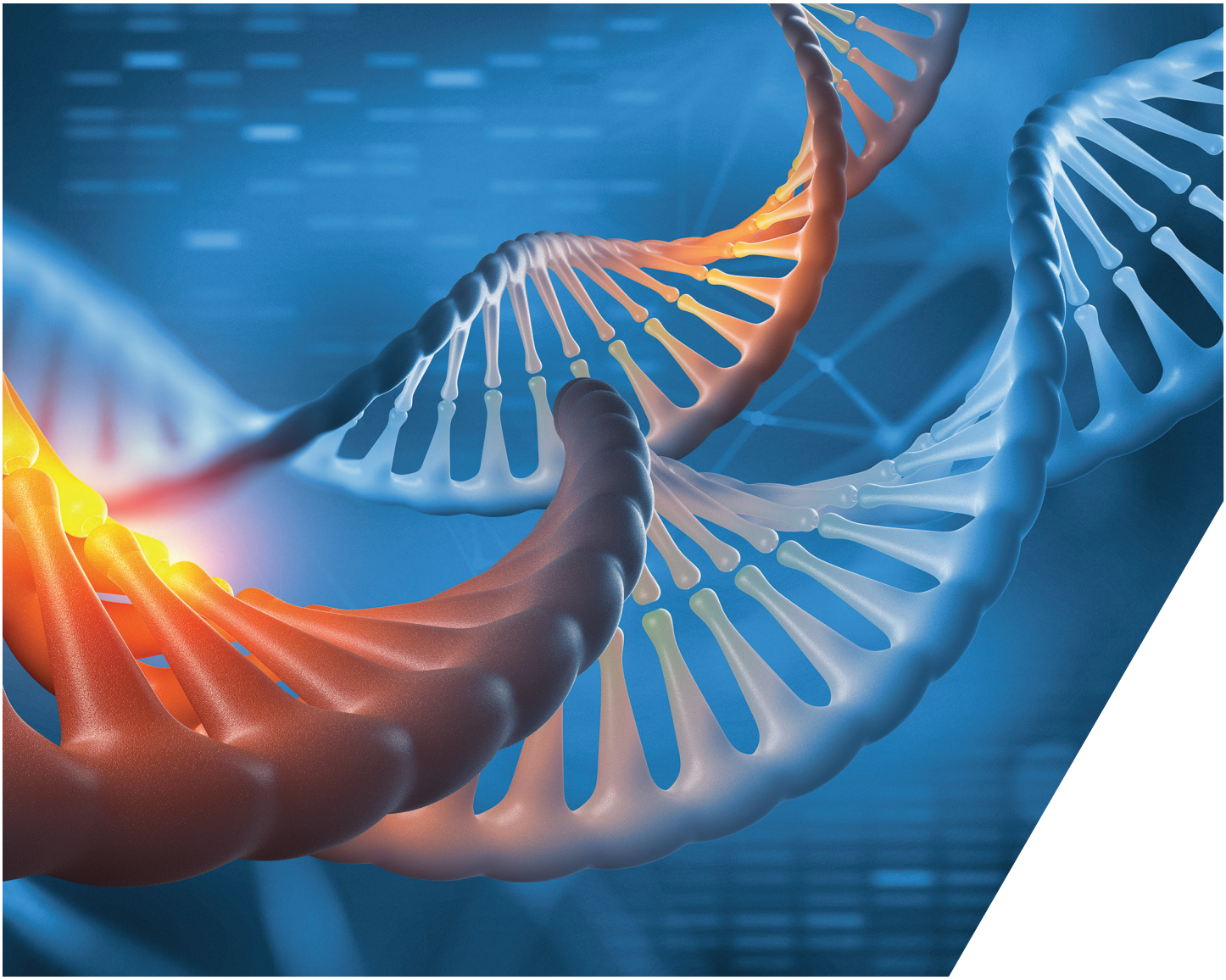
CUSTOMER STORY



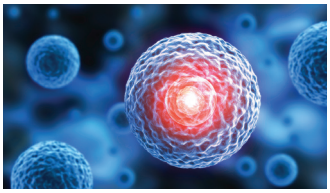
PROFILING I/O FOR GENOME PIPELINES

SANGER INSTITUTE GETS FAST, AGILE, AND CLOUD-READY WITH MISTRAL AND BREEZE

> “Improving run time often doesn’t require extensive rewrites. Knowing where to look is key.
>
> Keiran Raine, Cancer Researcher, Wellcome Sanger Institute

About the Customer

The Wellcome Sanger Institute is one of the premier centers of genomic discovery and understanding in the world. It leads ambitious collaborations across the globe to provide the foundations for further research and transformative healthcare innovations. The Institute’s Cancer Genome Project uses high-throughput genome sequencing to identify somatically acquired mutations with the aim of characterizing cancer genes, mutational processes, and patterns of clonal evolution in human tumors.



The Wellcome Sanger Institute used Mistral and Breeze to profile I/O, optimize its genome pipeline, and reduce expenses

Their Challenge

According to Cancer Research UK, 1 in 2 people born after 1960 will get cancer at some point in their lives. Carrying out genome projects to find cures is a necessity, and the Wellcome Sanger Institute is on the front lines of genomic research. The amount of data generated by such projects is immense, and each cancer sample generates around 250GB of data after initial processing. **Because so much data storage is required, optimization is crucial.**

The team at the Sanger Institute needed to make one of their cancer pipelines portable and tune it for cloud deployment. Most pipelines are written and tested on local machines and then run in parallel on compute clusters with shared storage. However, I/O behavior is very different on clusters and, unless the bioinformatician has access to comprehensive I/O profiling tools, there are likely to be inefficient I/O patterns that harm storage performance and potentially prevent others from getting their work done.

Our Solution

The Wellcome Sanger Institute used Altair Mistral™ to profile the pipeline and look for inefficient I/O patterns. The pipeline had been optimized in some areas, but Mistral revealed the need for additional improvement. It pinpointed a large number of small reads — up to a million 1-byte reads per second — which can harm computational performance and create suboptimal I/O patterns on shared storage. Optimizing small reads allows storage to run at maximum bandwidth with minimum impact on other jobs. **The team also used Altair Breeze™ to profile the containerized workload in the cloud** on Amazon Web Services (AWS). Breeze determined that the default storage option was the best value over faster, more expensive options.

Results

By using the Breeze and Mistral I/O profiling tools, the Wellcome Sanger Institute saved both time and money during a complex and high-value project. **The Institute discovered that it could save a significant 10% of project costs** by choosing a less expensive storage option without compromising performance. In addition, **run time was reduced from 32 hours to 18 hours**. The profiling work enabled the Institute's team to optimize its pipeline by employing lots of memory, profiling file I/O, and avoiding small reads and writes. When scaling up to full genomes run in parallel, speed and cost savings become increasingly important. The team put a lot of effort into tuning the pipeline and making it portable and easy to run, but their jobs were made significantly easier with easy wins that could only have been identified by measuring with the right tools.