Integrating Nextflow and AWS for Large-Scale Genomic Analysis: A Hypothetical Case Study

Ángel Canal-Alonso1, Pedro Jiménez1 and Noelia Egido1, Javier Prieto, Juan Manuel Corchado1

[1]Department of Bioinformatics and Computational Biology, AIR Institute, Carbajosa de la Sagrada, Spain

Email: acanal@air-institute.com

Summary

This article explores the innovative combination of Nextflow and Amazon Web Services (AWS) to address the challenges inherent in large-scale genomic analysis. Focusing on a hypothetical case called "The Pacific Genome Atlas", it illustrates how a research organization could approach the sequencing and analysis of 10,000 genomes. Although the "Pacific Genome Atlas" is a fictional example used for illustrative purposes only, it highlights the real challenges associated with large genomic projects, such as handling huge volumes of data and the need for intensive computational analysis. Through the integration of Nextflow, a workflow management tool, with the AWS cloud infrastructure, we demonstrate how these challenges can be overcome, offering scalable, flexible and cost-effective solutions for genomic research. The adoption of modern technologies, such as those described in this article, is essential to advance the field of genomics and accelerate scientific discoveries.

Keywords: Next-Generation sequencing, Cloud computing, Distributed Computing

Introduction

DNA sequencing

DNA sequencing is a process that determines the exact order of nucleotides (adenine, thymine, guanine and cytosine) in a DNA molecule. Since its invention in the 1970s, DNA sequencing has revolutionized the field of genomics and provided invaluable information about the genetics of numerous organisms, including humans.

History and Development

The initial development of DNA sequencing relied on manual techniques that were laborious and time-consuming. The most famous method of this era was the chain termination method developed by Frederick Sanger in 1975, which earned him a Nobel Prize. This method, known as Sanger sequencing, was the standard technique for many years and was used, for example, in the Human Genome Project, which aimed to sequence the entire human genome.

Over time, DNA sequencing techniques have evolved, giving way to what is known as Next Generation Sequencing (NGS). NGS is capable of sequencing millions of DNA fragments simultaneously, making it much faster and more efficient than previous techniques.

Importance of DNA Sequencing

DNA sequencing has had a profound impact on biology and medicine. It has allowed scientists to discover thousands of human genes, identify genetic variants associated with diseases and disorders, and better understand the evolution and diversity of life on Earth.

In medicine, DNA sequencing has led to the development of genetic tests for inherited diseases, the identification of new therapeutic targets, and the personalization of treatments based on an individual's genetic profile.

DNA sequencing also has applications in ecology, forensics, agriculture, and many other fields, demonstrating its versatility and relevance in contemporary science.

Sequencing Challenges

Despite advances, DNA sequencing is not without challenges. The massive volume of data generated by techniques such as NGS requires sophisticated data storage and analysis solutions. This is where distributed computing techniques come into play, providing tools and methodologies to efficiently handle, process and analyze large sequencing data sets.

Next generation sequencing

Next Generation Sequencing (NGS), also known as massively parallel sequencing, has revolutionized the field of genomics by allowing the simultaneous sequencing of millions of DNA fragments. Unlike Sanger sequencing, which sequences one piece of DNA at a time, NGS can analyze an entire genome in a single experiment. Here's how these techniques work:

Before sequencing, the DNA of interest (which can be the entire genome or specific regions) is fragmented into smaller pieces. These fragments are ligated with specific adapters that allow amplification and subsequent sequencing.

A common feature in NGS techniques is the amplification of DNA fragments. One of the most widely used techniques is bridge amplification, where DNA fragments are attached to a solid surface and copied in multiple iterations, generating a "cluster" of identical fragments.

Here are some of the most popular sequencing techniques:

- Sequencing by Synthesis (SBS): This technique, used on Illumina platforms, involves the sequential incorporation of fluorescently labeled nucleotides. Each time a nucleotide is incorporated, a fluorescent signal is emitted that is detected and recorded. The intensity and color of the fluorescence determine the specific nucleotide that has been incorporated.

- Ion Sequencing: Used in Ion Torrent platforms, this technique detects the ions released during the incorporation of a nucleotide into the growing DNA chain. No fluorescent markers required; Instead, a sensor detects the change in pH caused by the release of a hydrogen ion each time a nucleotide is added.

- Nanopores: In this technique, used in the Oxford Nanopore platforms, DNA molecules pass through nanopores in a membrane. As each nucleotide passes through the nanopore, it causes changes in electrical current that can be detected and translated into a sequence.

Given the large amount of data generated in a single NGS experiment, data analysis is a critical step. The process generally includes:

- Alignment/mapping: Sequence fragments, known as "reads", are aligned or mapped to a reference sequence, such as the human reference genome.
- Identification of variants: Once the reads are mapped, genetic variants can be identified, such as single nucleotide polymorphisms (SNPs) or insertions and deletions (indels).
- Functional analysis: Databases and bioinformatics tools can be used to interpret the potential impact of the identified variants.

NGS techniques have opened new possibilities in genomic research, allowing everything from the study of genetic diversity to the identification of mutations associated with diseases. However, the massive volume of data generated requires advanced computing solutions, and this is where distributed computing plays a crucial role.

Distributed computing

Distributed computing is a field of computer science that studies distributed systems and how they can be designed and coordinated to achieve a common goal. A distributed system consists of multiple autonomous computers connected through a network, where each computer has its own local resources, such as CPU, memory, and storage. Unlike a centralized system where all operations and resources are concentrated in a single machine or location, in a distributed system, operations are distributed among different machines, and they work together as a single cohesive system.

One of the main benefits of distributed computing is its ability to handle large amounts of data and perform complex calculations efficiently. By splitting a task into smaller threads and distributing them across multiple machines, it is possible to process large volumes of data in parallel, significantly speeding up processing time. This feature is especially valuable in fields such as genomics, where the data sets are huge and the analysis can be computationally intensive.

Another advantage is scalability. As compute or storage demands increase, more machines can be added to the distributed system, allowing the system to grow and adapt to changing needs. This contrasts with centralized systems, where scalability may be limited by the capacity of the central machine.

Fault tolerance is another key benefit. In a distributed system, if one machine fails, the others can continue working. Data and tasks can be replicated across multiple machines, so if one machine goes offline or fails, another machine can take over its work, ensuring system continuity and availability.

However, distributed computing also presents challenges. Coordination between machines, ensuring data consistency and fault management are complex aspects that must be carefully managed. Additionally, communication between machines on a network can introduce latency, which can impact system performance.

In the context of next-generation sequencing (NGS), distributed computing has proven to be a valuable tool for managing and analyzing large genomic data sets. By distributing the analysis across multiple machines, it is possible to significantly speed up processing time, allowing researchers to obtain results more quickly and perform more complex analyzes that were not possible before.

In the world of distributed computing, there are various architectures and models that have been developed to address different types of problems and needs. Below are some of the most common architectures:

1. Client-Server Systems:
- In this architecture, there is a clear distinction between service providers (servers) and service requesters (clients).
- Servers maintain and offer services and resources to clients, who in turn request and consume these services.
- Classic example: web servers and browsers.

2. Peer-to-Peer (P2P) Systems:
- In a P2P system, all nodes have similar capabilities and can act as both clients and servers.
- It is designed to be decentralized, eliminating the need for a central server.
- They are widely used for file sharing, as in the case of BitTorrent.

3. Architectures based on Grids (Grid Computing):
- Designed to solve problems that require large amounts of computational resources, combining resources from multiple locations and organizations.
- Nodes in a grid do not necessarily have a constant or long-term relationship, and are often added and removed dynamically as needed.
- Example: SETI@home project for the search for extraterrestrial intelligence.

4. Computer Clusters:
- A cluster is a group of computers connected to each other, working together to offer high availability, high computing capacity, or both.
- Often, machines in a cluster are similar and located physically close to each other, connected by a high-speed local network.
- Example: Hadoop clusters for big data processing.

5. Cloud Computing:
- Provides computing resources (CPU, memory, storage) as a service over the Internet.
- Resources can be dynamically scaled according to needs.
- Examples: Services such as Amazon Web Services (AWS), Google Cloud Platform and Microsoft Azure.

6. Systems based on Data Volumes (Data-intensive Computing):
- Designed specifically for problems that require processing large amounts of data.
- Often uses distributed storage and processing systems, such as Hadoop and its distributed file system (HDFS).

7. Architectures based on Microservices:
- Splits an application into small independent services that run as separate processes and communicate with each other via lightweight mechanisms, usually HTTP.
- Each microservice is responsible for a specific function and can be developed, deployed and scaled independently.

These architectures can be applied in different contexts according to the specific needs of the problem to be solved. For next-generation sequencing and genomic analysis, the ability to process and store large amounts of data efficiently is essential, making certain architectures, such as cloud computing and volume-based systems, data, are particularly relevant.

The confluence between NGS and Distributed Computing

Incorporating distributed computing into Next Generation Sequencing (NGS) is essential to address the computational and storage challenges presented by this technology. The data generated by NGS platforms is massive, and the analysis of this data is computationally intensive. Distributed computing offers solutions to overcome these challenges. The following describes the ways in which distributed computing can be applied in NGS:

1. Distributed Storage:

- NGS platforms generate terabytes of data in a single experiment. Storing this data on a single machine or server is neither practical nor efficient.
- Distributed file systems, such as the Hadoop Distributed File System (HDFS), allow large data sets to be stored by distributing the data across multiple machines, providing redundancy and high availability.

2. Parallel Processing:

- NGS analysis algorithms, such as read alignment and variant identification, are computationally intensive.
- Distributed processing frameworks, such as Hadoop and Spark, allow these tasks to be divided into smaller subprocesses that run in parallel on multiple nodes in a cluster, significantly speeding up the process.

3. Scalability:

- As data volume or computing demands increase, more nodes can be added to the distributed system.
- This scalability is essential to keep up with the exponential growth of genomic data.

4. Fault Tolerance:

- In a distributed system, if one node fails, work can be redistributed to other nodes, ensuring that progress is not lost and that analysis can continue without interruption.

5. Cloud Analysis:

- Cloud computing solutions, such as AWS, Google Cloud, and Microsoft Azure, offer "on-demand" distributed computing capabilities.
- Researchers can rent resources as needed, avoiding investment in expensive infrastructure and allowing access to cutting-edge tools and algorithms.

6. Workflow Optimization:

- Workflow management platforms, such as Apache Airflow or Nextflow, allow you to design, coordinate and optimize complex workflows in distributed environments, ensuring that tasks are executed in the correct order and making the most of available resources.

7. Collaboration and Data Sharing:

- Distributed systems facilitate data sharing and access between researchers and organizations, promoting collaboration and enabling joint analysis of large data sets.

In summary, distributed computing is essential to take full advantage of the capabilities of Next Generation Sequencing. By providing solutions for the storage and processing of large data sets, as well as for optimization and collaboration, distributed computing has expanded the possibilities of genomics research and accelerated discoveries in this field.

Use Case: Implementation of Genomic Analysis with Nextflow and AWS

Genomics has undergone an unprecedented revolution in terms of data generation and analysis complexity. Researchers and data scientists are challenged to rapidly process and analyze terabytes of sequencing data, and this requires robust, flexible and scalable computing infrastructures. In this context, cloud-based tools and services, such as AWS, and workflow management systems, such as Nextflow, have emerged as leading solutions to address these challenges.

In this use case, we will explore how a hypothetical genomics research organization implements an NGS analysis workflow using Nextflow to coordinate and optimize the process, and AWS to provide the necessary computational and storage infrastructure. The combination of these two technologies allows the organization to dynamically scale its resources, optimize costs and accelerate time to discovery.

This scenario illustrates a practical and modern approach to genomics research, demonstrating how distributed and cloud computing solutions can be effectively implemented to overcome the computational challenges inherent to the field of genomics.

Context and Need

The Pacific Genomics Organization (PGO) is a prominent research center located on the west coast of the United States. For more than a decade, OGP has led research in the areas of human genomics, microbiome, and evolutionary genomics. With a team of more than 200 scientists, the organization has produced some of the most cited studies in the field of genomics.

In recent years, the OGP has initiated an ambitious project: "The Genomic Atlas of the Pacific". This project seeks to sequence and analyze the genomes of 10,000 individuals from the Pacific region to study genetic diversity, identify rare variants, and better understand the evolutionary adaptations of populations in this unique geographic area.

Given the magnitude of the "Pacific Genomic Atlas" project, the OGP faced several challenges:

1. Data Volume: Each complete genome sequenced generates around 100-150 GB of data in raw format. With 10,000 genomes, the generation of more than a petabyte of data is anticipated.

2. Intensive Computational Analysis: Analysis of sequencing data is not a trivial task. It includes steps such as alignment, variant identification, and annotation, each of which requires significant computational resources.

3. Time: With traditional infrastructures, processing a single genome can take days. With 10,000 genomes, this translates into years of processing time, which is unacceptable for the organization.

4. Costs: Storing and analyzing large volumes of data can result in prohibitive costs if not properly managed and optimized.

Therefore, OGP needed a solution that could not only handle the massive volume of data and computational analysis, but was also cost-effective and significantly reduced overall processing time. Adoption of distributed computing techniques and cloud solutions emerged as the most viable solution to address these challenges.

Tool Selection

Given the need of the Pacific Genomics Organization (PGO) to efficiently process and analyze large volumes of genomic data, a process of evaluation and selection of suitable tools was initiated. Several solutions and platforms were considered, but the selected tools had to meet specific criteria, such as scalability, flexibility, ease of use, integration capabilities, and cost-effectiveness.

1.Nextflow:

Reasons for choice:

- Flexibility: Nextflow allows you to write complex workflows using a simple, easy-to-read script language. This makes it easy to adapt and modify workflows according to the needs of the project.
- Portability: Nextflow is infrastructure agnostic, meaning workflows can run on different platforms without modification, from a local laptop to high-performance clusters or the cloud.
- Parallelization and Optimization: Nextflow automatically handles the parallelization of tasks and optimizes resource utilization.
- Reproducibility: Allows consistent and reproducible execution of analysis, ensuring that the results are reliable and repeatable in different environments.

2. Amazon Web Services (AWS):

Reasons for choice:

- Scalability: AWS offers a wide range of services that can be dynamically scaled as needed. This is crucial to handle demand spikes during intensive processing phases.
- Variety of Services: From EC2 for computing, S3 for data storage, to more specialized services such as AWS Batch for the execution of batch jobs, AWS covers all the needs of the OGP.
- Cost-Effectiveness: With its pay-per-use model, AWS allows the OGP to optimize costs, since it only pays for the resources it uses. Additionally, AWS offers spot and reserved instance options that can further reduce costs.
- Integration with Nextflow: AWS integrates seamlessly with Nextflow, allowing workflows to be deployed and executed directly in the cloud without additional effort.

After careful consideration and preliminary testing, the combination of Nextflow and AWS emerged as the most suitable solution for the needs of the Pacific Genome Atlas project. This combination not only addresses scalability and performance challenges, but also offers a cost-effective solution for large-scale genomic analysis.

Workflow Design with Nextflow

The design of the workflow in Nextflow for the OGP "Pacific Genomic Atlas" project focused on addressing the main steps of NGS sequencing analysis, from receiving data in raw format to the identification and annotation of genetic variants. Below is a detailed description of the designed workflow:

1. Data Entry:

- Raw Reads: Raw sequencing reads, generally in FASTQ format, are the main input. These readings can be stored locally or in an S3 bucket on AWS.

2. Quality Control (QC):

- Tool: FastQC.
- A quality assessment of the raw reads is performed to identify potential problems such as DNA contamination or degradation.
- Reports generated by FastQC are consolidated for detailed review.

3. Preprocessing of Reads:

- Tools: Trimmomatic and Cutadapt.

- Elimination of low quality adapters and bases from the readings.
- The resulting clean readings are used for the next steps of the analysis.

4. Reading Alignment:
- Tool: BWA-MEM.
- Reads are aligned against a reference genome (e.g. human genome GRCh38) to obtain an aligned BAM file.
- Alignment allows readings to be mapped to their corresponding genomic location.

5. Post-Alignment Processing:
- Tools: SAMtools and GATK.
- Various operations such as deduplication, base quality recalibration and local realignment are performed to optimize the quality of the BAM file.

6. Identification of Variants:
- Tool: GATK HaplotypeCaller.
- Genetic variants, such as SNPs and indels, are identified, generating a VCF file.
- This step is crucial for the main objective of the project: to discover genetic variants in the Pacific population.

7. Annotation of Variants:
- Tool: ANNOVAR or SnpEff.
- Identified variants are annotated to provide information on their potential impact, genomic location, amino acid changes (if applicable) and other relevant characteristics.

8. Output and Report:
- Detailed reports and analysis summaries are generated.
- Processed data and results can be stored back to S3 or another suitable storage system for further analysis or sharing.

The workflow designed with Nextflow is structured into individual processes, where each process represents a specific step of the analysis (e.g., quality control, alignment, variant identification). Nextflow automatically handles task parallelization, dependency management, and error monitoring, ensuring a robust and efficient workflow. Additionally, thanks to its integration with AWS, computational resources are dynamically scaled according to the needs of each step of the analysis.

Infrastructure on AWS

For the "Pacific Genome Atlas" project carried out by the Pacific Genomics Organization (PGP), the infrastructure on Amazon Web Services (AWS) was designed to maximize efficiency, scalability and cost-effectiveness. The infrastructure selected and configured in AWS is detailed below:

1. Amazon S3 (Simple Storage Service):
- Use: Data storage.
- Raw sequencing reads in FASTQ format, as well as all intermediate and final data, are stored in S3 buckets.
- S3 offers durability, high availability and scalability for large data sets.

2. Amazon EC2 (Elastic Compute Cloud):
- Use: Processing and analysis.
- EC2 instances provide the computing power needed to run the Nextflow workflow.
- Different types and sizes of instances are used according to the specific needs of each step of the analysis. For example, instances with high memory can be used for variant identification and instances with high CPU performance can be used for alignment.

3. AWS Batch:
- Use: Management and execution of jobs.
- AWS Batch manages the execution of jobs in containers, automatically scaling the number of EC2 instances as needed.
- It is integrated with Nextflow, allowing workflow jobs to run efficiently in the cloud.

4. Amazon RDS (Relational Database Service):
- Use: Storage of metadata and analysis results.
- RDS provides relational databases that are used to store metadata about samples as well as summarized analysis results.

5. Amazon EFS (Elastic File System):
- Use: Shared storage.
- EFS provides a highly available, scalable file system that can be mounted on multiple EC2 instances. It is useful for data that must be accessible from multiple instances simultaneously.

6. AWS Lambda and Step Functions:
- Use: Automation and coordination.
- AWS Lambda allows code execution in response to specific events, such as the completion of a job or the arrival of new data in S3.
- Step Functions are used to coordinate microservices and automate workflows.

7. Amazon CloudWatch:
- Use: Monitoring and alerts.
- CloudWatch monitors the performance of AWS resources, providing real-time metrics, logs, and alerts. This is

essential to ensure the system is operating optimally and to quickly detect and address any issues.

8. AWS Identity and Access Management (IAM):
- Use: Security and access control.
- IAM is used to define permissions and policies, ensuring that only authorized users and services can access AWS resources.

The combination of these services on AWS provides a robust and scalable infrastructure for the project. Thanks to the flexibility and variety of services available on AWS, OGP was able to design a solution that perfectly fits its needs, ensuring efficiency in the processing and analysis of large-scale genomic data.

Nextflow integration with AWS

Nextflow's integration with AWS enables genomic workflows to run efficiently on cloud infrastructure, taking advantage of the scalability, flexibility, and computing power of AWS. Below is how this integration was carried out:

1. Credential Configuration:
- To allow Nextflow to interact with AWS services, you need to configure AWS credentials. This is done using AWS Identity and Access Management (IAM) roles and policies to ensure secure access.
- Credentials are stored securely and provided to Nextflow, either through environment variables or configuration files.

2. Running in AWS Batch:
- Nextflow has a specific runner for AWS Batch, making it easy to run jobs on AWS infrastructure.
- Using the AWS Batch runner, Nextflow automatically submits jobs for execution on AWS, handles task parallelization, and manages dependencies between tasks.

3. Storage in S3:
- Input, intermediate, and output data are stored in Amazon S3 buckets.
- Nextflow can directly read data from S3 and also write results to S3, eliminating the need for manual data transfers.

4. EFS for Shared Storage:
- For certain workflow steps that require access to data from multiple instances, Amazon EFS is used.
- Nextflow is configured to mount EFS volumes on EC2 instances, allowing shared access to data.

5. Monitoring with CloudWatch:
- Nextflow is integrated with Amazon CloudWatch to monitor the performance and status of jobs in real time.
- Metrics and logs generated by Nextflow are sent to CloudWatch, allowing alerts and monitoring in case of failures or bottlenecks.

6. Automation with Lambda:
- AWS Lambda functions are used to automate certain post-processing aspects, such as notification of the completion of a workflow or cleaning up temporary resources.
- Nextflow can trigger these Lambda functions upon completion of certain stages of the analysis.

7. Cost Optimization:
- Thanks to the integration with AWS, EC2 spot instances can be used to run jobs, which can significantly reduce costs.
- Nextflow is configured to request and use these spot instances when they become available.

Nextflow's integration with AWS offers a powerful and cohesive solution for running large-scale genomic analyzes in the cloud. By combining Nextflow's workflow management capabilities with AWS infrastructure, OGP was able to create a system that maximizes efficiency, reduces analysis time, and provides excellent value for money.

Conclusions

The "Pacific Genomic Atlas" project, carried out by the Pacific Genomics Organization (PGP), represents one of the most ambitious initiatives in the field of genomics in the Pacific region. The magnitude of the data generated and the complexity of the analysis required infrastructure and tools that were scalable, efficient and cost-effective. The solution found in the combination of Nextflow and Amazon Web Services (AWS) proved to be the right choice to face these challenges.

The main conclusions of the project are:

1. Scalability and Performance: The integration of Nextflow with AWS allowed OGP to process large volumes of genomic data in a significantly reduced time compared to traditional infrastructures. AWS's ability to dynamically scale resources ensured that analysis was performed optimally, regardless of data volume.

2. Flexibility and Adaptability: The modular and adaptable nature of the workflows in Nextflow, combined with the wide range of services offered by AWS, allowed OGP to adapt and modify the analysis according to the emerging needs of the project.

3. Cost-Effectiveness: By leveraging specific AWS services, such as EC2 spot instances and optimized S3 storage, OGP was able to keep costs under control, achieving an excellent return on investment.

4. Reproducibility and Consistency: One of the main advantages of using Nextflow is the guarantee of reproducibility. Each analysis performed can be replicated precisely, ensuring the validity and reliability of the results obtained.

5. Collaboration and Sharing: The cloud-based infrastructure facilitated the sharing of data and results with collaborators and stakeholders, promoting greater collaboration and transparency in the project.

6. Future Vision: With the infrastructure and workflows already established in AWS and Nextflow, OGP is well positioned for future projects and expansions. The adaptability of the solution ensures that the organization will be prepared to face future challenges in the field of genomics.

In summary, implementing genomic analysis using Nextflow and AWS has proven to be a powerful and effective combination. This use case serves as an excellent example of how modern technologies can be used to advance genomic research, providing innovative solutions to traditional challenges in the field.

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6

Acknowledgments

The present study has been funded by the AIR Genomics project (file number CCTT3/20/SA/0003) through the 2020 call for R&D Projects Oriented towards Excellence and Competitive Improvement of CCTT by the Institute of Business Competitiveness of Castilla y León and FEDER funds .