Minireview

Plant genomics

Nancy Terryn[a], Pierre Rouzé[b], Marc Van Montagu[a,*]

[a] *Laboratorium voor Genetica, Departement Plantengenetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent, K.L. Ledeganckstraat 35, B-9000 Gent, Belgium*

[b] *Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Universiteit Gent, B-9000 Gent, Belgium*

Received 19 April 1999

Abstract The rapidity with which genomic sequences of the model plant *Arabidopsis thaliana* and soon of rice are becoming available has strongly boosted plant molecular biology research. Here, two main genomic fields will be discussed: the progress in different structural genome projects, such as mapping, sequencing, genome organization and comparative genomics, and the so-called functional genomics approaches to analyze the genome using such molecular tools as transcript profiling, micro-arrays, and insertional mutagenesis. In addition a section on bioinformatics is included.

Key words: Genome; Mapping; Plant; Sequencing

1. Introduction

Despite many years of research, very little is known about the composition, organization, and evolution of higher plant genomes. Plant genomes vary in size, ploidy, and chromosome number. Sizes can range from 120 Mb for *Arabidopsis thaliana* to 50 000 Mb for some lilies.

Genomics has emerged as a science of its own, being much more than the sum of genome-wide methods. This is particularly the case when considering plant genome studies. The wealth of data that is becoming available on trait-conferring genes, on the number of gene copies, and soon on hotspots for recombination, is of high importance to plant breeding.

The first genome sequencing of a higher plant, that of *Arabidopsis thaliana*, a common weed of the Brassicaceae family, has now reached the halfway stage. This progress has been greatly stimulated by other genome projects, from human to yeast. Meanwhile, tools are being developed to analyze the wealth of information that becomes available once an entire genome has been sequenced. Indeed the structure and function of genes in a genomic sequence can still only be predicted with great difficulty. Thus, there remains a huge task for functional genomics to assess gene function on a genome-wide level.

2. Structural genomics

2.1. Mapping and sequencing

The basis for genome analysis in plants is often a genomic map, which can be either a genetic map based on information from both visible and molecular markers, or a physical map, in which yeast artificial clones (YACs) and bacterial artificial clones (BACs) are aligned with the chromosomes. Such maps are available for a wide range of plants, although only a genetic map is available for a limited number of agriculturally important crops, because the genome has not yet been cloned into YAC or BAC libraries. For an overview of maps of different plant species, the reader is referred to, for example, the Agricultural Information web site at http://probe.nalusda.gov:8000/.

In addition, the sequencing of plant genomes has been initiated. Because knowledge in plant molecular biology is lagging somewhat behind the yeast and animal field, the prime interests are the transcribed regions of the genome. Therefore, sequencing projects were initiated with expressed sequence tags (ESTs), which are single sequence reads on randomly selected cDNA clones.

2.1.1. EST sequencing. For both *Arabidopsis* and rice, large EST collections of approximately 35 000 clones are already available [1–3], as well as some small EST collections of approximately 5000 clones, for instance for poplar [4] and soybean (http://129.186.26.94/soybeanest.html). In the dbEST section of GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html), the number of ESTs available for other plants can be followed. For the time being, in contrast to genomic sequencing, the growth in base pairs is rather slow, but there is a remarkable expansion in the number of new (crop) species for which an EST collection is available. It might be mentioned here that in several companies large private collections of ESTs can be consulted on request for specific academic searches.

2.1.2. Genome sequencing. *Arabidopsis* was chosen for the first plant genome sequencing project, mostly because of the small size of its genome (120 Mb), and the fact that it has become the model plant for a wide range of studies in plant sciences [5,6]. For recent data on the genomic progress as well as for general information on *Arabidopsis* the web site of the *Arabidopsis* database can be consulted at http://genome.www.stanford.edu/Arabidopsis/. The first analysis of the *Arabidopsis* data has shown that this genome is extremely gene-rich and poor in repeated elements. On average, there is one gene every 4–5 kb, the average number of introns is five, with a mean size of approximately 160 bp [7,8]. The total gene content of *Arabidopsis* is estimated to be approximately 20 000–25 000. About half of the predicted genes can be assigned to a functional category based on their similarity to known proteins or motifs. In addition, the genome sequences have taught us that even in a relatively small genome, such as *Arabidopsis*, evidence for ancient genome duplications can be found [8].

The rice genome sequencing program is also taking shape.

*Corresponding author. Fax: (32) (9) 2645349.
E-mail: mamon@gengenp.rug.ac.be

PII: S0014-5793(99)00591-8

Rice has been selected as the second model plant because it is a monocotyledonous species with a small genome size (roughly 4-fold that of *Arabidopsis*) and because of its importance as a crop. An international team is setting the primary guidelines and the first data are currently being released. A rice genome web site is available at http://www.staff.or.jp/.

More plant genomes will be sequenced in the following years, probably at low coverage, with the coming of a new generation of high throughput sequencers, some based on the principle of mass spectrometry. To be considered are a nitrogen-fixing legume, such as *Medicago sativa*, and trees, such as eucalyptus (450 Mb). Also the US National Science Foundation is funding a mapping and cloning project that sets the stage for a large genome sequencing project of corn [9]. The focus will be on the regions in which valuable information is probably present rather than on sequencing the whole genome (3 billion bases full of repetitive DNA). When several plant genomes will have been sequenced in the coming years insights into species evolution will be gained from comparing genes and genome organization.

2.2. Genome organization

Recent studies indicate common aspects in the characteristics of many plant genomes, including in the structure of chromosomal components, such as telomeres and centromeres [10]. A study by Paul and Ferl [11] in both *Arabidopsis* and maize has indicated that the genome is grouped in specific domains that probably represent the structural loop domains created by the attachment of the chromatin to the nuclear matrix at loop basements.

However, plant genomes differ in their gene organization. Barakat et al. [12] analyzed the genome of *Arabidopsis* and have shown that its organization is different from that of the genomes of Gramineae. Genes in *Arabidopsis* are fairly evenly distributed, whereas studies of maize, rice, and barley have shown that the vast majority of genes are clustered in long DNA stretches (so-called gene space) that are separated by expanses of gene-empty DNA [13]. More than 50% of the maize genome is composed of interspersed repetitive DNAs, primarily retrotransposons that insert between genes [14].

Mobile DNAs in plants can be grouped in two classes: those that transpose as DNA molecules and those that transpose as an RNA intermediate [15]. The better-known elements such as *Ac* and *En*/*Spm* fall into the first class. Their copy number is usually rather low, with the exception of miniature inverted-repeat transposable elements (MITEs) that can be found at a few thousand copies per genome. The elements of the second class are usually referred to as retroelements. In large plant genomes, such as maize and barley, these elements, which contain long terminal repeats (LTRs), range from a few kb to 15 kb in size and make up the majority of the genome [16]. These retroelements are less abundant in plants with a smaller genome.

2.3. Comparative genomics

Comparative genomics is usually based on genomic maps and sequences and studies primarily the more complex genomes, such as those of grasses. Comparative genomics offers possibilities to link plant families through their genomes. These studies will provide keys to understanding how genes and genomes are structured and how they have evolved. Clear mapping data as well as some preliminary sequence data show the extent of synteny, i.e. conservation of gene order, between genomes of plants from the same family or from related ones [17]. This has been especially documented for monocotyledonous species in the grass family [18]. Through identification and mapping of synteny, it will be possible to isolate genes from crop plants with large genomes, using information on the homologous genes in a related plant with a smaller genome, such as *Arabidopsis*. Because of the importance of some of the domesticated grasses, such as rice, wheat, and maize, to the human food chain, developments in this area can lead directly to opportunities to improve the productivity of our food systems.

3. Functional genomics

3.1. Monitoring gene expression

Information on where and when a certain gene is expressed can provide indications of the biological function of the encoded protein. Classical approaches such as Northern blot analysis allow such a study on a single gene, but in view of the huge amount of data from the genome programs, from plants as well as from other organisms, strategies have been developed to look at expression on a genome-based level [19].

3.1.1. DNA micro-arrays. In principle DNA micro-arrays are a kind of 'reverse Northern blot' whereby DNA clones (cDNA clones, PCR-generated fragments, etc.) are spotted in a dense array and hybridized to RNA-derived probes. The hybridization signal can be quantified automatically and reflects the abundance of the corresponding mRNA in the total RNA pool. The value of this technique is in the miniaturization, whereby large numbers of clones can be analyzed simultaneously. The use of this technique in plants has recently been reviewed by Kehoe et al. [20]. Three techniques are currently in use: (1) DNA spotted on nitrocellulose filters and hybridized to radioactive probes [21], (2) DNA spotted on coated glass slides and hybridized to fluorescently labeled probes [22,23], and (3) DNA oligonucleotides synthesized in situ on a solid support and hybridized to fluorescent probes. This last technique is mostly referred to as DNA chips and has been developed by Affimetrix (Santa Clara, CA, USA). For an example from yeast, see Wodicka et al. [24].

To be consulted and queried by the scientific community, hybridization data of the micro-arrays should be stored in public expression databases and linked to sequence and mutant database entries with gene identity and structure. Particularly useful information, such as insights into gene regulation and interaction networks, will come from the clustering of the genes with the same expression profile and from the expression study of mutants.

3.1.2. Differential display. This method makes it possible to analyze and compare transcribed genes systematically in a bi-directional fashion in a one-tube reaction. It involves the isolation of RNA from the tissues to be compared, followed by PCR amplification of this RNA using random primers [25]. Samples are then analyzed by gel electrophoreses and the patterns of the amplified cDNAs compared. cDNAs that are found to be differentially amplified can then be eluted from the gel and cloned. The most significant advantages of differential display are its simplicity, the small amount of RNA needed as start material, and the possibility of detecting virtually all differentially expressed mRNAs if a large number of primer combinations are performed. Disadvantages, however,

are the problems with reproducibility and the reliability of the observed differences, because Northern blot data of the identified genes do not always confirm the data obtained by differential display.

3.1.3. Transcript profiling. Transcript profiling is also referred to as cDNA-AFLP. It is a technique based on AFLP [26], modified for use on mRNA material [27]. The principle is comparable to that of differential display, but the cDNA material is first cut with restriction enzymes, after which an adapter is annealed to these sites. Primers based on this adapter sequence can then amplify the cDNA, making use of extra selective nucleotides. When transcript profiling is done with all the possible primer combinations it should, in theory, yield information on about 80% of the transcripts, depending on the enzymes that are used in the protocol. In addition, if the differentially expressed AFLP bands are sequenced, one would be able to link expression data with the available genome information.

3.2. Gene knock-out and mutagenesis

The ultimate tool for studying gene function is gene knock-out via homologous recombination. Unfortunately in plants very little success has been realized with target-specific constructs for gene replacement [28]. Only for a moss, *Physcomitrella patens*, good results have been obtained [29].

To overcome the lack of a good knock-out system, a random mutational approach has been widely used and seems to be very promising in the field of plant functional genomics. Two classes of mutants can be distinguished: firstly, mutants generated by mutagens, such as ethyl methanesulfonate and X-rays and, secondly, insertional mutant collections whereby a piece of foreign DNA is randomly inserted into the genome.

3.2.1. Classical mutagenesis and map-based cloning. This approach has long been the search for a needle in the haystack. Indeed, once a certain mutant with an interesting phenotype was obtained by random mutagenesis it was difficult and at least time-consuming to map the position of the gene and 'walk' or 'land' on it [30]. However, recent techniques, such as AFLP [26], bulk segregant analysis [31], and the available maps and sequencing data allow a fast mapping of mutants [32].

3.2.2. Insertional mutagenesis. As the identification of classical mutants has long been a time-consuming effort, an approach has been followed to introduce foreign pieces of DNA into the plant genome to interrupt genes at random. The T-DNA of *Agrobacterium* as well as transposons have been used [33,34]. Because the sequence of the inserted element is known, these libraries can be screened by PCR-based strategies built on sequence information of the gene in which one wants to find a mutant. For instance, theoretically for every gene of *Arabidopsis* that has been identified within the sequencing program, an insertional mutant should be available in one of the world-wide collections.

4. Bioinformatics

The common feature in any genome-wide approach is the production of enormous amounts of raw data. Bioinformatics was soon seen as a way to help produce and deal with these data in specific databases. Nevertheless, the need to transform these data into information that biologists can use and query properly has probably been underestimated. As an effort toward building such knowledge, computer tools have been developed to decipher the gene architecture of the *Arabidopsis* genome and provide annotation of the genomic sequences. However, distinguishing coding from non-coding sequences and setting the proper gene structure remains a problem. The necessary EST/cDNA information is far from complete and the performance of computer prediction tools that are used instead has not been evaluated yet. More importantly, a proper computer-assisted environment for whole-genome annotation is still missing [35]. This lack largely explains why the present-day annotation is so faulty, limiting its use for data mining and experiment setting. The progress in functional and expression studies will increase the challenge to handle all data. It will certainly help solving the functional annotation problem, if the proper effort is made to settle an adequate ontology and develop the tools to analyze the data and turn them into knowledge on gene biological function that is properly linked to gene structure.

5. Conclusion

More exciting than the wealth of sequence and functional data that is being produced nowadays in plant genome research are the questions that remain unanswered and that will give us food for thought and experiments for the coming years. How are the plant genomes organized? What is the relation between different species? What is the rate of genomic evolution? What is the relation between genome organization and gene expression? The new molecular tools that are becoming available should allow us to, at least to some extent, answer these questions in the future.

References

[1] Newman, T., de Bruijn, F.J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Somerville, S., Thomashow, M., Retzel, E. and Somerville, C. (1994) Plant Physiol. 106, 1241–1255.
[2] Cooke, R., Raynal, M., Laudié, M., Grellet, F., Delseny, M., Morris, P.-C., Guerrier, D., Giraudat, J., Quigley, F., Clabault, G., Li, Y.-F., Mache, R., Krivitzky, M., Gy, I.J.-J., Kreis, M., Lecharny, A., Parmentier, Y., Marbach, J., Fleck, J., Clément, B., Philipps, G., Hervé, C., Bardet, C., Tremousaygue, D., Lescure, B., Lacomme, C., Roby, D., Jourjon, M.-F., Chabrier, P., Charpenteau, J.-L., Desprez, T., Amselem, J., Chiapello, H. and Höfte, H. (1996) Plant J. 9, 101–124.
[3] Yamamoto, K. and Sasaki, T. (1997) Plant Mol. Biol. 35, 135–144.
[4] Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T.T., Boerjan, W., Gustafsson, P., Uhlén, M., Sundberg, B. and Lundeberg, J. (1998) Proc. Natl. Acad. Sci. USA 95, 13330–13335.
[5] Meyerowitz, E.M. and Somerville, C.R. (1994) *Arabidopsis*, Cold Spring Harbor Monograph Series, Vol. 27, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[6] Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koornneef, M. (1998) Science 282, 662–682.
[7] Bevan, M., Bancroft, I., Bent, E., Love, K., Piffanelli, P., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema, W., Drost, L., Ridley, P., Hudson, S.-A., Patel, K., Murphy, G., Wedler, H., Wedler, E., Wanbutt, R., Weitzenegger, T., Pohl, T., Terryn, N., Gielen, J., Villarroel, R., De Clercq, R., Van Montagu, M., Lecharny, A., Kreis, M., Lao, N., Kavanagh, T., Hempel, S., Kotter, P., Entian, K.-D., Rieger, M., Scholfer, M., Funk, N., Muller-Auer, S., Silvey, M., James, R., Montfort, A., Pons, A., Puigdomenech, P., Douka, A., Voukelatou, E., Milioni, D., Hatzopoulos, P., Piravandi, E., Obermaier, B., Hil-

bert, H., Duesterhoeft, A., Moores, T., Jones, J., Eneva, T., Palme, K., Benes, V., Rechman, S., Ansorge, W., Cooke, R., Berger, C., Delseny, M., Volckaert, G., Mewes, H.-W., Schueller, C. and Chalwatzis, N. (1998) Nature 391, 485–488.
[8] Terryn, N., Heijnen, L., De Keyser, A., Van Asseldonck, M., De Clercq, R., Verbakel, H., Gielen, J., Zabeau, M., Villarroel, R., Jesse, T., Neyt, P., Hogers, R., Van den Daele, H., Ardiles, W., Schueller, C., Mayer, K., Déhais, P., Rombauts, S., Van Montagu, M., Rouzé, P. and Vos, P. (1999) FEBS Lett. 445, 237–245.
[9] Pennisi, E. (1998) Science 282, 652–654.
[10] Bennetzen, J.L. (1998) Curr. Opin. Plant Biol. 1, 103–108.
[11] Paul, A.-L. and Ferl, R.J. (1998) Plant Cell 10, 1349–1359.
[12] Barakat, A., Matassi, G. and Bernardi, G. (1998) Proc. Natl. Acad. Sci. USA 95, 10044–10049.
[13] Barakat, A., Carels, N. and Bernardi, G. (1997) Proc. Natl. Acad. Sci. USA 94, 6857–6861.
[14] Bennetzen, J.L., SanMiguel, P., Chen, M., Tikhonov, A., Francki, M. and Avramova, Z. (1998) Proc. Natl. Acad. Sci. USA 95, 1975–1978.
[15] Flavell, A.J., Pearce, S.R. and Kumar, A. (1994) Curr. Opin. Genet. Dev. 4, 838–844.
[16] Bennetzen, J.L. (1996) Trends Microbiol. 4, 347–353.
[17] Gale, M.D. and Devos, K.M. (1998) Science 282, 656–659.
[18] Bennetzen, J.L. and Freeling, M. (1997) Genome Res. 7, 301–306.
[19] Kozian, D.H. and Kirschbaum, B.J. (1999) Trends Biotechnol. 17, 73–78.
[20] Kehoe, D.M., Villand, P. and Somerville, S. (1999) Trends Plant Sci. 4, 38–41.
[21] Desprez, T., Amselem, J., Caboche, M. and Höfte, H. (1998) Plant J. 14, 643–652.
[22] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Science 270, 467–470.
[23] Ruan, Y., Gilmore, J. and Conner, T. (1998) Plant J. 15, 821–833.
[24] Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H. and Lockhart, D.J. (1997) Nature Biotechnol. 15, 1359–1367.
[25] Liang, P. and Pardee, A.B. (1992) Science 257, 967–971.
[26] Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. and Zabeau, M. (1995) Nucleic Acids Res. 23, 4407–4414.
[27] Bachem, C.W.B., van der Hoeven, R.S., de Bruijn, S.M., Vreugdenhil, D., Zabeau, M. and Visser, R.G.F. (1996) Plant J. 9, 745–753.
[28] Vergunst, A.C. and Hooykaas, P.J.J. (1999) Crit. Rev. Plant Sci. 18, 1–31.
[29] Schaefer, D.G. and Zrijd, J.-P. (1997) Plant J. 11, 1195–1206.
[30] Tanksley, S.D., Ganal, M.W. and Martin, G.B. (1995) Trends Genet. 11, 63–68.
[31] Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Proc. Natl. Acad. Sci. USA 88, 9828–9832.
[32] Vos, P., Simons, G., Jesse, T., Wijbrandi, J., Heinen, L., Hogers, R., Frijters, A., Groenendijk, J., Diergaarde, P., Reijans, M., Fierens-Onstenk, J., de Both, M., Peleman, J., Liharska, T., Hontelez, J. and Zabeau, M. (1998) Nature Biotechnol. 16, 1365–1369.
[33] Martienssen, R.A. (1998) Proc. Natl. Acad. Sci. USA 95, 2021–2026.
[34] Azpiroz-Leehan, R. and Feldmann, K.A. (1997) Trends Genet. 13, 152–156.
[35] Rouzé, P., Pavy, N. and Rombauts, S. (1999) Curr. Opin. Plant Biol. 2, 90–95.