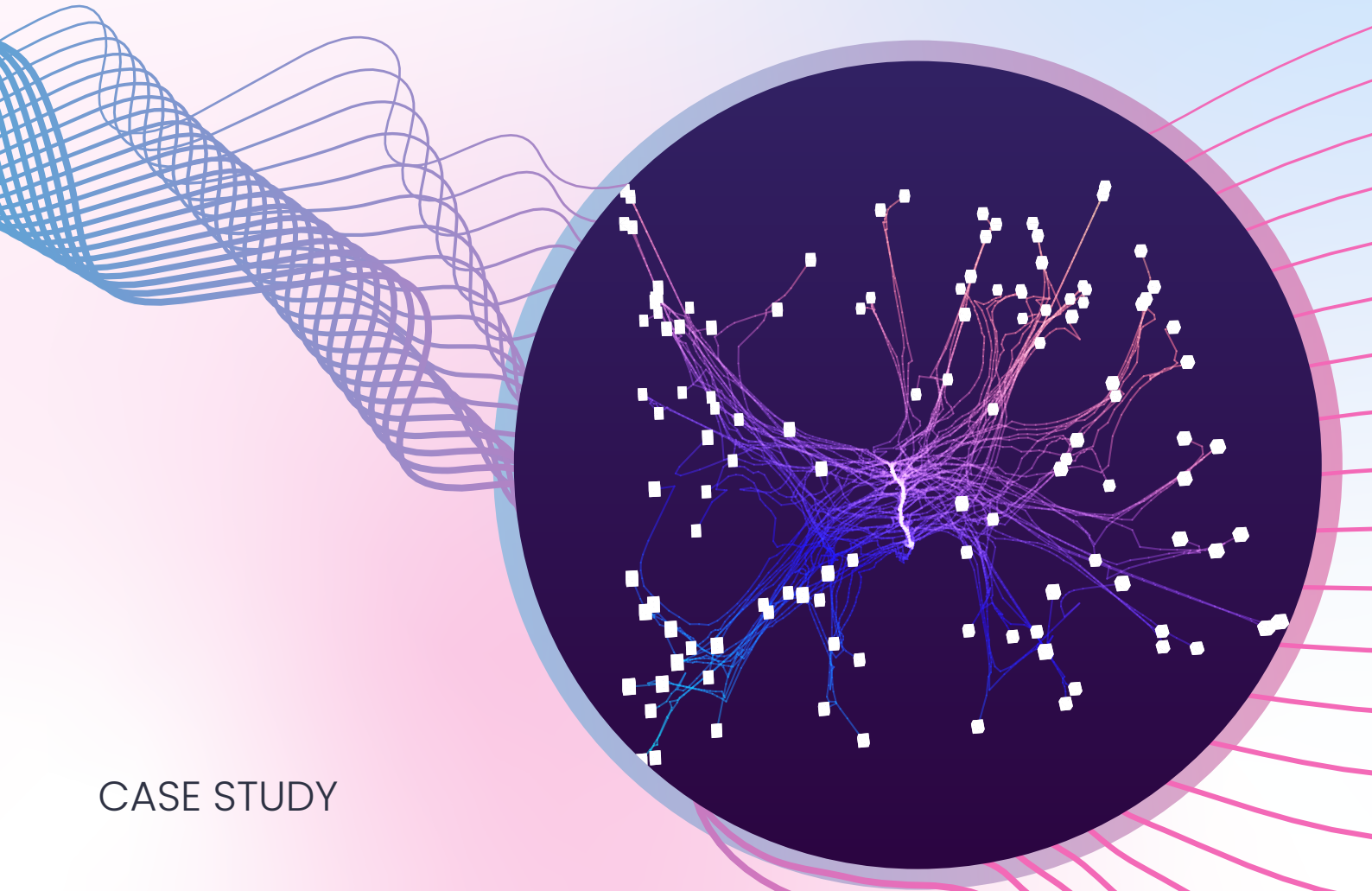
excelra

Development of a hit-calling algorithm on DEL selection data

Delivered as a containerized pipeline that accurately identifies true hits from DEL selection output data even if the raw counts are low in numbers



CASE STUDY

Client
Mid-sized pharma

Location
Europe

Services used
Data science

Results

- A fully functional, deployable, containerized pipeline
- A hit-calling algorithm that accurately finds candidate compounds or groups of compounds for resynthesis and testing
- Detailed documentation to ensure a smooth transition and optimized operation

Specification

The client is a European pharma company focused on discovering and developing small-molecule medicines with novel modes of action. A key stage of their research demands the processing of sequenced data to obtain a count matrix and quality control (QC) output, so they developed a next-generation sequencing (NGS) pipeline to process sequenced barcodes.

The challenge was due to the client's DNA-encoded library (DEL) size. Small libraries generally have higher coverage and a more straightforward enrichment method for identifying true hits. But in larger libraries like the client's, there is lower coverage (about 3-10 counts per compound), so a pragmatic statistical approach using differential gene expression was required to reliably detect true positives, and avoid the obstacles caused by the data.

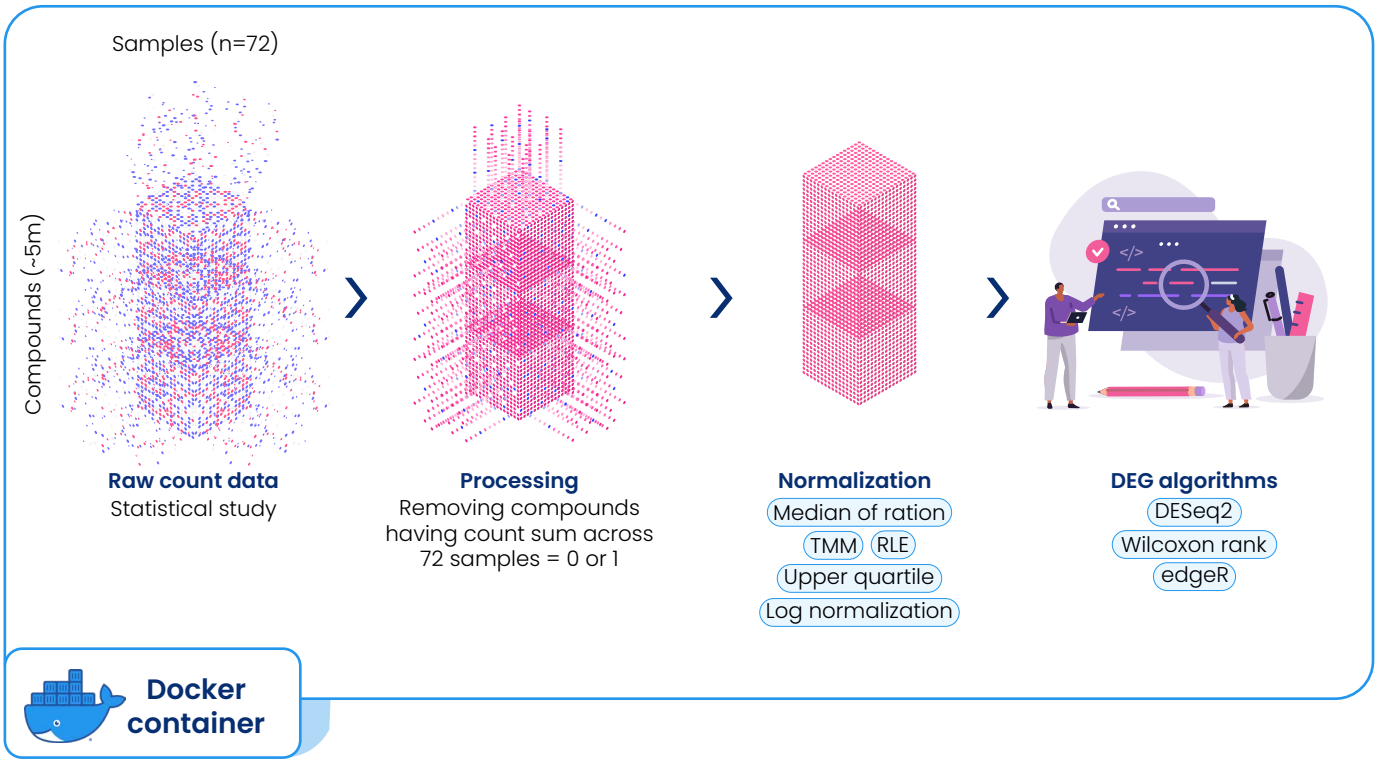
The ultimate goal was to develop a hit-calling algorithm to find candidate compounds or groups of compounds that could be resynthesized and tested. The algorithm would be given the following inputs:

- **Count matrix** (compound x sample)
- **Contrasts:** sample groups to compare
- **Sample annotation** (sample x properties)
- **Library:** compound properties (constituents, structure)

Our approach

To ensure the algorithm was optimized for the specific data involved, we started with a thorough analysis of data from 72 samples with raw counts for each of the 5 million compounds in the client's library. We also requested a contrast sheet from the client to conduct further analysis and ensure we were accurately focusing on the overall objective. Understanding that low raw count numbers increased the risk of selecting false positives or dropping false negatives, we put particular emphasis on normalizing the data set so that trends were maintained across all the generated data.

We reviewed various DEG algorithms based on the analyzed data and contrast sheet. After meticulous comparative analysis, it was clear that DESeq2 provided the most reliable and consistent results. With the algorithm selected, we built an effective and efficient automated pipeline, containerized with NextFlow. Finally, our data scientists rigorously tested the pipeline to ensure the outputs were generated in the form requested and to the standard demanded by the client before delivery.



The results

We delivered a fully functional, deployable, portable NextFlow pipeline in a containerized form that allows the client to accurately identify true hits from DEL selection output data even if the raw counts are low in numbers. The client was given detailed documentation to facilitate a smooth transition from our team to theirs, and to ensure optimized performance and minimal downtime. The client was satisfied with the pipeline's output, the quality of our work, and our ability to meet their deadline.

Conclusion

Our data scientists are supported by domain experts, so we're able to conduct all the necessary preparatory research before we start building pipelines. Understanding the complexity of the data involved and the output required helps us deliver effective and efficient pipelines optimized for our client's specific goals.

If your objectives require a close understanding of the scientific data you're working with, we're the partner you're looking for. Whatever your requirements, our data scientists and bioinformaticians can help.

Find out more about our data science services

Learn more →

Where data means more

excelra

BOSTON | UTRECHT | HYDERABAD

Connect with our experts: marketing@excelra.com

www.excelra.com