Evaluation of points of improvement in NGS data analysis

Ángel Canal-Alonso1, Noelia Egido1, Pedro Jiménez1, Javier Prieto[1], Juan Manuel Corchado[1]

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

Email: acanal@air-institute.com

Abstract

Sample text inserted for illustration. Replace with abstract text.Your abstract should give readers a brief summary of your article. It should concisely describe the contents of your article, and include key terms. It should be informative, accessible and not only indicate the general scope of the article but also state the main results obtained and conclusions drawn. The abstract should be complete in itself; it should not contain undefined abbreviations and no table numbers, figure numbers, references or equations should be referred to. It should be suitable for direct inclusion in abstracting services and should not normally be more than 300 words.

Keywords: term, term, term

1. Introduction

DNA sequencing is a fundamental technique in molecular biology that allows the exact sequence of nucleotides in a DNA sample to be read. Over the past decades, DNA sequencing has seen significant advances, evolving from manual and laborious techniques to modern high-throughput techniques. Despite these advances, interpretation and analysis of sequencing data continue to present challenges.

Artificial Intelligence (AI), and in particular machine learning, has emerged as an essential tool to address these challenges. The application of AI in the sequencing pipeline refers to the use of algorithms and models to automate, optimize and improve the precision of the sequencing process and its subsequent analysis.

The Sanger sequencing method, introduced in the 1970s, was one of the first to be widely used. Although effective, this method is slow and is not suitable for sequencing large amounts of DNA, such as entire genomes. With the arrival of next generation sequencing (NGS) in the 21st century, greater speed and efficiency in obtaining genomic data has been achieved. However, the exponential increase in the amount of data produced has created a bottleneck in its analysis and interpretation.

This is where AI comes into play. With its ability to process and analyze large sets of data quickly, AI can identify patterns, make predictions, and provide insights that would be difficult or impossible to obtain through conventional methods.

The application of AI in sequencing pipelines can be categorized into several stages:

- Raw Data Processing: DNA sequencing produces enormous amounts of raw data that require cleaning, alignment and assembly. Machine learning algorithms can optimize this process, eliminating errors, identifying regions of interest, and assembling sequences more efficiently.
- Identification of Variants: Once the DNA is sequenced and assembled, the next step is to identify the genetic variants. AI can help distinguish between true variants and sequencing errors, as well as predict the potential impact of variants on gene function.
- Functional Analysis and Interpretation: With the sequenced data in hand, it is essential to understand

the biological meaning behind it. AI can group genes by function, predict protein interactions, and associate genetic variants with specific diseases or traits.

Integrating AI into the sequencing process not only speeds up analysis, but also improves the accuracy and scope of the insights obtained.

As we progress through this analysis, we will further explore specific areas of improvement and how AI can contribute to the continued evolution of DNA sequencing.

Although sequencing technologies have advanced rapidly, significant challenges still exist:

- Genome Complexity: Many regions of the genome are difficult to sequence due to their repetitiveness or structure. These regions can hide crucial information about diseases and other genetic characteristics.
- Sequencing Errors: Despite improvements, errors in the sequencing process still persist. These may be due to the quality of the DNA, sample preparation, or the technology used.
- Interpretation of Variants of Uncertain Significance: Not all identified variants have known clinical significance, complicating the interpretation and clinical application of the data.

AI can contribute to these challenges with different approaches:

- Improved Sequencing Accuracy: AI algorithms, such as neural networks, can be trained to identify and correct sequencing errors, improving the accuracy and reliability of the results.
- Handling Complex Regions: AI techniques can be especially useful for analyzing complex genomic regions, identifying patterns in the data that may be too subtle to detect using conventional methods.
- Automated Interpretation: AI can help automatically interpret detected variants, associating them with disease databases and scientific literature, thus providing biological and clinical context.
- Predictive Analytics: With the use of AI, it is possible to perform predictive analytics to anticipate how certain variants may influence the development of specific diseases or traits in the future.

The incorporation of Artificial Intelligence into DNA sequencing pipelines represents a revolution in the way we approach and understand genomics. As these technologies continue to evolve and collaborate, we can expect significant advances in areas such as personalized medicine, genomic research, and evolutionary biology. It is essential to recognize and seize these opportunities to maximize the positive impact of AI on DNA sequencing and, ultimately, human health and well-being.

2. Preprocessing phase

Next-generation sequencing (NGS) has transformed genomics, enabling deep, high-throughput analysis of genomic samples. However, before NGS data can be interpreted and analyzed in depth, it must undergo a series of preprocessing steps to ensure its quality and reliability. AI has been incorporated in this phase to improve the efficiency and accuracy of preprocessing.

The NGS data preprocessing process is an essential stage in next-generation sequencing analysis. Before artificial intelligence was introduced in this area, this process relied heavily on specific bioinformatics tools and manual intervention. Traditional NGS data preprocessing aims to prepare the data for further analysis, ensuring that it is accurate, reliable, and free of artifacts or errors introduced during sequencing.

- Initial Quality Control (QC): Before any other step, it is essential to perform quality control of the reads obtained from the sequencer. Tools like FastQC are commonly used to evaluate the quality of reads along their lengths. This provides an overview of the status of the sequences, allowing researchers to identify problems such as contamination, sample degradation or problems arising from the sequencing process.
- Trimming and Filtering: Once the quality has been evaluated, the next step is trimming the readings. The goal is to remove adapters, primers, and other unwanted sequences, as well as discard low-quality reads. This is done to ensure that the data used for subsequent analysis is of the highest quality possible. Tools like Trimmomatic and Cutadapt are common for these purposes.
- Alignment or Mapping: With the clean and high-quality reads, we proceed to align or map these reads against a reference genome. This process seeks to determine the exact location in the genome where each of the reads come from. BWA and Bowtie are examples of alignment tools that have been widely adopted in the bioinformatics community.
- Alignment Post-Processing: After alignment, it is crucial to refine and optimize those alignments. This may include removing duplicates, which are identical reads that may have been caused by artifacts during library preparation. It can also include local realignment to correct misalignments and quality recalibration to obtain more accurate scores. The GATK (Genome Analysis Toolkit) has been a standard resource for these tasks.
- Variant Identification: After preprocessing, the data is ready for the identification of genetic variants, such as SNPs and INDELs. However, this step, although closely related, is considered more a part of the main analysis than preprocessing.

Preprocessing NGS data without AI intervention requires great attention to detail and deep knowledge of bioinformatics. Although these traditional tools and methods have proven effective, they are inherently slower and can be prone to human error or subjective interpretation, which led to interest in incorporating automated and optimized solutions through artificial intelligence.

Currently we can only find one machine learning tool applied to the preprocessing phase: AfterQC.

AfterQC is an automatic quality control (QC) tool that uses machine learning techniques to detect and correct common errors in NGS data.

Its main advantages are that it automates the quality control process and that it can correct errors in real time during the sequencing process.

However, it may not be suitable for all sequencing types or platforms and its effectiveness is greatly affected by the quality of the initial sample.

AI provides a series of general advantages in this preprocessing phase

- Automation: Reduces the need for manual intervention, which minimizes human errors and speeds up the process.
- Improved Accuracy: AI techniques can identify and correct errors that might be missed with traditional methods.
- Adaptability: AI tools can be trained and tuned for different types of data or experimental setups.

Regarding its limitations we can highlight:

- Need for High Computing Capabilities: Many AI tools require substantial computing resources, which may not be accessible to all laboratories.
- Learning Curve: Although they automate many processes, AI-based tools can have a learning curve for those unfamiliar with them.
- Training Data Dependency: The quality and quantity of training data are essential for the performance of AI models.

The preprocessing phase is critical in the NGS pipeline, and the introduction of AI-based tools in this process has significantly improved its effectiveness and efficiency. While these tools offer numerous benefits, it is essential to be aware of their limitations and ensure that they are used appropriately and in the correct context. As technology and research in this field advance, we can expect more optimized solutions and more robust tools that continue to revolutionize the world of genomic analysis.

The alignment phase, also known as read mapping, is a crucial stage in the analysis of next-generation sequencing (NGS) data. This phase focuses on determining the exact position of a DNA read or fragment within a reference genome or transcript. The fundamental purpose is to establish a correspondence between each read obtained from the sequencer and its probable location in a known genome. Through this process, it is possible to identify genetic variants, structure the genome of an organism or analyze genetic expression, among other objectives.

The alignment process begins, broadly speaking, by taking the sequence reads produced by the sequencing machines and comparing them to reference genomes that have been previously sequenced and well characterized. The idea is that by aligning reads against a reference, you can identify regions of the genome where these reads "fit" best. This procedure is essential, not only to validate the quality of the sequence produced, but also to infer relevant biological information from the location and nature of the reads.

3. Alignment phase

The alignment process is not trivial, especially given the enormous amount of data generated by modern sequencing techniques. Reference genomes, especially for complex organisms like humans, are vast, containing billions of bases. Alignment tools must be extremely efficient to accurately map billions of reads in a reasonable time.

Among the most popular tools for aligning NGS reads are BWA (Burrows-Wheeler Aligner) and Bowtie. BWA is particularly suitable for aligning short reads to long genomes, such as the human genome. It uses the Burrows-Wheeler transformation algorithm to index the genome, allowing for fast and efficient searches. Bowtie, on the other hand, is extremely fast and also employs the Burrows-Wheeler transform, but is designed with a focus on memory efficiency, allowing it to handle even larger genomes with limited computational resources.

Once the reads have been aligned against the reference genome, files, commonly in SAM or BAM format, are generated that detail the position and orientation of each read relative to the reference genome. These files become the basis for many downstream analyses, such as genetic variant detection.

It is essential to mention that alignment is not always perfect. Reads may map to multiple locations, or not map at all, due to variations in the genome or errors in the reads. Therefore, interpretation of alignment results and subsequent analysis steps require meticulous care and a deep understanding of the biological and technical context of the data.

3.1 Contribution of AI

The use of artificial intelligence (AI) in the alignment phase is motivated by the need to improve the accuracy, speed and efficiency of read mapping, especially when faced with large or complex genomes. Advanced machine learning techniques and neural networks offer a unique opportunity to address these challenges and have shown potential in transforming the genomic alignment landscape.

Convolutional neural networks (CNN) have been highlighted as an effective architecture in the read alignment domain. CNNs are known for their ability to process and learn features from spatially structured data, such as images. In the context of genomic alignment, genome reads and sequences can be conceptualized and transformed into “images” or matrices, where CNNs can learn sequence-specific patterns and features. These learned features are used to predict the most likely position of a read within the reference genome.

For example, DAVI (Deep learning-based tool for alignment and single nucleotide variant identification) is an algorithm that uses CNNs for the alignment of genomic reads. By treating alignment as a pattern recognition problem, DAVI seeks to capture the complexities of genomic sequences through deep learning.

AI-based techniques have the potential to improve alignment accuracy by being able to recognize subtle patterns in sequences that might be missed with traditional algorithms. Additionally, AI can adapt to new forms of errors or genetic variations as they are trained on more data, offering a more dynamic and evolutionary approach to alignment. Additionally, deep learning techniques, when run on suitable hardware such as GPUs, can offer significant speedups in terms of processing time.

Despite its promising applications, integrating AI into genomic read alignment presents challenges. First, AI-based solutions require large, highly accurately labeled training data sets, which is not always easy to achieve in genomics. Second, neural networks, particularly deeper architectures, can be computationally intensive and require specialized hardware. This may not be accessible to all laboratories or research centers. Finally, unlike traditional methods, where logic and decisions are clearly defined and interpretable, decisions made by AI models can be less transparent, which can complicate the interpretation and validation of results.

The introduction of artificial intelligence into the alignment phase of genomic sequencing offers exciting opportunities to improve the accuracy and efficiency of read mapping. However, it is essential to address and overcome the limitations associated with the implementation of AI in this domain to fully realize its potential and ensure reliable and reproducible results.

4. Variant search phase

The variant searching phase, often referred to as “variant calling,” is a critical stage in the analysis of next-generation sequencing (NGS) data. The main objective of this phase is to identify places in the genome where there are differences with respect to a standard reference. These differences, or variants, can range from changes in a single base, known as single nucleotide polymorphisms (SNPs), to insertions, deletions and rearrangements of longer segments of DNA. These variants can have significant biological consequences, from disease predisposition to specific phenotypic characteristics.

Upon alignment of the NGS reads to the reference genome, a complete map of how each read relates to the reference is obtained. However, to discover variants, it is essential to examine areas where the readings differ from the reference. This is done by evaluating the stack of aligned reads at each position in the genome.

The variant search process can be divided into several stages. The first is the initial identification of places where the readings appear to differ from the reference. This preliminary identification can result in the detection of possible SNPs, insertions or deletions. These candidate variants are then further evaluated for quality and reliability. Multiple criteria are applied, such as the depth of coverage, the quality of the bases, and the balance between reads that support or refute the variant.

GATK, developed by the Broad Institute, is a sophisticated suite of tools designed for the analysis of genome sequencing data, with a particular focus on the identification of genomic variants. In the context of variant searching, GATK has set industry standards with its best practices.

The variant calling process in GATK is based on an algorithm called "HaplotypeCaller". Instead of examining each position in the genome in isolation, HaplotypeCaller jointly considers evidence across a short range of the genome to identify potential haplotypes (sets of variants that are commonly inherited together). By doing this, it reassembles the aligned reads in that interval and compares the reassembled haplotypes to the reference genome to identify variants.

HaplotypeCaller creates a probabilistic model to determine the probability that there are different variants at a given position, given the observed reads. Proposed variants are scored based on quality, and those that exceed certain thresholds are retained for further analysis. Additionally, GATK provides tools to filter and recalibrate variants based on multiple metrics to ensure that the final variants are of high quality.

SAMtools is a set of programs that allow you to interact with sequence alignments stored in SAM and BAM formats. In the context of variant searching, one of its key components is `mpileup`, which generates a summary of the alignment of reads relative to the reference genome.

The variant calling process in SAMtools begins with `mpileup` generating a set of aligned reads for each genome position. Based on the stack of reads, `mpileup` identifies positions where the reads differ from the reference. A probabilistic model is then used to determine whether these differences are significant enough to be considered genuine variants or whether they could simply be sequencing errors.

After this initial identification, variants are evaluated based on several metrics, including base calling quality and coverage depth. Variants that do not meet certain quality criteria are filtered. Unlike GATK, which uses a haplotype-based approach, SAMtools examines each position more independently, although it still considers contextual information, such as adjacent bases, to make decisions about variant calling.

Both GATK and SAMtools have robust and proven approaches for searching for variants, but they differ in their methodology and how they model and evaluate evidence from aligned reads. GATK, with its haplotype-based approach, attempts to capture broader context when calling variants, while SAMtools uses a more direct approach, examining positions and their immediate context.

It is crucial to note that while variant calling has been refined over the years, it remains an inherently challenging process. True variants can be difficult to distinguish from artifacts introduced by sequencing errors, incorrect alignments, or ambiguities in the reference genome. Therefore, a rigorous and careful approach is essential to ensure that the identified variants are true and biologically significant.

4.1 Contribution of AI

The integration of artificial intelligence (AI) in the variant search phase arises from the need to improve accuracy, reduce false positives and speed up the process, particularly in challenging genomic sequences. Advanced machine learning techniques and neural networks are being explored to transform and optimize this critical area of genomic analysis.

One of the most promising architectures in variant detection is the convolutional neural network (CNN). Although CNNs primarily originated in image analysis, their ability to detect local patterns makes them suitable for analyzing genomic sequences, where variations can be viewed as "patterns" in a genomic sequence "image." DeepVariant, developed by Google Brain, is a notable example that uses CNNs for variant calling. It takes a region of the genome, converts it into a color-coded image, and then uses a CNN to identify variants in that region.

The key advantage of using AI-based techniques, such as DeepVariant, is the potential improvement in accuracy. Because these techniques can learn and adapt to complexities in the data, they are less prone to systematic errors that could plague traditional heuristic methods. Machine learning can also be trained to recognize and correct specific sequencing artifacts, further improving accuracy. Additionally, AI-based analysis can be considerably faster, especially when implemented on optimized hardware such as GPUs.

However, there are challenges associated with using AI in this domain. First, training AI models requires large, high-quality data sets. In genomics, where true variants and sequencing errors can be difficult to distinguish, obtaining accurately labeled training data is challenging. Second, interpretability remains a concern. While traditional methods offer clear logic behind each decision, an AI model's decisions can be opaque, making results difficult to validate and understand. Additionally, the transferability of a model trained on one data type or sequencer to another may be limited.

The implementation of artificial intelligence in the variant search phase provides an opportunity to overcome some of the limitations of traditional approaches. While AI-based techniques offer improved accuracies and speeds, it is essential to be aware of the limitations and challenges inherent in these methods. The combination of human expertise, well-established heuristics and the power of AI appears to be the way forward for optimal variant calling in the future.

5. Annotation phase

The variant annotation phase is an essential stage following variant search and validation in next-generation sequencing (NGS) data analysis. Once variants are identified, it is imperative to understand their biological and clinical significance. Variant annotation is precisely this process of adding relevant information to detected variants, with the aim of providing context and potentially inferring the functional impact that these variants may have.

In essence, variant annotation is the act of enriching the raw list of variants with additional data. These data may include the genomic location of the variant (for example, whether it is within a gene, in an intergenic region, or in a regulatory region), the amino acid change (if applicable), and predictions about the effect of the variant on the function of the gene or protein. It may also include information about whether that variant has been previously observed in other populations or individuals and whether it has been associated with specific diseases or phenotypes.

Currently, variant annotation is performed using computational tools and public and private databases that store genomic, phenotypic, and clinical information. These databases contain vast repositories of known genomic variants and often have details on the frequency of those variants in specific populations, associations with diseases, and more.

One of the most widely used tools for annotation is ANNOVAR, which allows users to leverage a variety of databases and resources to annotate variants. It works by extracting relevant information from databases such as dbSNP, 1000 Genomes Project, ClinVar, among others. In addition to ANNOVAR, there are other programs and platforms, such as VEP (Variant Effect Predictor) and SnpEff, that also provide extensive functionalities for annotation.

Despite the advanced tools available, annotation is not without challenges. A variant that affects multiple transcripts or genes can have different effects depending on the context. Furthermore, predictions about the pathogenicity of a variant are not always definitive. The correlation between a specific variant and a phenotype can be complex, and not all variants found in disease-associated genes are necessarily pathogenic.

5.1. AI contribution

The increasing complexity and magnitude of genomic information, combined with the need for accurate and rapid interpretations, has created an environment conducive to the integration of artificial intelligence (AI) into the variant annotation phase. Using advanced machine learning techniques, we seek to optimize and improve the accuracy of predictions and the integration of multidimensional data during annotation.

Neural networks, particularly deep neural networks (DNNs), are emerging as potential tools for variant annotation. These networks are capable of processing information at multiple levels of abstraction, allowing for a deeper interpretation of the variant in context. CADD (Combined Annotation-Dependent Depletion) is an example that, although not based solely on AI techniques, uses a machine learning model to integrate multiple annotations into a single metric that predicts the deleteriousness of variants in the human genome.

DeepSEA, on the other hand, is a neural network-based tool that predicts the effects of non-coding variants using massive sequencing data. DeepSEA is a deep learning model designed to predict the functional effects of DNA sequence variations, particularly in non-coding regions of the genome. At a technical level, DeepSEA takes nucleotide sequences as input and, using its convolutional neural network architecture, transforms the sequence into a set of predictive features for various functional states, such as specific protein binding, epigenetic marks, and effects on gene expression. . When trained with massive genomic and epigenomic data, DeepSEA is capable of learning and recognizing patterns in DNA sequences that are associated with specific functions. This approach allows DeepSEA to identify how specific variants, even those in non-coding regions of the genome, can influence gene regulation and ultimately cell function or organism phenotype.

One of the main advantages of AI in variant annotation is its ability to handle large amounts of data and heterogeneous information sources. As more data accumulates, AI-based methods can constantly improve their accuracy. Additionally, AI models, particularly those based on deep learning, can capture nonlinear interactions and complex relationships that traditional approaches might miss. This is especially relevant when considering epistatic effects or interactions between variants.

However, implementing AI in this phase is not without challenges. The quality and quantity of training data remain crucial. Training AI models requires large sets of labeled data, and if that data contains errors or biases, those problems will be reflected in the resulting model. Additionally, the "black box" or opaque nature of AI models is a concern in clinical genomics, where traceability and justification of decisions are vital. Finally, the adaptability of these models can be a challenge; Models trained for a specific data set or population may not generalize well to other contexts.

Although incorporating AI into the variant annotation phase presents exciting opportunities to improve the accuracy and efficiency of the process, addressing its limitations and challenges is crucial for widespread adoption. With interdisciplinary collaborations and continued development of algorithms, AI has the potential to revolutionize the way we interpret and understand the human genome in clinical and research contexts.

6. Conclusions

Next-generation sequencing (NGS) has revolutionized the genomics landscape, offering an unprecedented view of the diversity and complexity of the human genome. Despite their enormous advances, there are areas in current tools and approaches that require refinement and optimization. Artificial intelligence (AI) is presented as an innovative solution, capable of addressing some of these challenges and further enhancing the capabilities of NGS.

Current tools for NGS, although advanced, still face challenges in terms of accuracy, speed, and ability to handle high-volume data. A notable concern is the handling of complex genomic regions, such as those rich in GC or regions with tandem repeats, where traditional tools often struggle to obtain accurate and consistent reads. Structural variants, such as duplications, deletions, and inversions, are also challenging to detect and accurately characterize using conventional approaches.

Furthermore, the interpretation of variants, especially those in non-coding regions or in understudied genes, can be ambiguous. Current tools can provide conflicting or inconclusive annotations, making it difficult to make clinical decisions based on this data.

Scalability is another challenge. As the volume of data generated by NGS continues to grow, tools must be able to process and analyze this data efficiently, without compromising accuracy. Furthermore, the integration of different types of data, such as transcriptomics, proteomics, and epigenomic data, is still a complex task with conventional tools.

Artificial intelligence has the potential to directly address many of the challenges mentioned above. For example, in complex genomic regions, deep learning models can be trained to recognize patterns and anomalies with much higher accuracy than heuristics-based methods. These models can be especially useful for the detection of structural variants, identifying subtle patterns and genomic contexts that traditional tools might miss.

In terms of interpretation, AI can integrate and analyze information from multiple sources simultaneously, offering a holistic view of the variant in question. With access to global databases and scientific literature, AI models can provide more informed, consistent and up-to-date interpretations of variants.

From a scalability standpoint, AI, especially when implemented on specialized hardware like GPUs, can process massive volumes of data at significantly faster speeds than traditional tools. Furthermore, AI's inherent ability to handle heterogeneous data facilitates the integration of different data sets, providing a richer, multidimensional understanding of the genome.

As the field of genomics advances, the need for more accurate, faster, and adaptive tools becomes increasingly pressing. Artificial intelligence, with its ability to learn and adapt, offers promising solutions to address the challenges inherent in existing NGS tools. This synergy between genomics and AI has the potential to not only improve our current analytical capabilities, but to open new frontiers in genomic medicine and research. With continued investments in research and development, we are likely to see significant advances in this interdisciplinary space in the coming years.

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

Acknowlegments

This study has been funded by the AIR Genomics project (with file number CCTT3/20/SA/0003), through the call 2020 R&D PROJECTS ORIENTED TO THE EXCELLENCE AND COMPETITIVE IMPROVEMENT OF THE CCTT by the Institute of Business Competitiveness of Castilla y León and FEDER funds.