MEMÓRIAS
DA
ACADEMIA DAS CIÊNCIAS
DE
LISBOA

CLASSE DE CIÊNCIAS

TOMO XLVII
Volume 1

Plant genomes. Scientific results, applications and debates

PERE PUIGDOMÈNECH

ACADEMIA DAS CIÊNCIAS
DE LISBOA

LISBOA • 2020

Plant genomes. Scientific results, applications and debates

Pere Puigdomènech[1]

INTRODUCTION

Genome has become a popular concept since the one from *Homo sapiens* was published in 2001. It was indeed a breakthrough result and one that would have seem impossible a decade before. This has been possible thanks to the acceleration of our methods to analyze and sequence DNA that has happened since the first sequences were obtained by the methods of Sanger (Sanger and Coulson, 1975) and Maxam and Gilbert (Maxam and Gilbert, 1977). The first plant sequences appeared in the 1980 decade, for instance the first maize storage protein sequence (Geraghty *et al.*, 1981). The use of cDNA and genomic cloning and sequencing was quickly adopted by those working in trying to understand the molecular basis of plant physiology and development and by plant geneticists. The first plant sequence published in Spain followed some years after (Prat *et al.*, 1985). We have at this moment more than 100 plant genomes in our databases and the questions is now to analyze the variability that exist in the genomes of different plant species and that means accumulating hundreds or thousands of genome sequences for a given species. This change has arrived during a lifetime for some of us.

The first plant genome to be published was that of the model species *Arabidopsis thaliana*, the first genome stretch of 1.8 Mb was published in 1998 (Bevan *et al.*, 1998) and the full genome in several articles in 2000 even preceding the publication of the sequence of the human genome that was celebrated as a scientific milestone by politicians and the media. After twenty years of these publications the present situation allows to confirm the importance of the results that started to be produced at this time. The landscape has changed for different reasons and the study of plant genomes and their use has become a routine in Plant Biology laboratories but also in Plant Breeding research both in public laboratories and in seed companies.

DNA SEQUENCING

The reasons for the explosive development of plant genomics are in the first instance methodological, in particular in DNA sequencing. The first methods developed by the end of the 70s were based on enzymatic or chemical reactions upon purified DNA fragments that produced radioactively-labeled pieces that were analyzed by polyacrylamide gel electrophoresis. The reading of the autoradiographs obtained by exposing the gels to x-ray films was done manually. Obviously that was very laborious. Automated DNA sequencers started to appear, essentially based on the Sanger method and optical

[1] Centre for Research in Agricultural Genomics. CSIC-IRTA-UAB-UB. Bellaterra. Barcelona, Spain

reading of fragments marked by fluorescent nucleotide derivatives and separated first by means of gel electrophoresis and afterwards by capillary electrophoresis that are still in use for a number of applications. Those were the systems used in the first genome projects. They needed to clone relatively short DNA fragments (around 1 kb) from larger fragments for sequencing. Assembling was relatively straightforward and was based in the knowledge of the structure of the fragments. The foreseen cost of the project carried out in the United States was 3000 million dollars but it was done with 10% less and two years before schedule. A similar thing happened with the Arabidopsis genome. This fact was due to the increased efficiency of the sequencing methods that were developed. In any case the methods used for sequencing allowed a very careful quality control and they produced a high quality genome sequences. At the same time, bioinformatics was being developed and it was used for the annotation and presentation of the final results.

Less than ten years after the publication of the human genome new developments of DNA sequencing based in the analysis of short (100-300 base pair) DNA segments appeared and they reduced the cost and speed of sequencing by orders of magnitude. The situation during these years is well presented in the figure elaborated by the American NIH where it compares the cost of DNA sequencing with the cost of microchips, known as the Moore's Law, as it is presented in Figure 1.

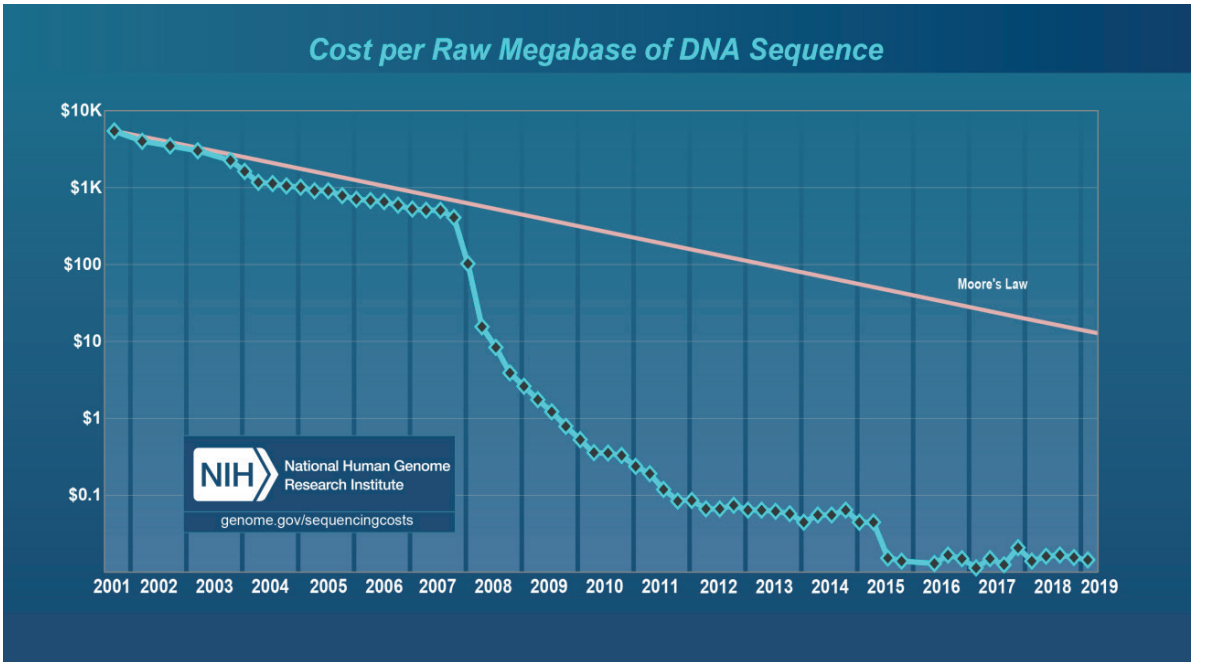

Figure 1.
The Cost of sequencing one Megabase from 2001 to the present. National Institutes of Health, USA, 2019.

SEQUENCES OF DIFFERENT PLANT GENOMES

As a consequence of these methodological developments regarding DNA sequencing the availability of plant genomes changed drastically. One of the reasons to start with Arabidopsis was that its genome was at that moment one of the smallest known among eukaryotic species, around 135 Mb and it was

reasonable to start with an easier example. The genetic map of the plant was also well known and that information helped to construct the complete sequence of the genome. The simplicity of the genome allowed also developing tools to identify the genes present in the DNA. At the beginning of the work only around 30% of the genes could be identified in relation to a possible function. The number of genes could also be deduced. One of the interesting results of the first analysis of this genome is that it contains around 27 000 genes a value that may be higher that the number of genes in the human genome that is 20 times larger. The tools needed for this analysis are essentially bioinformatic and the development of these tools has been constant during all this period facilitating the management of the large volume of data that has been generated.

After the Arabidopsis genome, the second to be published was that of rice. Two reasons explain this choice. One is the importance of rice as a crop. It is the basic food for millions of people, around the world, particularly in Asia, and it provides around one fifth of the calories needed for humans. It is also one of the simplest known among the cultivated cereals, around 430 Mb. For this reason an international consortium was formed to sequence the rice genome that was published in 2005 (International Rice Genome Sequencing Project, 2005). It was interesting that more or less at the same time two other publications present results of the rice genome, one funded by a private company and the other one by a network of Chinese scientists were published three years before but having a lower sequence quality. This fact witnessed the influence that seed companies and Chinese groups started to have upon plant genomics. After Arabidopsis and rice a number of other genomes were published. It is worth to be mentioned the genome of *Populus*, as an example of a tree and genomes of increased complexity but high scientific and economic importance such as soybean or maize. Most of these genome sequences were obtained using methods based on Sanger sequencing. The sequences of most of these reference genomes are available through databases such as Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) or Plaza (https://bioinformatics.psb.ugent.be/plaza/)

The first collection of plant genome offered an initial view of the complexity of the subject. This complexity gives rise, for instance, to an extraordinary diversity of genome sizes. We know that the genome of a plant can be contained in 80 Mb in the case of Urticularia that has a similar number of genes to Arabidopsis structured in a very compact way. On the other extreme, conifers span their genomes one thousand times this size with a similar number of genes. But the complexity may be the product of important reorganization of the genomes, essentially produced by duplications or fusions. A good example is the genome of maize, an ancient allotetraploid that gives rise to a genome of 3000 Mb, similar to the human genome and 53 000 genes. The presence of highly repetitive sequences, in particular different types of transposons, explains some of these differences in the size of plant genomes.

THE GENOMES OF CUCURBITS

The importance of studying and using the large collection of data that were being obtained with genome sequencing prompted the proposal of a plant genome initiative in Spain. The species proposed as a starting model was melon (*Cucumis melo*). The reasons for the choice were at the same time scientific (moderate genome size, interesting phenotypic and genetic variability, other interesting species in the genus) and economic (cucurbits are second only to *Solanaceae* among horticultural species in world

consumption and the seed is a valuable hybrid). Spain is the first world exporter of the species and it is a traditional fruit in many regions. For this reasons a consortium was formed between nine different public laboratories and five private companies that funded partially the project that had also the support of a state-funded foundation (Fundación Genoma) and five autonomous governments. A double-haploid line obtained from a cross of two distant melon varieties was chosen in order to minimize heterozygosity and incorporate variability in the sequence. It was interesting that the initial plan was to use Sanger sequencing and clone by clone analysis. However at this time the new methods of massive sequencing appeared and the strategy was changed towards a shotgun sequencing using one of the systems (454 analyzer) available and the project was able to reach its goals in half of the time and once the reference genome sequence was obtained, funds were left to sequence four additional varieties. That was the experience of many groups working in plant genomics in the field at the same time and the melon genome one of the first genomes of its size that was obtained using this strategy.

The melon genome was published in 2012 (Garcia-Mas *et al.* 2012). It was obtained by a combination of methods, mainly massive 454 sequencing, but also some Sanger sequencing. As a whole the quality of the genome had to be considered as very high when compared with other genomes that were published at the same time. The assembly obtained allowed to anchor the sequence with the available genetic maps and to identify 27 000 genes a number comparable to other plant genomes. The melon sequence published corresponds to a double haploid line DHL92 obtained from a crossing between two distant varieties, a cultivated Spanish melon (Piel de Sapo T111) and a Korean variety (PI 161375). The resequencing of the parental lines allowed to measure on one hand the number of single nucleotide polymorphisms that exist between these two distant lines and on the other hand to observe the recombination between the two parental chromosomes as it can be seen in Figure 2.

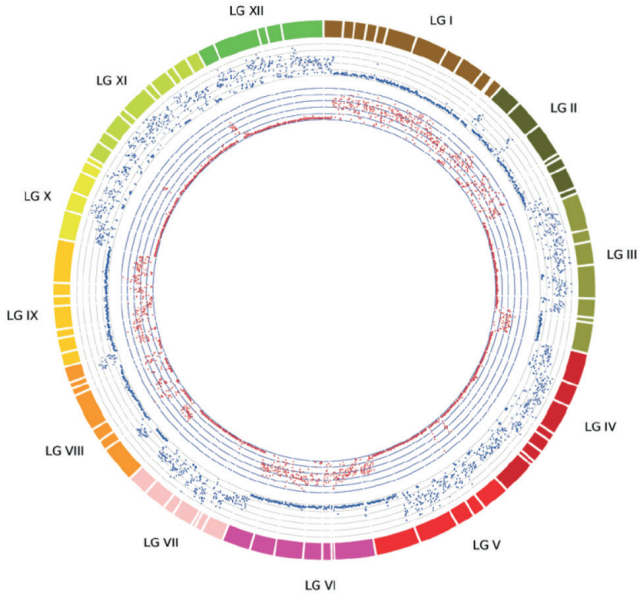

Figure 2.
Camparison of the genomic sequence of the melon double haploid line DHL92 with its two parental lines PS in blue and PI in red. The SNPs observed are plotted for each linkage group (Garcia-Mas *et al.*, 2012).

A cucumber genome was published one year before (Huang *et al.*, 2009) and that of watermelon in 2014 (Guo *et al.*, 2014). It was then possible to compare them and to conclude that no major genomic rearrangement occurred in this family of species neither in its origin nor during the period of speciation. The genome of melon was immediately used both by scientists interesting in studying the molecular basis of interesting characters. It has to be mentioned the analysis of the sex determination in plants studied in melon (Boualem, 2009; Martin, 2008) that demonstrated the importance of having the genome sequence. At the same time it was obvious that the comparison of the sequence of different melon varieties allowed discovering a large number of SNPs that are useful as genetic markers for breeders.

THE STUDY OF GENOME VARIABILITY

To obtain the sequence of a reference genome is a valuable goal in itself. It allows to identify the structure of the genome of a species, the genes it contains and to compare with other species providing information about their evolution. However, if we consider the sequencing of plant genomes as a tool that could be used for breeders, the most important information is the variability of the genome and its relation with characters of agronomic interest. For this reason once a reference genome is obtained, resequencing of varieties started immediately. Varieties to be studied can be either those that can be considered as the more distant among themselves and the results can inform about the variability existing in the whole species. This type of studies has resulted, for instance, in an increasingly rich view of how the process of domestication has occurred in the species of agronomic importance. A good example of this trend is rice. Already in 2012 an article on rice domestication published by scientists from China and Japan contained the sequences of 1083 rice genomes that allowed cartography of rice varieties and to formulate a hypothesis for its origin. (Huang *et al.*, 2012).

Another aspect that resulted from the comparison of different plant genomes and different varieties within plant species is the presence of different mechanisms acting upon these genomes. Those are, of course point mutations that create single nucleotide polymorphisms but also, the action of the different types of transposable elements, the deletion of duplication of genome regions, that may include coding sequences and even whole genome duplication. All these mechanisms have been acting upon plant genomes and they may help to explain how plants may adapt to changing conditions. One concept, for instance is the "pan-genome" hypothesis that proposes that a core genome exists in plant species that may include as few as a third of genes while the other ones may be disposable in some sense or variable among varieties (Morgante, 2007). This concept has been applied recently to explain the differences in flavor among tomato varieties (Gao, 2019). Other hypothesis may include the appearance on miRNAs in order to regulate important genes such as those providing resistance to pathogens (Gonzalez *et al.*, 2015).

THE MODIFICATION OF PLANT GENOMES

Even before plant genomes were available, the possibility to modify them through transformation technologies existed. In particular, the properties of Agrobacterium demonstrated from 1983 that it was possible to transfer DNA sequence to plant genomes allowing the modification of the set of characters in a given crop. The possibility was based in a number of microbiological and cellular methods that were

developed but also in the increasing wealth of knowledge of the molecular basis of characters interesting for agriculture. These technologies produced genetically modified crops that started to be used by farmers in 1994 and that in 2017 were grown in near 190 million hectares around the world as it is shown in Figure 3.

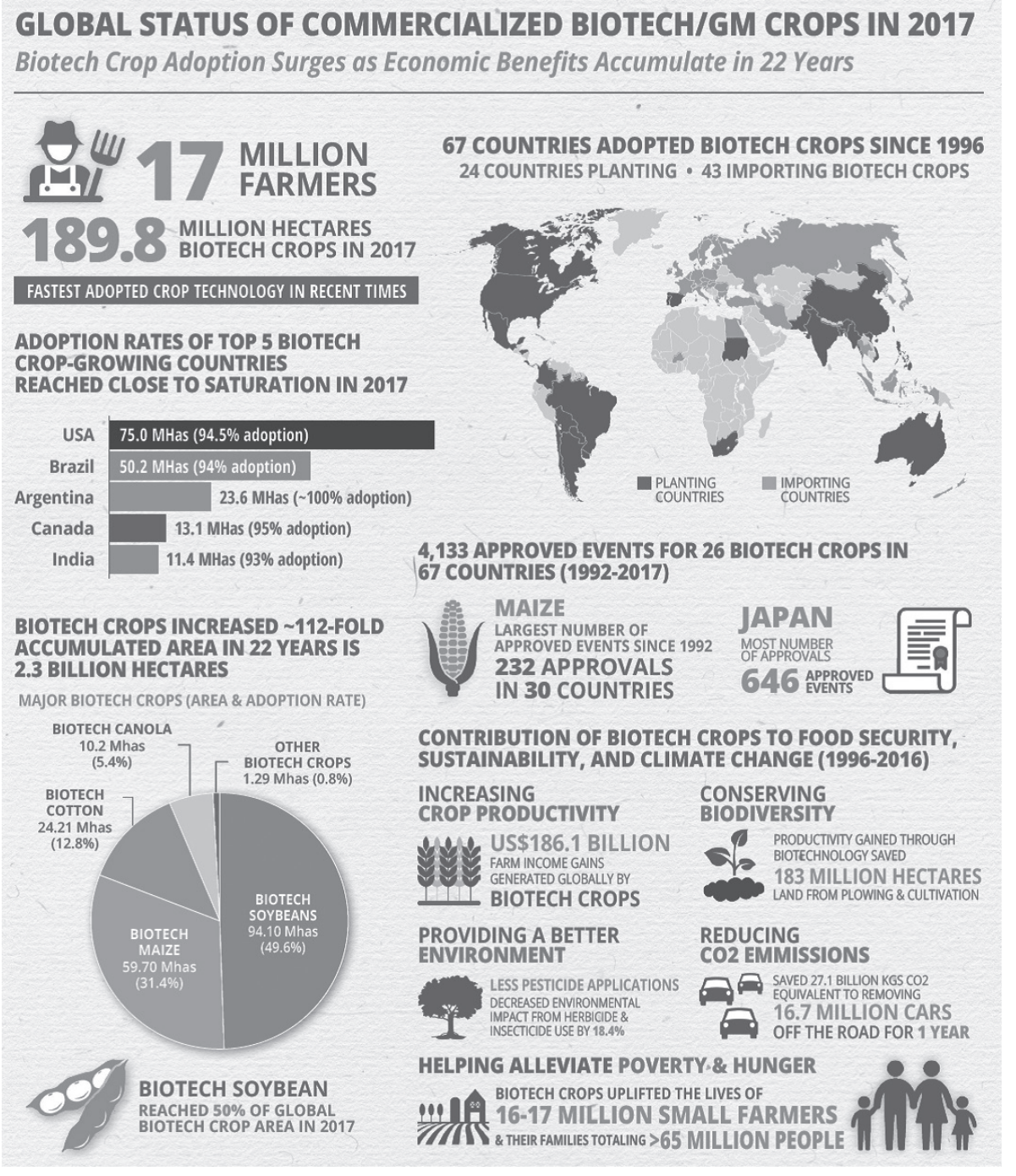

Figure 3.
The adoption of genetically modified crops in 2017, ISAAA 2019. (www.isaaa.org)

The use of GMO crops has been made within the framework of strict regulations in most countries, in particular in the United States, the European Union, Canada, Argentina, Brazil or Japan among others. These regulations have allowed that no major direct problem upon human or animal health or the environment has been produced by these new crops. However the costs of the regulatory burden, the labelling of the food products and in general the negative perception of these crops, especially in Europe have reduced the application of these methodologies.

GENOME EDITING

A new method to modify the genomes of animals and plants has been developed in the recent years using site-directed mutagenesis and in particular the CRISPR-Cas9 system. Examples in plants have been published since 2013 (Nekasov *et al.*, 2013; Shan *et al.*, 2013). The system is being widely used in plant Biology research and it has shown its value as a tool for molecular genetics. A number of new developments have been published. Among those it can be mentioned, the use of the CRISPR-Cas9 system for editing several gene sequence simultaneously (Ma *et al.*, 2015) that may lead to act upon quantitative genetic characters (Rodríguez-Leal *et al.*, 2017). Other developments have the goal to produce plants that contain only the desired modification of the genome and no other modification. It may be done through crossing the edited variety with other varieties to take out other modifications produced by the editing procedure (Chandasekaran *et al.*, 2017) or by using only the complex of the guiding RNA with the Cas9 protein avowing any DNA transformation (Woo *et al.*, 2015).

The developments of genome editing have produced a number of reactions. For one hand it has been increasingly clear the new possibilities that the new technologies open for plant breeding. On the other hand the experience with GMO use, the reactions that have been produced and the development of regulations especially in Europe prompted different types of reactions that have been summarized from the beginning different points of view (Voytas and Gao, 2014). Indeed the way how edited plants may be regulated has been a controversial question in many countries but specifically in Europe that has finally involved even the European Court.

From the academic community the situation has been object of important concern. From the beginning of agriculture itself, the plants (and animals) used for the production of food has been subjected to and important selection that has produced the species and varieties presently used in agriculture. Genetics from the beginning of the XXth century and other technologies have been decisive to face the challenges that the increase in population worldwide have presented. According to this line of thinking many scientists acknowledge that having a new tool such as genome editing and the developments that have been published recently open a number of possibilities that can be used to try to solve the future problems of food production in our present society. Solutions have been proposed to have ways for a responsible use of the new technologies and having a scientific analysis of the possible risks that they present (Casacuberta and Puigdomènech, 2018). A number of reports have been produced on this issue from different actors.

ACADEMIES

EASAC: https://easac.eu/publications/details/easac-and-the-new-plant-breeding-techniques/

The Royal Society: https://royalsociety.org/topics-policy/publications/2017/consultation-reponse--february-genomics-and-genome-editing/

Ethics Committees:

Nuffield Council of Bioethics: http://nuffieldbioethics.org/project/genome-editing/ethical-review--published-september-2016

Comité d'Éthique INRA-CIRAD-IFREMER: http://institut.inra.fr/Missions/Promouvoir-ethique-et-deontologie/Avis-du-comite-d-ethique/Questions-ethiques-et-politiques-posees-par-l-edition-du-genome-des-vegetaux

Scientific Community:

EPSO: https://epsoweb.org/epso/epso-statement-on-the-court-of-justice-of-the-eu-ruling-regarding-mutagenesis-and-the-gmo-directive/2019/02/19/

Plant Research Institutes: http://www.vib.be/en/news/Pages/Open%20Statement%20for%20the%20use%20of%20genome%20editing%20for%20sustainable%20agriculture%20and%20food%20production%20in%20the%20EU.aspx

All these reports indicate the interest of the use of new technologies for research and for their use to try to solve the problems faced by agricultural production. The presence of the Academic community is an example of their involvement when scientific results have an impact in society.

REFERENCES

BEVAN, M., BANCROFT, I., E. BENT, K. LOVE, P. PIFFANELLI, H. GOODMAN, C. DEAN, R. BERGKAMP, W. DIRKSE, M. VAN STAVEREN, W. STIEKEMA, L. DROST, P. RIDLEY, S.A. HUDSON, K. PATEL, G. MURPHY, H. WEDLER, R. WAMBUTT, T. WEITZENEGGER, T. POHL, N. TERRYN, J. GIELEN, R. VILLARROEL, R. DECLERCK, M. VAN MONTAGU, A. LECHARNY, M. KREIS, N. LAO, T. KAVANAGH, S. HEMPEL, P. KOTTER, K.D. ENTIAN, M. RIEGER, M. SCHOLFER, B. FUNK, S. MULLER AUER, M. SILVEY, R. JAMES, A. MONFORT, A. PONS, P. PUIGDOMÈNECH, A. DOUKA, E. VOUKELATOU, D. MILIONI, P. HATZOPOULOS, E. PIRAVANDI, B. OBERMAIER, H. HILBERT, A. DUESTERHOEFT, T. MOORES, J. JONES, T. ENEVA, K. PALME, V. BENES, S. RECHMAN, W. ANSORGE, R. COOKE, C. BERGER, M. DELSENY, G. VOLCKAERT, H W. MEWES, SCHUELLER, C., CHALWATZIS, N., "Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana", The EU Arabidopsis Genome Project, *Nature*, 391:485-488, 1998.

BOUALEM, A., FERGANY, M., FERNANDEZ, R. TROADEC, C., MARTIN, A. MORIN, H., SARI, M.A., COLLIN, F., FLOWERS, J.M., PITRAT, M., PURUGGANAN, M.D., DOGIMONT, C., BENDAHMANE, A., "A Conserved Mutation in an Ethylene Biosynthesis Enzyme Leads to Andromonoecy in Melons", *Science*, 321:836-8, 2008.

CASACUBERTA JM, PUIGDOMÈNECH P., "Proportionate and scientifically sound risk-assessment of gene-edited plant", *EMBO Reports*, e46907, 2018.

CHANDRASEKARAN, J., BRUMIN, M., WOLF, D., LEIBMAN, D., KLAP C., PEARLSMAN, M., SHERMAN, A., ARAZI, T., GAL--ON, A., "Development of broad virus resistance in non-transgenic cucumber using CRISPR/Cas9 technology", *Mol Plant Pathol*, 17:1140-53, 2016.

GAO, L., GONDA, I., SUN, H., MA, Q., BAO, K., TIEMAN, D.M., BURZYNSKI-CHANG, E,A., FISH, T.L., STROMBERG, K.A., SACKS, G.L., THANNHAUSER, T.W., FOOLAD, M.R., DIEZ, M.J., BLANCA, J., CANIZARES, J., XU, Y., VAN

DER KNAAP, E., HUANG, S., KLEE, H.J., GIOVANNONI, J.J., FEI, Z., "The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor", Nat Genet, 51:1044-1051, 2019, (em: https://www.ncbi.nlm.nih.gov/pubmed/31086351).

GARCIA-MAS, J., BENJAK, A., SANSEVERINO, W., BOURGEOIS, M., MIR, G., GONZÁLEZ, V.M., HÉNAFF, E., CÂMARA, F., COZZUTO, L., LOWY, E., ALIOTO, T., CAPELLA-GUTIÉRREZ, S., BLANCA, J., CAÑIZARES, J., ZIARSOLO, P., GONZALEZ-IBEAS, D., RODRÍGUEZ-MORENO, L., DROEGE, M., DU, L., ALVAREZ-TEJADO, M., LORENTE-GALDOS, B., MELÉ, M., YANG, L.,WENG, Y., NAVARRO, A., MARQUES-BONET, T., ARANDA, M.A., NUEZ, F., PICÓ, B., GABALDÓN, T., ROMA, G., GUIGÓ, R., CASACUBERTA, J.M., ARÚS, P., PUIGDOMÈNECH, P., "The genome of melon (*Cucumis melo* L.)", *Proc. Natl. Acad. Sci. USA,* 109:11872-11877, 2012.

GERAGHTY, D., PEIFER, M.A., RUBENSTEIN, I, and MESSING, J., "The primary structure of a plant storage protein: zein", *Nucleic Acids Research*, 9:5163-5174, 1981.

GONZALEZ, V.M., MUELLER, S., BAULCOMBE, D., PUIGDOMÈNECH, P., "Evolution of NBS-LRR Gene Copies among Dicot Plants and its Regulation by Members of the miR482/2118 Superfamily of miRNAs", *Mol. Plant*, 8:329-331, 2015.

GUO, S., ZHANG, J., SUN, H, …… FEI, Z., XU, Y. "The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions", *Nature Genetics*, 45:51-58, 2013.

HUANG, S., LI, R., ZHANG., …….. DU, Y., LI, S. "The genome of the cucumber, Cucumis sativus L", Nature Genetics, 41:1275-1281, 2009.

HUANG, X., KURATA, N.,WEI, X., WANG, Z., WANG, A., ZHAO, Q., ZHAO,Y, LIU, K., LU, H., LI, W., GUO, Y., LU, Y., ZHOU, C., FAN, D., WENG, Q., ZHU, C., HUANG, T., ZHANG, L., WANG, Y., FENG, L., FURUUMI, H., KUBO, T., MIYABAYASHI, T., YUAN, X., XU, Q., DONG, G., ZHAN, Q., LI, C., FUJIYAMA, A., TOYODA, A., LU, T., FENG, Q., QIAN, Q., LI, J., HAN, B., "A map of rice genome variation reveals the origin of cultivated rice", *Nature*, 490:497-501, 2012.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT. "The map-based sequence of the rice genome", *Nature*, 436:793–800, 2005.

MA, X., ZHANG, Q., ZHU. Q., LIU, W., CHEN, Y., QIU, R., WANG, B., YANG, Z., LI, H., LIN, Y., XIE, Y., SHEN, R., CHEN, S., WANG, Z., CHEN, Y., GUO, J., CHEN, L., ZHAO, X., DONG, Z., LIU, Y.G., "A Robust CRISPR/Cas9 System for Convenient, High-Efficiency Multiplex Genome Editing in Monocot and Dicot Plants" , *Mol Plant*, 8:1274-84, 2015, (em: https://www.ncbi.nlm.nih.gov/pubmed/25917172).

MARTIN, A., TROADEC, C., BOUALEM, A., RAJAB, M., FERNANDEZ, R., MORIN, H., PITRAT, M., DOGIMONT, C., BENDAHMANE, A., "A transposon-induced epigenetic change leads to sex determination in melon", *Nature*, 461:1135-8, 2009.

MAXAM, A.M., GILBERT, W., "A new method for sequencing DNA", *Proc. Natl. Acad. Sci. U.S.A.*, 74:560–564, 1977.

MORGANTE, M., DE PAOLI, E., RADOVIC, S., "Transposable elements and the plant pan-genomes", *Curr Opin Plant Biol*,10:149-55, 2007, (em: https://www.ncbi.nlm.nih.gov/pubmed/17300983).

NEKRASOV, V., STASKAWICZ, B., WEIGEL, D., JONES, J.D., KAMOUN, S., "Targeted mutagenesis in the model plant *Nicotiana benthamiana* using Cas9 RNA-guided endonuclease", *Nat. Biotechnol.*, 31:691-693, 2013.

PRAT, S., CORTADAS, J., PUIGDOMENECH, P., PALAU, J., "Nucleic acid (cDNA) and amino acid sequences of the maize endosperm glutelin-2", *Nucleic Acid Res,* 13:1493-1504, 1985.

RODRÍGUEZ-LEAL, D., LEMMON, Z.H., MAN, J., BARTLETT, M.E., LIPPMAN, Z.B., "Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing", *Cell,* 171:470-480, 2017.

SANGER, F., COULSON, A.R., "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase", *J. Mol. Biol*, 94:441–228, 1975.

SHAN, Q., WANG, Y., LI, J, ZHANG, Y, CHEN, K., LIANG, Z, ZHANG, K., LIU, J., XI, J., QIU, J.L., GAO, C., "Targeted genome modification of crop plants using a CRISPR-Cas system", *Nature Biotechnology*, 31:686–688, 2013.

VOYTAS, D.F., GAO, C., "Precision genome engineering and agriculture: opportunities and regulatory challenges", *PLoS Biol,* 10:12, 2014, (em: https://www.ncbi.nlm.nih.gov/pubmed/?term=Voytas%20DF%5BAuthor%5D&cauthor=true&cauthor_uid=24915127).

WOO, J.W., KIM, J., KWON, S.I., CORVALÁN, C., CHO, SW., KIM, H., KIM, S.G., KIM, S.T., CHOE, S., KIM, J.S., "DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins", *Nat Biotechnol,* 33:1162-4, 2015.

(Comunicação apresentada à Classe de Ciências
na sessão de 18 de abril de 2017)