

# **BCCM GEN-ERA**

## **BCCM collections in the genomic era**

**Contract - B2/191/P2/BCCM GEN-ERA**

### **SUMMARY**

#### **Context**

The Belgian Coordinated Collections of Microorganisms (BCCM) is a unique initiative of the Belgian Federal Government and its Science Policy Office (Belspo). BCCM is a Biological Resource Centre (BRC) that preserves and provides microbial and genetic resources to support life sciences and the biotechnology sector in the field of fundamental and applied research.

Prokaryotic and eukaryotic microorganisms represent most of the biodiversity present on Earth and can be found in every environment capable of supporting life. They play a major role in a huge range of functions, from ecosystems to humans, and contribute to countless applications. Nevertheless, the current knowledge in microbiology barely represents the tip of the iceberg and microbiological research still needs to tackle the considerable task of finding and understanding the yet undiscovered microbial capabilities. Culture collections play an essential role in this endeavour by isolating, cultivating, identifying, preserving and distributing the cultivable diversity.

Since its creation in 1983, BCCM developed and maintained international leadership among European BRCs through the implementation of a number of features including a website, an ISO 9001 certification, an online catalogue, a laboratory information management system and its recognition as international depository authority. The study, valorisation and documentation of the microbial resources also need to stay up to date with the current developments in microbiology and modern technologies to analyse microorganisms in an efficient manner. Nowadays, microbial research is greatly facilitated by new approaches in genomics including whole genome sequencing. The latter provides the entire genetic information of an organism and is increasingly requested in many disciplines. Acquiring know-how in genomics is thus crucial for BCCM in order to remain a major BRC, for future national and international collaborations and to answer upcoming research questions.

#### **Objectives**

The first objective of the BCCM GEN-ERA project was to implement expertise in genomics in the BCCM collections for which the real challenge was the handling and analysis of genomic big data. This required the installation of specific bioinformatics structures (hardware) and software for which the BCCM scientists had to be trained to ensure a long-term implementation. The focus was primarily on whole genome sequencing because the determination of the entire gene repertoire of microorganisms supports the leading expertise of the collections in the field of taxonomy and phylogeny while allowing potential functional analyses. Moreover, offering strains having their whole genome sequenced is necessary to meet the needs of BRCs users and is thus an added value for the visibility and attractiveness of BCCM.

The BCCM GEN-ERA project also aimed at answering specific research questions covering the microbial diversity of the BCCM collections (i.e., bacteria, mycobacteria, cyanobacteria, yeasts and moulds) and more specifically on selected microorganisms having a societal impact (i.e., associated with pollinating insects, causing pathologies or producing bioactive compounds). The project implicated five out of the seven BCCM collections and was a collaboration with the laboratory of Eukaryotic Phylogenomics of the University of Liège that brought the necessary expertise in bioinformatics and (phylo)genomics.

## Conclusions

Two different bioinformatics infrastructures were envisaged for the handling and analysis of the sequencing data, namely Galaxy and Nextflow. Both were tested and compared for their respective performance, appropriateness, user-friendliness and compliance with the FAIR principles (i.e., Findable, Accessible, Interoperable, Reusable).

Galaxy is a bioinformatics web platform whose objective is notably to make genomic analyses available to all researchers, even with little computing skills, thanks to a user-friendly graphical user interface. The installation of a private “BCCM” Galaxy was tested but experienced several security gaps that complicated the deployment. Therefore, it appeared that a system administrator would be needed to maintain the infrastructure. These difficulties also limited the interoperability and reuse of the bioinformatics workflows. Moreover, some mandatory programs required for modern bioinformatics practices were not available. For these reasons, Nextflow, designed to perform bioinformatics using command lines instead of a graphical user interface, was preferred. In total, 14 Nextflow workflows, sustained by 11 Singularity containers were implemented. They cover the common genomics needs for microbial taxonomy and metabolic modelling of microbial collections like BCCM. They can be used on prokaryotes and small eukaryotes in a completely reproducible manner. The workflows are provided to the users as a program that can be run with a single command line, increasing the reproducibility of the analyses. This “GEN-ERA toolbox” was made freely available from the GitHub repository <https://github.com/Lcornet/GENERA> which also includes a large documentation for the users. Nextflow thus fulfilled most of the criteria for a long-term and FAIR utilization of the bioinformatics infrastructure at BCCM. The only disadvantage, as compared to Galaxy, is the user-friendliness. Working with command lines is indeed less intuitive and necessitated dedicated trainings, but could be reduced to a minimum thanks to the Singularity containers.

The Nextflow infrastructure implemented at BCCM in collaboration with the University of Liège was used to investigate case studies for which pending research questions could be addressed by genomic analyses. In particular, we investigated fungal pathogens causing skin infections, non-*tuberculosis* mycobacteria, bacteria and yeasts from the gut of pollinating insects (bees and bumblebees) as well as cyanobacteria displaying biological activities. These analyses provided breakthroughs in their respective fields and opened new perspectives for future researches.

The BCCM GEN-ERA project established expertise in genomics at BCCM by setting a bioinformatics framework, by bringing genomic tools and by developing genomics skills of the BCCM scientists. This investment was thus performed in view of a long-term implementation allowing genomics to become a continuous activity within the BCCM consortium. In this respect, an essential milestone was achieved with the BCCM GEN-ERA GitHub repository which can be regarded as an open gateway to use, reuse and learn to use bioinformatic tools for genomic analyses. It was designed to offer online and free access to analysis programs for the handling of genomic data from prokaryotic and eukaryotic microorganisms. It is a web portal that centralizes software for notably genome assembly, genome annotation, phylogenomics or metabolic modelling of genomes. It works as a common platform that can be used by all BCCM collections and scientists but also by other facilities (e.g. BRCs, microbiological laboratories) that are interested in the same topics.

## Keywords

Genomics; BCCM; microorganisms; genomes; culture collection; bioinformatics; taxonomy; phylogeny; molecular evolution; biodiversity.