Photo: DC Studio (AdobeStock)

10

Applications of bioinformatics in agriculture

Adhemar Zerlotini Neto| Antonio Nhani Junior | Fábio Danilo Vieira | Leandro Carrijo Cintra | Maurício de Alvarenga Mudadu | Paula Regina Kuser Falcão | Poliana Fernanda Giachetto

Introduction

Biotechnology has been fundamental for the progress observed in Agriculture over the last 30 years. Bioinformatics, the multidisciplinary area responsible for analyzing the large volume of data resulting from genomic technologies, was essential in this progress. With the arrival of next-generation sequencing technologies, an extraordinarily large volume of genomic data that needed to be analyzed was produced. In the era of digital transformation, the ability to generate biological data more rapidly, more affordably and in greater volume, produces an enormous amount of data, Big Data. This large and growing volume of data requires solutions in at least three spheres: scalable infrastructure, data management and intelligent use of that data.

Bioinformatics uses computational tools to answer complex biological questions and contribute to innovative results. The theme involves the use of high-performance computing infrastructure and tools to organize, analyze, integrate, process, simulate and store large volumes of data derived from *in vivo* and *in vitro* experiments. A challenge for bioinformatics is to integrate the heterogeneous data generated by the "omics" sciences (both with each other and with the data generated by traditional sciences), allowing discoveries that go beyond what is possible in each of the individual disciplines. Several new layers of omics, such as analysis of genomes, metabolomes, transcriptomes, or interactomes, have become important for research advances. The integration of all this information allows making discoveries and improving the knowledge of biological systems.

Access to high storage and processing capacity, with powerful indexing algorithms, as well as machine learning applications, is crucial for the execution of bioinformatics activities. More importantly, a trained and constantly updated team to assist in planning data generation processes, data analysis and extraction/acquisition of new knowledge from Big Data is what will enable Embrapa to be a relevant actor in this area of knowledge.

In this context, in 2011, Embrapa's Multiuser Laboratory of Bioinformatics (LMB) was created to provide bioinformatics support to RD&I projects aligned with Embrapa's strategic objectives. Since its creation, the LMB has already contributed in a broad portfolio of projects, within three operating guidelines:

- **Access to the computing park**, hence its high-performance infrastructure.
- **Consulting in the analysis of biological** data that require high-performance computing, whether due to the volume of data or the complexity of the analyses.
- **Training** for multiplying skills through courses and other training actions.

The LMB has performed research projects at Embrapa and partner institutions that involve more than 20 crops and livestock systems studied in more than 50 research projects. An important aspect in bioinformatics is that each project is unique, and the LMB team works to meet these demands. This work in bioinformatics is based on the following areas: analysis of gene expression, assembly and analysis of genomes, identification of molecular markers, analysis of transcriptomes and metagenomes, evolution studies, modeling of biological systems, prediction of protein structures and molecular interaction, interaction or inhibition of molecules, among other activities.

LMB computational infrastructure to support bioinformatics projects applied to agriculture

Bioinformatics projects require a differentiated computational infrastructure, and most of them are very difficult or even impossible to carry out using common computational equipment. These requirements can be understood considering the computational complexity of the algorithms and the volume of biological data analyzed.

The objective of this session is to present the computational infrastructure used for storage and processing of large volumes of data produced by the biotechnology research projects of Embrapa and its partner institutions. This infrastructure focuses on making available processing and memory capacity as well as storage of large volume of data.

To deal with the various algorithms with high computational complexity in bioinformatics, it is standard to use computational clusters of computers for data processing. For those less familiar with the field of high-performance computing, a computer cluster is a set of computers connected in a network with a central coordination node that work together to solve computational problems. The main advantage of a cluster is to provide computing power of tens, hundreds and, in some extreme cases, thousands of processing nodes in a transparent way for the user, that is, without the user having to interact and trigger data analysis in each of the machines individually. The jobs to be executed in the system are activated from a management node that remain in one or more execution queues and are automatically sent to a suitable processing node, when available.

With the arrival of multicore computing, each processing node in modern clusters has a few dozen cores; in some exceptional situations, each node can reach hundreds of processing cores. Therefore, a very important question for processing in bioinformatics is: how much memory should each processing node have? The answer requires careful consideration regarding that as the amount of memory is directly proportional to the number of cores in the processing node. In addition, it should be considered that this proportion has increased with the development of new biological investigation techniques, which generate significantly increasing amounts of data. Thus, until recently, it was recommended that each processing node should have 8 Gb of RAM for each available core. With the significant increase in the volume generating biological data, this amount was updated, and new processing platforms for bioinformatics activities are being developed with 16 Gb of RAM for each available CPU core in the computing node.

Another relevant issue in biological data processing platforms is related to data storage and preservation. Basically, the most significant bottleneck that has to be addressed is the amount of data to be stored. The speed of data accessing does not significantly impact the performance of the platforms, as in general, the tools and programs executed to perform the analyses will load the data into memory and execute the analyses for a significant amount of time. A delay in the initial load does not considerably impact the total execution time of the task. However, a restriction on the storage capacity of the computing environment will have a wide range of negative occurrences. It is not possible to execute several projects at the same time, as they commonly demand a few hundred gigabytes, and can reach a few tens of terabytes for raw data storage for some exceptional projects. During the analyses, it is necessary to store intermediate data, possibly up to an order of magnitude of the original data size. Therefore, currently the platforms for processing biological data commonly use storage systems with capacity of a few petabytes.

The processing environment available today has a cluster with a head node and 14 processing nodes. Of these, 13 have 64 cores and 512 Gb of RAM each. There is also a special node that is used to perform jobs that require a large amount of memory. This node has 2 Tb of RAM and 160 processing cores. In total, the cluster provides 992 processing cores. For managing tasks in the cluster, a queue management system is used, initially developed by Sun Microsystems, known as Sun Grid Engine (SGE). For bioinformatics analyses, a high performance cluster is specially useful as the analyses, in general, involve multiple datasets to be processed in pipelines consisting of multiple stages, enabling the execution of computing tasks in parallel on separate machines. Computational analyses with such characteristics are ideal to be executed in clusters of computers.

The following data storage servers are available: an SGI Infinite storage with 150 Tb capacity in a RAID 6 configuration and an IBM DS3412 storage capable of storing 51 Tb in a RAID 5 configuration. In addition to primary storage, it is critical to have a backup policy that ensures data security on the platform. Due to the volume of data constantly received and generated, the most cost-effective methodology for backup involves the use of LTO tapes. Currently, the platform has a tape library available with capacity for 44 LTO6 drives. As each LTO6 tape provides, on average, 6.25 Tb of data storage, the total library is capable of handling up to 275 Tb of online backup.

This type of computational infrastructure is essential for carrying out data analysis of bioinformatics research projects in agriculture.

Applications

Bioinformatics and the tambaqui production chain

Embrapa's first strategic objective is "to develop knowledge and technologies for the adequate management and sustainable use of Brazilian biomes." Historically, Embrapa has always been concerned with regional development, actively performing on front lines where scientific or economic risks were discouraging factors for the private sector. This exceptional role played by Embrapa has guaranteed the use of the Cerrado biome for agriculture, bringing development and wealth to the region. The North region of Brazil has a fish production chain with an annual native fish production of 290 thousand tons, according to the 2019 Pisciculture Yearbook, and the main product is tambaqui (*Colossoma macropomum*). To promote the development of this important production chain, among other equally relevant objectives, Embrapa, through the BRS Aqua[1] project, identified critical points for increasing the production that, if properly resolved, would increase the competitiveness and sustainability of the tambaqui production chain.

One of the critical points identified by Embrapa in the tambaqui production chain[2] was the occurrence of crossbreeding between related matrices. Many fish farmers do not know this, but the simple choice of matrices for crossbreeding can, if wrongly done, reduce the final weight of fish by 10% to 30%. In other words, using the same amount of feed in the food, the producer could lose up to 30% of food conversion. In the scientific literature, this phenomenon is known as inbreeding depression, and few fish producers are aware of this. To measure the size of the problem, let us observe the case of native fish, which are highly appreciated in the North region. As mentioned, in 2019, the production was 290 thousand tons, assuming a conservative estimate, as the inbreeding of related matrices may have negatively impacted production by at least 30 thousand tons.

In addition to inbreeding depression, the inbreeding between related matrices causes yet another harmful phenomenon, scientifically known as lethal alleles. In any population, lethal alleles are rare; however, when they occur in homozygosis, they impair embryo development. That is, these alleles cause deformities in embryos or abort their development when inherited from both the father and the mother. Hence the recommendation to avoid consanguineous pairings. If these alleles are rare in the population as a whole, within families carrying these alleles the occurrence of homozygosity is significantly more frequent, reaching up to 25%. That is, in consanguineous breeding, up to 25% of embryos can be lost or have birth defects. Both inbreeding depression and lethal alleles are critical problems in the fish production chain.

In addition to inbreeding depression and lethal alleles, another critical point is the existence of fertile hybrids in the breeding stock. In biology classes, one learns that when two different species interbreed, the result is an infertile animal. Unfortunately, this is not always true for fish. For example, the tambaqui can crossbreed with the pacu (*Piaractus mesopotamicus*), and the hybrid is a fertile animal. However, many producers crossbreed tambaqui with pacu because the hybrids gain more weight than purebred animals and the flavor of the meat is not significantly affected. In the literature, this phenomenon is known as "hybrid vigor", and it is widely used in grain production, for example. The problem occurs when hybrids are wrongly chosen to compose the breeding stock. While this choice may seem unlikely at first, it occurs because selection is often based on external characteristics, and because of hybrid vigor it is not uncommon for a hybrid to be wrongly selected because it weighs more, for example. In this case, as the hybrids are fertile, the error of this choice will only be discovered during crossbreeding, when the producer observes the natural segregation that entails a great deal of variability in the economic interest characteristics, such as slaughter weight. Producers who sell fingerlings for fattening may have their credibility affected by selling low quality animals, as the segregation variability greatly affects fattening.

[1] The BRS Aqua project is financed by the BNDES/Funtec Technological Fund, the Fisheries and Aquaculture Secretariat (SAP) of the Ministry of Agriculture, Livestock and Supply (MAPA), CNPq, FAPDF and Embrapa. In this part of the BRS Aqua project, the following Units largely participated: Embrapa Genetic Resources and Biotechnology, Embrapa Fisheries and Aquaculture and Embrapa Agricultural Informatics.

[2] This critical point occurs in all fish production chains where it is not possible to identify the relationship between the matrices.

Once these problems were identified, Embrapa researchers developed two DNA chips that solve such issues in an innovative, efficient and low-cost way. These chips have molecular markers, known as single nucleotide polymorphisms, or simply SNPs, that can provide enough information to determine the degree of relatedness and purity of the species. In the case of kinship, the markers must have considerable variability in the population studied. Mathematically, this means requiring the Minor Allele Frequency (MAF) to be close to 0.5. The principle is exactly the same as a paternity test, except that this application can identify any degree of kinship to avoid inbreeding, reducing inbreeding depression and minimizing the occurrence of lethal alleles. The scientific challenge is to precisely choose such SNPs molecular markers. In the case of tambaqui, the lack of a publicly available reference genome was the first obstacle to be overcome. This led Embrapa to carry out an internal Tambaqui Genome Project, and the LMB was responsible for assembling the Tambaqui Genome. The genome contains approximately 1.3 billion nucleotides divided into 27 chromosomes (or linkage groups). Once the genome was ready, the next step was to select a representative subpopulation of the tambaqui population and then sequence the DNA from the pool of that subpopulation. The result of this sequencing was mapped to the reference genome, and finally the discovery of SNPs was carried out. Although a minimum coverage of 150X was required, more than 2 million SNPs were identified (Ianella et al., 2019). The task of selecting 96 SNPs to compose the kinship chip took into account the MAF, the spacing within the chromosomes, the functional annotation and, finally, the absence of genomic variations in the flanking regions of the candidate SNP. As noted, the bioinformatics work was very intense in order to carry out all these tasks, which justifies the need for an infrastructure like the LMB. After the validation phase of the SNPs in a different population from that used in the previous phase, the validated SNPs were incorporated into the chip, which proved to be extremely efficient in determining the degree of relatedness and is currently being used in the tambaqui production chain. In other words, the producer already has an innovative tool to eliminate inbreeding depression and lethal alleles, thus avoiding silent damage caused by inbreeding.

The DNA chip for purity determination, however, required more complex analyses. This is because it was necessary to include in the analysis two more species that crossbreed with tambaqui and produce fertile hybrids, namely, the pacu and the caranha (*Piaractus brachypomus*). As none of these species has a reference genome, it was necessary to use the tambaqui genome as a reference. This procedure is important because, in addition to intraspecies variations, there are also interspecies variations (tambaqui x pacu / tambaqui x caranha), which increases the degree of complexity of the analyses. Even at the stage of mapping the reads in the reference genome, the similarity requirement had to be reduced due to interspecific differences. Differently from relatedness SNPs, SNPs to measure the purity of the species must be “fixed”, that is, they must not show species variation, that is, MAF = 0.

An example can help to better understand the problem. If at a specific position in the genome there is an “A” nucleotide fixed on the tambaqui, and in that same position there is the “C” nucleotide fixed on the pacu, then this genome position is a serious candidate to compose the purity chip of species, because, in a DNA test, an “A” result would mean “tambaqui” and a “C” would mean pacu. And to further reduce costs, genomic markers capable of simultaneously separating the tambaqui from the other two species were explored. In the previous example, this would mean that the caranha also had a “C” fixed to that same genomic position[3]. Thus, with a single DNA chip, it is possible to assess the purity of tambaqui in relation to the two main species that produce hybrids[4]. Once again, using allelic frequency, physical spacing in the genome and functional annotation, 96 SNPs were selected to compose the chip, and after the validation

[3] SNPs are biallelic markers, which enable separating one species from two others simultaneously. There are triallelic SNPs, but they are very rare, and therefore it is not possible to produce a single genotyping chip that separates the three species two by two simultaneously.

[4] From what has already been shown, this purity chip does not separate pacu from caranha.

phase in independent populations, the validated SNPs were incorporated into the purity measurement chip. This genomic tool enables us to eliminate all the hybrids that were wrongly chosen to compose the breeding stock.

Economic impact studies carried out by Embrapa, assuming an average production of 150 thousand tons of tambaqui, forecast additional gains between US$ 1,8 million and US$ 5,5 million for producers[5]. Each sample analysis for purity and relatedness currently costs US$ 12.00. For a producer with 100 matrices, this would be equivalent to an investment of US$ 2,4 thousand. As each matrix has a useful life of three years, this amount is amortized over an equal period. These two technologies, named TambaPlus[6], have already been adopted by producers in five states: Mato Grosso, Tocantins, Roraima, Amazonas and Rondônia, and more than 1,500 tests have already been performed. The TambaPlus is so important that the technology was selected to compose a select group of technologies that were highlighted at the 47th Anniversary of Embrapa[7].

Research on the tambaqui production chain will continue. There is still plenty of room to improve fish production. In any genetic improvement program, there are two main phases, namely, the Selection and the Crossing phase. Tambaqui is still at an earlier stage, known as pre-breeding. The main concern was to first avoid inbreeding and the presence of hybrids in the breeding stock.

Bioinformatics in vaccine development: reverse vaccinology

In animal production, the use of vaccines is an effective and low-cost alternative for preventing or reducing the severity of diseases that affect livestock. Vaccination contributes to maintaining animal health and welfare, increasing the efficiency of food production and reducing the transmission of zoonoses. Compared to other forms of control, for instance the use of antibiotics and pesticides, vaccines have advantages, such as the non-contamination of the environment and animal products (meat, milk and eggs).

Following conventional vaccine development methodology, the pathogen is cultivated in vitro in the laboratory and used in its attenuated form (in which it loses the ability to cause disease) or killed to elicit a protective immune response in the host. Alternatively, purified components of the pathogen can also be used as antigens in the subunit vaccines (Rappuoli; Covacci, 2003).

Although conventional obtained vaccines are among humanity's most important inventions, which comprise a powerful tool in the fight against disease-causing biological agents, not all pathogens can be cultivated *in vitro* and used in the development of vaccines, in its conventional form. In addition, conventional methods are quite time-consuming, and it may take five to 15 years to obtain an effective vaccine (Vernikos, 2008).

Reverse vaccinology, a methodology first published by Rappuoli (2000), emerged as an alternative strategy for the discovery of protective antigens for developing vaccines based on the analysis of the target pathogen's genome. Made possible by large-scale gene sequencing, along with the development

[5] News provided in a video conference entitled TambaPlus®: Genomic tools for the analysis and management of tambaqui matrices for the production of fingerlings, available on the Agrotins platform. Available at: https://agrotins.to.gov.br/programacao/tambaplus-ferramentas-genomicas-para-analise- e-gestao-de-matrizes-de-tambaqui-destinadas-a-produca.html.

[6] Available at: https://www.embrapa.br/busca-de-noticias/-/noticia/46203188/ferramentas-genomicas-ajudarao-a-evitar-cruzamentos-consanguineos-entre-matrizes-de-tambaqui

[7] Available at: https://www.embrapa.br/47-anos/solucoes-tecnologicas-em-destaque?link=47-anos

of bioinformatics tools, reverse vaccinology uses *in silico* prediction tools to identify targets (antigens) for developing vaccines. Through these tools, genomes, transcriptomes and proteomes are examined *in silico*, predicted proteins are selected based on desirable attributes – which can induce an immune response capable of protecting against a given disease, and the targets are then identified. Based on them, different types of vaccines can be designed and developed within an interval of 1 to 2 years.

Commercial vaccines obtained through this methodology are already a reality. A vaccine developed against invasive meningococcal disease, caused by the bacterium *Neisseria meningitidis* serogroup B, was released for use in Europe in 2014 (Andrews; Pollard, 2014). In this vaccine, the immune response is triggered by epitopes – specific sequences of amino acid residues present in the antigen that directly participate in the interaction with antibodies, which were identified using bioinformatics tools. Epitopes have been considered particularly interesting in vaccine development, as it has been shown that vaccines composed of these peptides can optimize or exceed the protection potential induced by the cognate native protein (Kao; Hodges, 2009). In contrast to live attenuated vaccines, a vaccine containing a synthetic epitope is not able to reverse the virulence of a pathogen (Palatnik-De-Sousa et al., 2018). Furthermore, epitope-based vaccines are more specific, do not induce undesirable immune responses, are capable of generating long-lasting immunity and are less expensive than conventional vaccines (Ahmad et al., 2016).

In the reverse vaccinology approach, the protein sequences of an organism are analyzed using *in silico* prediction programs. These proteins, however, are mostly predicted from the sequencing of genomes and transcriptomes, using bioinformatics tools. This is because large-scale genetic sequencing, made possible by new technologies that have dramatically reduced the cost of generating sequences, as well as exponentially increasing the number of sequences generated from a sample, has accumulated an unprecedented amount of genomic and transcriptomic data. On the other hand, a technological advance that would allow a large-scale development of protein sequencing techniques with high sensitivity has not yet taken place. Methodological progress for obtaining expressed gene sequences caused the subsequent evolution in analysis methodologies. A list of programs can be accessed on the "List of RNA-Seq bioinformatics tools" page (Wikipedia, 2020). We will now provide a brief commented description of the methodology applied to obtain differentially expressed genes in the salivary gland of the bovine tick (Andreotti et al., 2018). All tools cited are obtained through an academic license or government research institution or are freely distributed.

In order to better understand the host-parasite interaction and identify possible genes and mechanisms involved, a study initiated in 2015, funded by Embrapa, generated more than 600 million sequences from RNA sequencing (using the RNA-Seq methodology) from larvae, nymphs, salivary gland, intestine and ovaries of the cattle tick, *Rhipicephalus (Boophilus) microplus* (Andreotti et al., 2018).

In addition to the characterization of transcriptomes from different tissues through *de novo* assembly, our research group also identified the differentially expressed genes (DEG) between ticks grown in resistant cattle (Nellore), susceptible cattle (Holstein) and crossbred animals with intermediate resistance to the parasite (Nellore x Holstein). The analysis of this dataset, using tools that inform the function of proteins predicted from DEG and the biological pathways in which they act, brought new discoveries about the cattle tick interaction and pointed out potential candidates that can be used as antigens in the development of vaccines to control the cattle tick (Giachetto et al., 2020).

The first step in RNA-Seq analysis is to check the quality of the generated sequences. Tools like FastX Toolkit (FastX-GitHub, 2020) and FastQC (FastQC-GitHub, 2020) check several parameters, highlighting the following:

- Average quality of bases and average quality per sequence. For a good result, the sequence must have a "Phred score" greater than 30.

- GC content (%GC). The percentage of the presence of Guanine and Cytosine nucleotide bases in the sequence must be close to the normal distribution, since the very high GC content prevents the synthesis and, often, the clustering of sequences during the acquisition and assembly processes.
- Number of indeterminate bases (%N). Indeterminate bases make the contingency process difficult. They can occur at the beginning of the sequencing, where there is a saturation of reagents; in the end, by decreasing the concentration of reagents; or in a region with high %GC, which hinders reading the region by the polymerase.
- Presence of adapters. Adapters are short nucleotide sequences used for library preparation and sequencing. Its presence impairs contingency, giving rise to chimeric sequences. To eliminate them, tools such as Trimmomatic (Bolger et al., 2014) and Trim Galore (TrimGalore-GitHub, 2020) are often used.

As we deal with a large number of sequences, a great tool to group and visualize the data obtained in the quality analysis (and subsequent steps) is the MultiQC (Ewels et al., 2016), which organizes the results obtained in a Web page.

After verifying the quality of the sequences, we proceeded to obtain the transcriptome, through sequence-to-sequence comparison and their contingency by similarity. Several tools can be used in this step, such as QUAST (Gurevich et al., 2013), which is recommended for the analysis of metagenomes. The tool of choice for analyzing this work was the Trinity program (Grabherr et al., 2011). This tool is, in fact, a pipeline that brings together, through scripts developed in the programming languages Perl[8] and Python[9], various analysis tools for quality, sequence contingency and statistics to identify DEGs, with the differential of being able to identify isoforms (the same as transcribed) of the same gene, resulting from alternative splicing. Different tissues can express different isoforms in different amounts. Identifying the locally expressed isoform allows to better understand the expression of a particular gene in a particular metabolic pathway or tissue.

Once the transcriptome is obtained, the next step is to verify the quality of the assembly. An initial approach is to map the sequences used for assembly back to the transcriptome obtained. In a good setup, more than 80% of the sequences map the transcriptome. A second assessment consists of identifying and quantifying complete sequences, through similarity analysis against curated databases, such as SwissProt or TrEMBL (The UniProt Consortium, 2019), or searching for orthologs present in the closest classification of the studied organism, in this case, the arthropods, using the BUSCO software (Seppey et al., 2019).

Several factors influence the experimental design of an RNA-Seq assay for the identification of DEGs:

- In the preparation of samples, from extracting total RNA to obtaining libraries for sequencing, the batch effect may occur, in which they are included from using different solutions (made on different days) to the person who prepares them (Conesa et al., 2016).
- The sequencing depth (the number of generated sequences), which influences the number of sequences obtained and, therefore, the quantification of the number of identified DEGs (Conesa et al., 2016; Lamarre et al., 2018).

[8] Available at https://www.perl.org

[9] Available at: https://www.python.org

- The number of technical replications (how many times the same sample is sequenced), which influences the statistical power for detecting DGEs, recommending no less than three repetitions (Conesa et al., 2016), although a higher number (about six repeats) can increase the representativity of transcriptome sequences (Lamarre et al., 2018). An assay with triplicates is commonly accepted, as the increase in replicates implies an increase in assay costs.
- The preparation of a biological repetition. Conesa et al. (2016) point out that biological variability is particular to each assay, and although difficult to control, it is important for a study involving populations, suggesting that the biological sample be done in triplicate. Lamarre et al. (2018) point to the detection of up to 20% of DEGs due to biological variability, which may not justify raising the costs of the assay.

The correlation between the samples used in the assay is also an important measure of the quality of the assembly and the libraries constructed. The principal component analysis allows visualizing correlations between technical and biological replicates, which should preferably form not too distant clusters. A discrepancy between samples from the same group may indicate contamination, sample mixture, sequencing error or batch effects, which must be considered for discarding that sample. Also important is the fact that without a technical triplicate, a biological duplicate must be discarded, impairing the entire analysis.

With a good quality transcriptome, the differentially expressed sequences were identified. Trinity incorporates several statistical tools for this purpose. In this case, we chose to use RSEM (Li; Dewey, 2011), which estimates the quantity of each transcript by realigning the sequences of each library (or experimental treatment) to the generated transcriptome – a reason for the importance of quality and the relationship between replicates - and edgeR (Robinson et al., 2010), a package developed in the statistical program R (R Core Team, 2020) and part of the Bioconductor Project (Huber et al., 2015) for the analysis of biological data, which performs the pairwise comparison sequences generated between all samples and identifies those with differential expression.

The penultimate step is the annotation (or identification) of each differentially expressed sequence, through similarity analysis in nucleotide and protein sequence databases, looking for homology to already known sequences, and in databases of metabolic pathways that inform in which one the gene participates. This is followed by a manual analysis of each result, the bibliographic basis seeking the role of such a gene in the development of the tick's life cycle, and the selection of possible targets for vaccine development.

The existence of commercial vaccines available for the control of bovine ticks demonstrates they can act effectively in the control of infestations, reducing the application of acaricides. However, the adoption of these vaccines has been limited, mainly because they are not effective against all life stages of the parasite, in addition to their low efficacy against some regional strains of *R.* (*B.*) *microplus* (Andreotti, 2006). The results obtained in a test conducted by Embrapa with a regional tick isolate showed an efficacy of 46.4% and 49.2%, respectively, for the TickGARD® and GavacTM vaccines (Andreotti, 2006). Thus, based on the database described above, our team is currently coordinating a study that foresees the identification of candidate immunogenic epitopes for the development of vaccines against cattle ticks, using the methodology of reverse vaccinology, based on the predicted proteins of the transcriptomes of the parasite. Executing a pipeline containing a series of analysis tools, candidate target genes for vaccine production are analyzed for the presence of epitopes that can interact with the bovine immune system for the production of antibodies, helping to fight to tick infestation.

A highly effective vaccine, integrated in cattle tick control strategies, can considerably reduce herd infestations and the implications related to the use of acaricides, which include, in addition to cost and environmental contamination, a growing concern with food safety, which has increasingly led to the consumption of food free of chemical residues, obtained from sustainable production systems. Also, by validating the pipeline we are proposing, the LMB will be able to apply the reverse vaccinology methodology in the identification of targets for the control of other problems of interest in agriculture.

Bioinformatics tools

As advocated by digital agriculture, to be transformed into useful knowledge, information generated from biological experiments must be accessible and, when possible, made available on the Internet. Bioinformatics and computational biologists have been dealing with this scenario for more than a decade in an environment with adequate infrastructure like the one described above, and have implemented software libraries, toolkits, platforms and databases to achieve success in this matter.

Several data analysis tools are used in Embrapa's LMB, and a search for a data integration solution became necessary. Analysis results are carefully stored in a directory structure and reports are generated. Some tools generate results in a format already available for the Internet or they can be executed directly online. Two tools under development have greatly contributed to the integration of the generated data and the transformation of these data into information.

Machado: a genomic data integration framework

In 2017, the PlantAnnot project was started to discover candidate proteins to use in pipelines for the development of transgenic plants (Prado et al., 2014; Napier et al., 2019) that are resistant to abiotic stresses. It aimed to develop a bioinformatics system applied to the discovery of genes related to abiotic stresses in plants, focused on climate change. In this project, a large volume of genomic data was extracted from public databases. The extracted dataset corresponds to 53 plant genomes, totaling more than 1.8 million genes and more than 2.3 million proteins. These data were used to perform computational analyses in order to select 72,000 proteins of interest for the pipelines. One of the project goals was to store and make available the data and analyses performed.

To solve this problem, the Machado open-source software was developed. Machado is a genomic data integration framework written in Python[10] that allows research groups to store genomic data, and also offers interfaces for navigation, searches, and visualization. Machado uses the BioPython library (Cock et al., 2009) which supports the vast majority of file formats and programs used in bioinformatics. In addition, Python has become one of the main programming languages in the data sciences area (Millman; Aivazis, 2011), and Machado can also benefit from the tools in this area. This framework uses the Chado database schema and therefore should be very intuitive for current developers to adopt or execute Machado on existing databases.

GMOD's biological relational database schema, Generic Model Organism Database Project[11], known as Chado (Mungall; Emmert, 2007), is one of the few open-source initiatives that has achieved relative success in the community. Many software systems can connect to it, such as Gbrowse (Stein et al., 2002), Jbrowse (Skinner et al., 2009) and Apollo (Lee et al., 2013), which are important tools for visualization and

[10] Available at: https://www.python.org

[11] Available at: http://www.gmod.org

annotation of genomes. There are some data integration tools that use Chado as a database schema or that can extract data from that schema, but they were developed in programming languages not often used in bioinformatics (Kalderimis et al., 2014; Spoor et al., 2019).

Machado has several data loading tools for genomic data and for analysis results from known software in the biological environment (BLAST, InterproScan etc.) (Altschul et al., 1990; Quevillon et al., 2005), and its web interface contains a powerful search tool that allows users to quickly filter and sort the results.

Within the scope of the PlantAnnot project, a tool called Plant Co-expression Annotation Resource was created using Machado to store and make available the data of the project[12]. This tool is an implementation of Machado, which is an example of its usefulness for researchers who need to store and make accessible a large volume of genomic data.

As an example, one of the uses of the Plant Co-expression Annotation Resource is to enable navigation through the genome of 53 species of angiosperm plants, which allows visualizing details about genes, proteins and RNA through the JBrowse genome browser. This tool is also used to perform keyword searches and use filters. The user can then perform simple searches for genes, proteins and RNA, using keywords of interest. Furthermore, it can also add more complex filters to the search results, producing more specific result lists. For example, a set of proteins with no known function, candidates for the creation of transgenic plants resistant to abiotic stresses, such as drought, heat, cold, among others.

Machado is meant to be a modern object-relational framework that uses the latest Python modules to produce an effective open-source program for genomic research and can be an engaging project for new developers, contributors and users. Thus, we created a corporate account for LMB on GitHub, which we believe is the first Embrapa account on this platform[13]. A demo version of the system was also created[14].

Machado will undergo improvement phases for ongoing projects at Embrapa, such as the project "The Hologenome of Nelore: Implications for Meat Quality and Food Efficiency" which is focused on genomic improvement of cattle, led by Embrapa Southeast Livestock. This project intends to identify molecular mechanisms related to meat tenderness, therefore, several data sets that need to be integrated were produced, such as genomes, transcriptomes, proteomes, genotyping, among others.

DBGAP: web system for retrieving information on pedigree, phenotypes and genotypes

The development of large-scale genotyping technologies of molecular markers such as the Single Nucleotide Polymorphisms (SNP) – to estimate the genomic profile of animals – has enabled developing genome-wide association studies – GWAS at a genomic scale, as well as the introduction of genomic selection technology in genetic improvement programs. Current technologies for generating molecular data are capable of genotyping tens to hundreds of thousands of SNP markers, in a single assay for each individual, with enormous speed and automation (Caetano, 2009).

On the other hand, this situation implies the need to store an enormous volume of data, not only of genotypes, but also the phenotypes and pedigree of an increasing number of animals. Therefore, performing the proper storage and extracting useful knowledge from this amount of data is a major challenge. Given the volume of data stored, an important matter to consider when developing a computational solution is the suitability of database modeling for the desired application, as this will

[12] Available at: https://www.machado.cnptia.embrapa.br/plantannot

[13] Available at: https://github.com/lmb-embrapa

[14] Available at: https://www.machado.cnptia.embrapa.br/demo_machado

directly impact the query and writing times in relational database management systems (RDBMS) where this information will be stored.

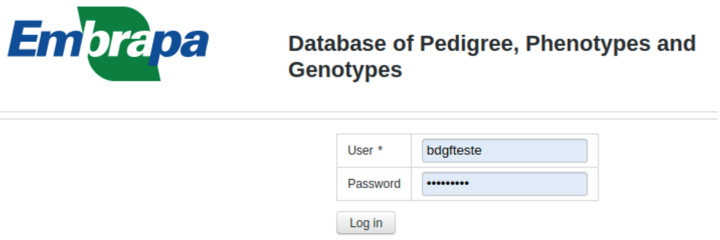
Therefore, in order to provide a solution that would be efficient both in storage and in the integration and querying of this high volume of data, the Database of Pedigree, Phenotypes and Genotypes (DBGAP) system was developed. The purpose of this system is to integrate the data sent in various formats, so they can be analyzed in genetic/genomic evaluation software. The DBGAP was initially developed using a data diagram proposed by Higa and Oliveira (2015). This diagram was redesigned in such a way as to allow implementing the JavaScript Object Notation (JSON) type. With the implementation of JSON and text types in some tables, it was possible to use the Not Only SQL[15] (NoSQL) approach to store part of the data, streamlining queries that would need to perform joins with other tables.

To develop the system, information technology components were chosen within the philosophy of using free software. The database management system chosen was PostgreSQL[16], as it is a reliable DBMS, widely used in the market. The version control software, GitLab[17], hosted at Embrapa, was used. The programming language chosen was Java[18] and its components of the Java Enterprise Edition (Java EE) technology.

Among the Java EE technologies available and used by DBGAP, the Java Server Faces (JSF) framework stands out. The architecture of the JSF framework employs the MVC model (Model, View, Controller), which separates the presentation and application layers. The application server chosen to host the DBGAP system was WildFly[19].

The system development project used some concepts from Scrum, which is an agile framework to perform complex projects. Scrum combines monitoring and feedback activities, generally through quick and daily meetings with the entire team, in order to identify and correct any deficiencies in the development process. In addition, the Scrum method is based on fundamentals such as: small teams, unknown requirements and short iterations, these are called sprints (Schwaber, 2004).

The DBGAP system has many features implemented and is currently in the process of user's approval. Through its web interface, it is possible to query and import phenotypic, genotypic and pedigree data of various animal species. When accessing it, the login page will be displayed (Figure 1):



Database of Pedigree, Phenotypes and Genotypes

User * bdgfteste

Password •••••••••

Figure 1. BDPFG20 system login screen.

Available at: http://www.dbGaP.cnptia.embrapa.br

[15] Available at http://nosql-database.org

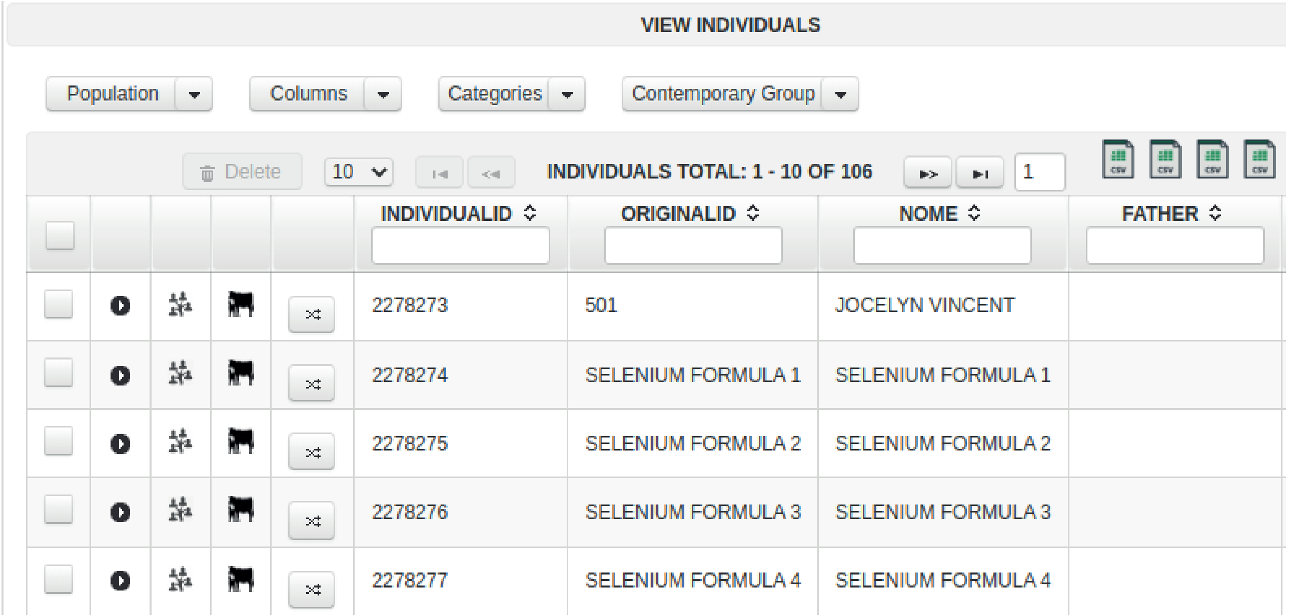
[16] Available at: https://www.postgresql.org

[17] Available at: https://gitlab.com

[18] Available at: https://www.oracle.com/br/java

[19] Available at: http://wildfly.org/downloads

One of its important features concerns the visualization of animal data (Figure 2). On this screen, the user finds various information about the individual, such as individual identifier code, original name, father, mother, date of inclusion in the population, population and other information contained in the JSON variables related to the type of individual (beef cattle, poultry, etc.). However, it is important that the variables of the phenotypes related to the species considered by the system must be previously registered, which are imported from the Embrapa Experiments System - SIEXP (Apolinário et al., 2016), where they were defined for the species the user will work with in the user group (e.g., beef cattle, poultry, etc.).



□	INDIVIDUALID	ORIGINALID	NOME	FATHER
□	2278273	501	JOCELYN VINCENT	
□	2278274	SELENIUM FORMULA 1	SELENIUM FORMULA 1	
□	2278275	SELENIUM FORMULA 2	SELENIUM FORMULA 2	
□	2278276	SELENIUM FORMULA 3	SELENIUM FORMULA 3	
□	2278277	SELENIUM FORMULA 4	SELENIUM FORMULA 4	

Figure 2. Screen showing registered individuals in the system.

Available at: http://www.dbGaP.cnptia.embrapa.br

One can also import data from files with columns separated by tabs (TSV). These files must follow a standardized format. After importing the data, the pedigree of an animal listed on the animal view page can be viewed. The pedigree window can be expanded to facilitate viewing the animals and their relatives.

The database provides several filters so that the user can check the data that has been uploaded and then export it to the format of the evaluation software. Generally, the data are exported in tabular format to be analyzed in the R program, as they are extensive tables with measurements of animal characteristics. It is also possible to export the data of these animals (phenotypes, pedigree) to files in CSV format and operate them in Excel. Existing filters allow queries by population, category, animal name, father's name, mother's name. Another tool, perhaps the most important in the system, is the one for identification of duplicated animals, allowing the user to associate duplicated animals in a single animal.

The DBGAP system is part of a computational solution proposed in other Embrapa projects (MaxiDep and MaxiPlat). The goal of these projects was to combine efforts to structure a computing solution (of which DBGAP is one of the components) to support routine genetic evaluation of beef cattle breeding programs, within the scope of the Embrapa-Geneplus program. This effort included both the development of assets to support the organization of data used in genetic evaluations (DBGAP system) and the development of a national solution for the resolution of genetic-statistical models (brBlup software).

A comparison by a search in other software systems with web interface developed by Embrapa Digital Agriculture (Vieira, 2012a, 2012b), with functionality to store genotypes and phenotypes and that includes basic queries to molecular data (SNPs), shows that a simple query in about 800 animals and 700 thousand SNP markers took at least an hour to be processed in the other software systems developed. A similar query performed in the DBGAP database takes less than a minute, as using the JSON type fields and text in the tables removes part of the necessary normalization of the traditional model, speeding up the searches.

Final considerations

The research reported in this chapter is ongoing and will continue to other stages. Regarding the research on tambaqui, with the progress of production in the near future, it will be possible to start the genetic improvement itself. The genomic tools presented in this chapter may evolve to assist in the matrix selection stage, focused on improving some characteristic of economic interest, like for instance, the slaughter weight. In the beef production chain, genomic selection is already a reality, and the results are excellent. The same can occur with the fish production chain. With the growing status of fish protein on the world menu, perhaps the Amazon region may soon become a major producer and, possibly, even an exporter of native fish. There is still a long way to go, but Embrapa has already made a significant contribution by showing and opening the way, and bioinformatics plays a fundamental role.

Validating a methodology that includes the identification of antigens through a reverse vaccinology pipeline and obtaining a multi-epitope vaccine is underway at Embrapa, with the participation of the LMB, which is aimed at controlling the bovine tick. The infestation of cattle herds by this parasite is considered today one of the most significant problems in livestock farming in economic terms, affecting all countries with tropical and subtropical climates.

In Brazil alone, annual losses due to tick infestation are in the order of US$3.24 billion (Grisi et al., 2014). Obtaining an effective vaccine will certainly contribute towards controlling the parasite, reducing the application of acaricides, as well as the environmental and economic damage resulting from this practice. Moreover, once validated, there are several possible applications of the methodology, including the identification of targets for the control of other problems of interest to agriculture involving animal health and welfare.

Machado tool will assist other ongoing projects at Embrapa. There is already a program for its use in the Genomics Applied to the Optimization of Genetic Improvement Programs for Tropical Forage Species, led by Embrapa Cerrados, with a focus on forage plant improvement. This project predicts the sequencing of reference genomes for six tropical forage species, with the characterization of broad sets of genomic variants, and Machado will probably be used as a basis for the implementation of a portal to access the generated genomic data.

DBGAP database is being structured to allow its use in other data collections, with some specific changes for each project.

As shown in the research reported here, bioinformatics has become fundamental and will be even more important in the innovation agendas towards the digital transformation of agriculture. The existence of multi-user structures is crucial to support research projects that do not have the necessary structure for complex analyses, allowing better use of the resources. With bioinformatics relying on the availability of a specialist team and adequate infrastructure, the management of the structure that supports research projects must be focused on keeping both aspects up to date.

References

AHMAD, T. A.; EWEIDA, A. E.; SHEWEITA, S. A. B-cell epitope mapping for the design of vaccines and effective diagnostics. **Trials in Vaccinology**, v. 5, p. 71-83, 2016. DOI: 10.1016/j.trivac.2016.04.003.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403-410, Oct. 1990. DOI: 10.1016/S0022-2836(05)80360-2.

ANDREOTTI, R. Performance of two Bm86 antigen vaccine formulation against tick using crossbreed bovines in stall test. **Revista Brasileira de Parasitologia Veterinária**, v.15, p. 97-100, 2006.

ANDREOTTI, R.; GIACHETTO, P. F.; CUNHA, R. C. Advances in tick vaccinology in Brazil: from gene expression to immunoprotection. **Frontiers in Biosciences**, v. 10, p. 127-42, Jan. 2018. DOI: 10.2741/s504.

ANDREWS, S. M.; POLLARD, A. J. A vaccine against serogroup B Neisseria meningitidis: dealing with uncertainty. **The Lancet Infectious Diseases**, v. 14, n. 5, p. 426-434, May 2014. DOI: 10.1016/s1473-3099(13)70341-4.

APOLINÁRIO, D. R. de F.; QUEIROS, L. R.; VACARI, I.; CRUZ, S. A. B. da. **SIExp – Sistema de Informação de Experimentos da Embrapa**. Versão v. 1.7.6. Campinas: Embrapa Informática Agropecuária, 2016.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, Aug. 2014. DOI: 10.1093/bioinformatics/btu170.

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, p. 64-71, 2009. Número especial. DOI: 10.1590/s1516-35982009001300008.

COCK, P. J. A.; ANTAO, T.; CHANG, J. T.; CHAPMAN, B. A.; COX, C. J.; DALKE, A.; FRIEDBERG, I.; HAMELRYCK, T.; KAUFF, F.; WILCZYNSKI, B.; DE HOON, M. J. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422-1423, June 2009. DOI: 10.1093/bioinformatics/btp163.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; SZCZEŚNIAK, M. W.; GAFFNEY, D. J.; ELO, L. L.; ZHANG, X.; MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v. 17, article number 13, 2016. DOI: 10.1186/s13059-016-0881-8.

EWELS, P.; MAGNUSSON, M.; LUNDIN, S.; KÄLLER, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047-3048, Oct. 2016. DOI: 10.1093/bioinformatics/btw354.

FAStQC-GitHub. Available at: https://github.com/s-andrews/FastQC/releases. Accessed on: 7 May 2020.

FAStX-Github. Available at: https://github.com/agordon/fastx_toolkit. Accessed on: 7 May 2020.

GIACHETTO, P. F.; CUNHA, R. C.; NHANI JUNIOR, A.; GARCIA, M. V.; FERRO, J. A.; ANDREOTTI, R. Gene expression in the salivary gland of Rhipicephalus (Boophilus) microplus fed on tick-susceptible and tick-resistant hosts. **Frontiers in Cellular and Infection Microbiology**, v. 9, p. 477, Jan. 2020. DOI: 10.3389/fcimb.2019.00477.

GRABHERR, M. G.; HAAS, B. J.; YASSOUR, M.; LEVIN, J. Z.; THOMPSON, D. A.; AMIT, I.; ADICONIS, X.; FAN, L.; RAYCHOWDHURY, R.; ZENG, Q.; CHEN, Z.; MAUCELI, E.; HACOHEN, N.; GNIRKE, A.; RHIND, N.; DI PALMA, F.; BIRREN, B. W.; NUSBAUM, C.; LINDBLAD-TOH, K.; FRIEDMAN, N.; REGEV, A. Full-length transcriptome assembly from RNA-seq data without a reference genome. **Nature Biotechnology**, v. 29, n. 7, p. 644-652, 2011. DOI: 10.1038/nbt.1883.

GRISI, L.; LEITE, R. C.; MARTINS, J. R. de S.; BARROS, A. T. M. de; ANDREOTTI, R.; CANÇADO, P. H. D.; LEÓN, A. A. P. de; PEREIRA, J. B.; VILLELA, H. S. Reassessment of the potential economic impact of cattle parasites in Brazil. **Revista Brasileira de Parasitologia Veterinária**, v. 23, n. 2, p. 150-156, Apr./June 2014. DOI: 10.1590/S1984-29612014042.

GUREVICH, A.; SAVELIEV, V.; VYAHHI, N.; TESLER, G. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072-1075, Apr. 2013. DOI: 10.1093/bioinformatics/btt086.

HIGA, R. H.; OLIVEIRA, G. B. **Banco de Dados de Genótipos e Fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal**. Campinas: Embrapa Informática Agropecuária, 2015. 30 p. (Embrapa Informática Agropecuária. Documentos, 133). Available at: https://ainfo.cnptia. embrapa.br/digital/bitstream/item/138127/1/Doc133.pdf. Accessed on: 7 May 2020.

HUBER, W.; CAREY, V. J.; GENTLEMAN, R.; ANDERS, S.; CARLSON, M.; CARVALHO, B. S.; BRAVO, H. C.; DAVIS, S.; GATTO L.; GIRKE, T.; GOTTARDO, R.; HAHNE, F.; HANSEN, KD.; IRIZARRY, R. A.; LAWRENCE, M.; LOVE, M. I.; MACDONALD, J.; OBENCHAIN, V.; OLE'S; A. K.; PAG'ES, H.; REYES, A.; SHANNON, P.; SMYTH, G. K; TENENBAUM, D.; WALDRON, L.; MORGAN, M. Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, v. 12, n. 2, p. 115-121, Jan. 2015. DOI: 10.1038/nmeth.3252.

IANELLA, P.; YAMAGISHI, M. E. B.; VARELA, E. S.; VILLELA, L. C. V.; PAIVA, S. R.; CAETANO, A. R. Tambaqui (Colossoma macropomum) single nucleotide polymorphism discovery by reduced representation library deep sequencing. In: AQUACULTURE, 2019, New Orleans. **Abstracts**. [S.l.: s.n.], 2019. p. 491. Available at: https://ainfo.cnptia.embrapa.br/digital/ bitstream/item/208846/1/CNPASA-2019-Aqua2.pdf. Accessed on: 7 May 2020.

KALDERIMIS, A.; LYNE, R.; BUTANO, D.; CONTRINO, S.; LYNE, M.; HEIMBACH, J.; HU, F.; SMITH, R.; ŠTĚPÁN, R.; SULLIVAN, J.; MICKLEM, G. InterMine: extensive web services for modern biology. **Nucleic Acids Research**, v. 42, n. W1, p. W468-W472, July 2014. DOI: 10.1093/nar/gku301.

KAO, D. J.; HODGES, R. S. Advantages of a synthetic peptide immunogen over a protein immunogen in the development of an anti-pilus vaccine for Pseudomonas aeruginosa. **Chemical Biology & Drug Design**, v. 74, p. 33-42, 2009. DOI: 10.1111/j.1747-0285.2009.00825.x.

LAMARRE, S.; FRASSE, P.; ZOUINE, M.; LABOURDETTE, D.; SAINDERICHIN, E.; HU, G.; LE BERRE-ANTON, V.; BOUZAYEN, M.; MAZA, E. Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. **Frontiers in Plant Science**, v. 9, article 108, Feb. 2018. DOI: 10.3389/fpls.2018.00108.

LEE, E.; HELT, G. A.; REESE, J. T.; MUNOZ-TORRES, M. C.; CHILDERS, C. P.; BUELS, R. M.; STEIN, L.; HOLMES, I.H; ELSIK, C.G.; LEWIS, S.E. Web Apollo: a web-based genomic annotation editing platform. **Genome Biology**, v. 14, n. 8, article number R93, Aug. 2013. DOI: 10.1186/gb-2013-14-8-r93.

LI, B.; DEWEY, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. **BMC Bioinformatics**, v. 12, n. 1, article number 323, Aug 2011. DOI: 10.1186/1471-2105-12-323.

MILLMAN, K. J.; AIVAZIS, M. Python for scientists and engineers. **Computing in Science & Engineering**, v. 13, n. 2, p. 9-12, Mar. 2011. DOI: 10.1109/MCSE.2011.36.

MUNGALL, C. J.; EMMERT, D. B. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. **Bioinformatics**, v. 23, n. 13, p. i337-i346, July 2007. DOI: 10.1093/bioinformatics/btm189.

NAPIER, J. A.; HASLAM, R. P.; TSALAVOUTA, M.; SAYANOVA, O. The challenges of delivering genetically modified crops with nutritional enhancement traits. **Nature Plants**, v. 5, n. 6, p. 563-567, June 2019. DOI: 10.1038/s41477-019-0430-z.

PALATNIK-DE-SOUSA, C. B.; SOARES, I. da S.; ROSA, D. S. Epitope discovery and synthetic vaccine design. **Frontiers in Immunology**, v. 9, p. 826, 2018. DOI: 10.3389/978-2-88945-522-5.

PRADO, J. R.; SEGERS, G.; VOELKER, T.; CARSON, D.; DOBERT, R.; PHILLIPS, J.; COOK, K.; CORNEJO, C.; MONKEN, J.; GRAPES, L.; REYNOLDS, T.; MARTINO-CATT, S. Genetically engineered crops: from idea to product. **Annual Reviews of Plant Biology**, v. 65, n. 1, p. 769-790, Apr 2014. DOI: 10.1146/annurev-arplant-050213-040039.

QUEVILLON, E.; SILVENTOINEN, V.; PILLAI, S.; HARTE, N.; MULDER, N.; APWEILER, R.; LOPEZ, R. InterProScan: protein domains identifier. **Nucleic Acids Research**, v. 33, p. W116-W120, July 2005. Issue suppl_2. DOI: 10.1093/nar/gki442.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2020. Available at: https://www.R-project.org. Accessed on: 7 May 2020.

RAPPUOLI, R. Reverse vaccinology. **Current Opinion in Microbiology**, v. 3, n. 5, p. 445-450, Oct. 2000. DOI: 10.1016/s1369-5274(00)00119-3.

RAPPUOLI, R.; COVACCI, A. Reverse vaccinology and genomics. **Science**, v. 302, n. 5645, p. 602, Oct. 2003. DOI: 10.1126/science.1092329.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, n. 1, p. 139-140, Jan. 2010. DOI: 10.1093/bioinformatics/btp616.

SCHWABER, K. **Agile project management with scrum**. United States: Microsoft Press, 2004. 163 p.

SEPPEY, M.; MANNI, M.; ZDOBNOV, E. M. BUSCO: assessing genome assembly and annotation completeness. In: KOLLMAR, M. (ed.). Gene prediction. New York: Humana, 2019. p. 227-245. (Methods in molecular biology, v. 1962). DOI: 10.1007/978-1-4939-9173-0_14.

SKINNER, M. E.; UZILOV, A. V.; STEIN, L. D.; MUNGALL, C. J.; HOLMES, I. H. JBrowse: a next-generation genome browser. **Genome Research**, v. 19, n. 9, p. 1630-1638, 2009. DOI: 10.1101/gr.094607.109.

SPOOR, S.; CHENG, C. H.; SANDERSON, L. A.; CONDON, B.; ALMSAEED, A.; CHEN, M.; BRETAUDEAU, A.; RASCHE, H.; JUNG, S.; MAIN, D.; BETT, K.; STATON, M.; WEGRZYN, J. L.; FELTUS, F. A.; FICKLIN, S. P. Tripal v3: an ontology-based toolkit for construction of FAIR biological community databases. **Database**, v. 2019, 2019. DOI: 10.1093/database/baz077.

STEIN, L. D.; MUNGALL, C.; SHU, S.; CAUDY, M.; MANGONE M.; DAY, A.; NICKERSON, E.; STAJICH, J. E; HARRIS, T. W.; ARVA, A.; LEWIS, S. The generic genome browser: a building block for a model organism system database. **Genome Research**, n. 516, p. 1599-1610, 2002. DOI: 10.1101/gr.403602.

THE UNIPROT CONSORTIUM. UniProt: a worldwide hub of protein knowledge. **Nucleic Acids Research**, v. 47, p. D506-D515, Jan. 2019. Issue D1. DOI: 10.1093/nar/gky1049.

TRIMGALORE-GITHUB. Available at: https://github.com/FelixKrueger/TrimGalore. Accessed on: 7 May 2020.

WIKIPEDIA. **List of RNA-Seq bioinformatics tools**. 2020. Available at: https://en.wikipedia. org/wiki/List_of_RNA-Seq_bioinformatics_tools. Accessed on: 7 May 2020.

VERNIKOS, G. S. Genome watch: overtake in reverse gear. **Nature Reviews Microbiology**, v. 6, n. 5, p. 334-335, 2008. DOI: 10.1038/nrmicro1898.

VIEIRA, F. D. **Sistema Bife de Qualidade**. Versão 1.6. Campinas: Embrapa Informática Agropecuária, 2012a. 1 CD-ROM.

VIEIRA, F. D. **Sistema Suínos**. Versão 1.1. Campinas: Embrapa Informática Agropecuária, 2012b. 1 CD-ROM.