Molecular Biology, Vol. 35, No. 3, 2001, pp. 285–293. Translated from Molekulyarnaya Biologiya, Vol. 35, No. 3, 2001, pp. 339–348.

REVIEWS

UDC 575.113

Introduction into Plant Genomics

A. V. Zelenin, E. D. Badaeva, and O. V. Muravenko

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia;
E-mail: azel@genome.eimb.relarn.ru

Received October 18, 2000

Abstract—The success in complete sequencing of "small" genomes and development of new technologies that markedly speed up the cloning and sequencing processes open the way to intense development of plant genomics and complete sequencing of DNA of some species. It is assumed that success in plant genomics will result in revolutionary changes in biotechnology and plant breeding. However, the enormous size of genomes (tens of billions of base pairs), their extraordinary abundance of repetitive sequences, and allopolyploidy (the presence in a nucleus of several related but not identical genomes) force us to think that only few "basic" plant species will undergo complete sequencing, whereas genome investigations in other species will follow the principles of comparative genomics. By the present time, sequencing of the Arabidopsis genome (125 Mbp) is completed and that of the rice genome (about 430 Mbp) is close to its end. Studying the genomes of other plants, including economically valuable ones, already began on the basis of these works. The peculiarities of plant genomes make extraordinarily important our detailed knowledge on plant chromosomes which, in its turn, calls for expansion of research in this direction and development of new chromosome technologies, including the DNA-sparing methods of high-resolution banding.

Key words: plant genome, genomics, chromosome analysis, chromosome banding

INTRODUCTION

The impressive achievements of the Human Genome program, which have brought near completion its most important stage—determination of the nucleotide sequence of the genome euchromatic part, as well as the success in deciphering "small" (viruses, bacteria, yeasts) and "medium" (*Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster*) genomes [1–4] have put a problem of starting broad-scale studies of plant genomes and establishment of genomics of economically important plants (crop genomics). The problem was formulated at the Plant Genome Conference in 1997 in the USA. The materials of this Conference were published in 1998 in Proceedings of the National Academy of Sciences of USA [5] (see also [6–9]).

It is expected that the main attention will be focused on the following problems: discovering new genes, which is important for the development of plant biotechnology; identification, cloning, and sequencing of genes responsible for variability and resistance to unfavorable environmental factors and regulating chromosome pairing in polyploid plants. This will open new possibilities for improving the breeding process. Up to now, the number of cloned, sequenced, and mapped plant genes (except Arabidopsis and rice) is extremely low, but it has been announced [7, 8] that sequencing of "large" or "super-large" plant genomes, such as maize, barley, and wheat, is either planned or already under way.

GENOMES AND GENOMICS

The term "genome" was introduced by Winkler at the beginning of the XX century to designate a haploid set of chromosomes with their genes [10]. This definition is still in common use [11], but with the development of molecular genetics the sense of this term has changed significantly. At present, it is applied both in the above-mentioned narrow and in the modern sense to designate the total genetic material of an individual organism (unicellular, multicellular, or virus) that is not an allopolyploid, i.e., does not contain several related but not identical genomes [12].

The term "genomics" is not related to any definite scientist. It means the science concerning genomes. It includes studying genomes at molecular, chromosomal, biochemical, and phenotypic levels [12, 13].

Already at the start of chromosome investigations, there began comparative studies of genomes of related plants, including analysis of meiotic conjugation of chromosomes in interspecies hybrids. The improvement of existing and development of new technologies, integration of knowledge obtained in the related fields of science, such as molecular biology, genetics, and cytology, have significantly expanded the sphere of genomics.

0026-8933/01/3503-0285$25.00

To characterize genomes of individual species, researchers began to use chromosome technologies (karyotype analysis, chromosome banding, *in situ* hybridization), biochemical methods (electrophoretic analysis of proteins and immunochemical methods), as well as approaches based on DNA analysis (determination of DNA content, restriction profiles, collinearity of molecular markers, and finally, determination of complete sequence, i.e., structural genomics) [14].

Genomics is subdivided into functional genomics closely linked with the new science of proteomics (investigation of the protein population in a cell), comparative genomics, evolutionary genomics, ethnogenomics, paleogenomics, etc. Naturally, ethnogenomics, like pharmacogenomics or cardiogenomics that are more often distinguished as special sections, belong only to human genomics. However, all these directions are based on structural genomics, i.e., knowledge of the primary structure (nucleotide sequence) of the whole genome or its separate parts.

PRIMARY STRUCTURE OF DNA AS BASIS FOR GENOMICS

Knowledge of the nucleotide sequence of DNA is important and in many cases essential for development of structural genomics, including plant genomics. However, for some reasons, first of all those connected with genome peculiarities (see below for details), structural genomics of plants is developed much slower than studies of other genomes (viruses, microorganisms, man). Owing to this, we should shortly analyze the situation in structural genomics at the time of preparation of this article. This reservation is necessary, because the impetuous development of genomics may at any time introduce extremely important changes into its separate sections.

The possibility and scientific expediency of sequencing nucleic acid fragments of different length was formulated in the middle of the last century and it was first actualized in the second half of the sixties as determination of the primary structure of three transport RNA, including the one for valine [15].

However, determination of the complete structure of genomic nucleic acids became possible only after development of efficient techniques for DNA sequencing. The strategy of genome sequencing is based on two principles that may be arbitrarily called classical or gradual, and total. The first approach includes preliminary splitting of the genome into parts, obtaining genetic and molecular markers, and stepwise study of different chromosome regions (clone-to-clone technology), their cloning, sequencing of separate clones, compilation of local contigs, and finally, establishing the full nucleotide sequence of chromosome regions and entire chromosomes.

From the very beginning of works on the Human Genome program, the investigations followed a classical way. The natural genome division into 24 parts, i.e., "chromosome by chromosome," was used. Pure fractions of individual chromosomes were obtained by chromosome sorting using flow fluorimetry, or from the human–rodent hybrid cells. This material was used to create chromosome-specific DNA libraries.

The largest human chromosome (first) consists of 300 Mbp, the smallest one (Y) of 50 Mbp. This is much more than in viral genomes that were sequenced completely by that time. For this reason, human chromosomes were split into parts using radiation hybrids and/or microdissection, and the arm-, locus-, and band-specific DNA libraries were obtained.

In this way, the first success was achieved in the large-scale sequencing of human genome (chromosomes 21 and 22) [3, 8, 16]. Thus, DNA of chromosome 22 contains 54 Mbp, 38 of which pertain to the sequenced euchromatic part. Powerful international consortiums used the same classical approach for total sequencing of the first eukaryotic genome of yeast *Saccharomyces cerevisiae* (13.5 Mbp) [17] and genomes of first multicellular organisms: a roundworm *Caenorhabditis elegans* (97 Mbp) [18] and a dicotyledonous plant *Arabidopsis thaliana* (125 Mbp) [19–25].

The second approach is based on simultaneous splitting of the genome into parts (shotgun technology), cloning of the obtained fragments in a virus vector, obtaining a significant amount of individual clones, their partial sequencing, design of a contig, and a final step, determination of the total sequence of genomic DNA.

In such a schematic presentation, the total cloning technology seems rather simple. However, in practice it runs into a number of quite serious difficulties because of the necessity of obtaining enormous amounts of clones (it is assumed that the genome under study or its region should be overlapped by clones no less than ten times), a huge volume of sequencing, and extraordinarily complicated work on clone analysis and design of contigs. That is why structural studies of most genomes used the classical way.

An exception was the works on sequencing "small" genomes. The total technology enabled determination of primary structures of nucleic acids of phages and viruses, and lately many microorganisms were added (genome size from thousands to several million nucleotides) [17]. The success of these studies is based on the fact that the major part of genomes of these organisms consists of the genes proper and regulatory sites: they are virtually free of the so-called nonsense DNA represented by repetitive sequences of

different length and complexity. It is these "repeats" that are the main and in many cases a practically insurmountable difficulty in forming extended contigs.

During the past years, qualitative changes happened both in the sequencing technique proper and in the material preparation (cloning). The process has been practically put on an industrial scale. It is claimed that at least in some specialized centers the sequencing rate reaches hundreds of million base pairs a day. This allowed a well known American molecular biologist Crag Ventor to announce about two years ago that the Center of Genomic Investigations organized by him—Celera Genomics—will very soon perform full sequencing of human genome using the improved strategy of total cloning [26]. In the middle of 2000, mass media [27] have published Ventor's statement that this titanic work was over, and the full primary structure of human genome will be available to the world community by the end of 2000, but there are still no publications on this subject in scientific journals. However, only after detailed publications it will become clear whether it is possible to apply the total sequencing strategy to deciphering extensive genomes, in particular, such "large" and "superlarge" ones as most plant genomes.

By the present time, the structure of one relatively extensive "medium" genome of fruit fly *Drosophila melanogaster* deciphered with participation of Celera Genomics was published. The deciphering of this genome was based on the results of its total sequencing [28, 29]. The sequenced part is 120 Mbp, containing about 13,500 genes. The data of Celera Genomics were compared with the results of studies of Drosophila genome obtained by the world scientific community during nine decades of the XX century [30–32]. It appears that the majority of genes included in the described contig were revealed earlier. However, it covers only 2/3 of the total length of the genome, its euchromatic part. The remaining (heterochromatic) third consists of different repetitive sequences that are not amenable to such sequencing.

Deciphering of the Drosophila genome has shown the efficiency of combining the two approaches, "classical" and "total" [30]. It can be assumed that such a combination of technologies will be an important feature of the long-term strategy of structural genomics for both intermediate and large genomes.

Concluding this section, it seems reasonable to discuss the term "complete genome sequencing." During many years it implied determination of the linear sequence of all nucleotides in the genome. In particular, just this was kept in mind when the Human Genome program was created.

Such a concept accounts for the situation concerning "supersmall" genomes of viruses and bacteria. It also corresponds to reality as applied to the genome of *C. elegans* (96 Mbp) in which the linear arrangement of 97% of all nucleotides was determined. Approximately the same can be said about Arabidopsis (125 Mbp), in which the structure of 94% of genome is known. It was proposed to call such a level of genome sequencing "essentially completed."

Different figures characterize the level of the Drosophila genome sequencing (180 Mbp). The designed contig with about 13,500 genes covers only 120 Mbp, i.e., its euchromatic part. The rest 33% of the genome belong to extended heterochromatic regions whose sequencing at the present-day level of knowledge is practically impossible and hardly reasonable [29]. It is generally agreed to call such a level of genome sequencing "substantially completed."

In the *Caenorhabditis–Arabidopsis–Drosophila* triad, there is a linear correlation between the size of the whole genome and that of its sequenced part. However, a different situation is observed as soon as we begin to analyze data obtained for the human genome, or more exactly, for the sequenced chromosomes. Thus, chromosome 22 consists of 56 Mbp, 33.5 Mbp of which or about 60% is sequenced. Thus, the ratio of the sequenced (euchromatic) and nonsequenced (heterochromatic) parts of this chromosome is quite close to that for Drosophila, although its genome is 18 times smaller. Further works on sequencing the rice and human genomes will allow one to find out whether there is a correlation between the size of the whole genome and its sequenable part. It should be stressed that at the present level of knowledge the term "completeness" of sequencing small and intermediate genomes suggests only determination of the primary structure of their euchromatic part.

However, it is already clear that for all genomes, beginning with the small ones, only the sequence of their euchromatic part, i.e., the linear arrangement of genes in individual chromosomes and in the genome as a whole, rather than the complete sequence, is and probably will be the point at least in the visible future.

THE SITUATION IN PLANT GENOMICS

The solution of problems of structural genomics of plants is connected with significant difficulties caused by the peculiarities of plant genomes.

1. Huge genomes reaching tens of billions of base pairs. Here we shall give approximate values characterizing the haploid genomes in some organisms [32–38]:

Escherichia coli, 4.5 Mbp;

yeast (*S. cerevisiae*), 13.5 Mbp;

a roundworm (*C. elegans*), 97 Mbp;

a small dicot flowering plant (*A. thaliana*), 125 Mbp;

a fruit fly (*D. melanogaster*), 180 Mbp;

rice (*Oryza sativa*), 430 Mbp;

man (*Homo sapiens*), 3.2 Gbp;

rye (*Secale cereale*), cultivated barley (*Hordeum vulgare*), and diploid wheat (*Triticum monococcum*), 6–7 Gbp;

durum (tetraploid) wheat (*Triticum durum*), 12–13 Gbp;

soft (hexaploid) wheat (*Triticum aestivum*), 16 Gbp;

Liliaceae (*Lilium*), 50–60 Gbp.

2. Numerous polyploid forms.

3. Abundance (up to 99%) of so-called nonsense (noncoding, i.e., containing no structural genes) DNA, which seriously hinders sequencing, gene mapping, and design of gene contigs [33–37, 39].

4. Poor (compared with the human and Drosophila genomes) morphological, genetic, and physical mapping of chromosomes.

5. A large number of "small-chromosomal" plants in which the chromosome length does not exceed 3 μm.

6. The difficulty of chromosomal mapping of individual genes using *in situ* hybridization [40].

7. A huge, not yet exactly determined number of plant species, for most of which the number of chromosomes and DNA content in a haploid genome is still unknown [36].

In spite of the outlined difficulties, the molecular-biological investigation of plant genomes has been going on for a long time. These works were and are carried out now mainly under the strategy based on chromosome mapping using genetic and physical markers [41]. Physical markers are, in particular, obtained by microdissection (see below for details). Selective sequencing was successfully used to characterize individual sites of chromosomes. In parallel, methods of comparative, evolutionary, and functional genomics are widely used [37, 42–48].

Meanwhile, the works on total sequencing of plant genomes have begun quite some time ago, considering two plants with small genomes: Arabidopsis (96 Mbp), a member of dicots; and rice (430 Mbp) from monocots [9, 37, 39, 42, 49].

By the present time, the work on complete sequencing of the *A. thaliana* genome has been finished and published [19–25]. This plant (small mustard) from the Brassicaceae (Cruciferae) family grows in diverse climatic and ecological conditions. Arabidopsis is a typical plant that is close to other members of the plant world, but its genome is characterized by a small amount of excessive DNA (about 10%), which made much easier its sequencing. By the end of 1999 the structure of chromosomes 2 and 4 was determined [19–21], which made 31% of the whole *A. thaliana* genome, and by December 2000 the remaining three chromosomes 1, 3, and 5 were sequenced [16, 22–25, 50].

The Arabidopsis genome consists of 125 Mbp, whereas its sequenced euchromatic part covers 115 Mbp, i.e., 92% of the whole genome. In addition to the euchromatic part, centromeric regions of chromosomes (about 3 Mbp) were also sequenced for the first time. Three heterochromatic regions are still not sequenced (three gaps). For comparison, in the Drosophila genome the number of such "gaps" reaches 1200 [50].

The Arabidopsis genes are compact, contain several exons (with an average size of 250 bp) separated by short introns. The genes are tightly linked to each other (intergene spacers make about 4.6 kb). Sequence analysis revealed in the Arabidopsis genome about 15,000 individual genes [22, 38], and the functions of some 1000 are now known [51]. Up to 70% of genes are duplicated, i.e., the total number of genes in the Arabidopsis genome is 25,000 [22, 50]. It was suggested [52] that this mass gene duplication is the result of at least four large-scale events that happened 100–200 million years ago.

Significant success was achieved in sequencing the rice genome containing 430 Mbp divided between 12 chromosomes. These works have been carried out for many years by Japanese scientists and are based mainly on the "chromosome-by-chromosome" genome splitting [9, 37, 39, 42, 49] using laser microdissection and obtaining the chromosomal and smaller-size libraries. In parallel, works based on total cloning are also under way [38, 53].

There were numerous reports in mass media about sequencing of the rice genome claiming that if it is not over then very close to completion, but these claims have not been confirmed by scientific publications. One such report appeared on January 26, 2001 on the Associated Press site in the Internet (www.sunone.org). It dealt with the work carried out by the American company Syntenta and Myriad Genetics, Inc. It reported cloning of the rice genome in bacterial artificial chromosomes (BAC), with several thousand BAC clones used to design several hundreds of contigs. The situation with sequencing of the rice genome is described in more detail in [38].

There is no more concrete information about sequencing results. However, the involvement in the work on deciphering the rice genome of an additional research group will undoubtedly contribute to the successful finish of an international project scheduled by the end of 2003. Already now the information about the partial sequence of the rice genome obtained under this project is successfully used in comparative plant genomics, in particular, for comparison of individual functional sites of Arabidopsis and rice genomes [38].

Arabidopsis has no economic significance [54]. Its genome will be used and is used now as a "basic" or "reference" genome as well as a donor of genes for biotechnological manipulations. The already available data on the rice genome are also successfully used for the same purpose. For example, it was recently found by comparative mapping that the gene *Vrn-A1* responsible for vernalization is located on chromosome 5A of hexaploid wheat and the corresponding gene *Hd-6* is located on the rice chromosome 3 [55].

No doubt, the number of works on comparative mapping of plants will sharply grow. Thus, a topical problem is isolation and cloning of the wheat gene *Ph1* responsible for homologous chromosome pairing in meiosis. It is believed that isolation of this gene will make a valuable contribution to cereal breeding. The gene is mapped on the wheat chromosome 5B, but neither obtaining a physical marker nor microdissection of this chromosomal region gave practical results [56]. One may hope that success will be achieved after comparative mapping of the corresponding loci of the wheat and Arabidopsis or rice genomes.

Phenotypic characters show that Arabidopsis and rice contain roughly the same number of genes, 25,000–30,000, and thus they differ three- or fourfold in the content of excessive DNA [20], although it is possible that the amount of duplicated genes in rice is somewhat higher than in Arabidopsis. Most likely, just the difference in the excessive DNA content is responsible for the delay in determining the total structure of the rice genome in spite of enormous effort put into the problem.

For this reason, speaking about "large" and "superlarge" plant genomes containing tentatively the same 25,000–30,000 genes [20, 39, 42, 57], one can rather definitely state that determination of their complete primary structure will meet the more difficulties the larger is the genome. To support this statement, one should remember that the genome of barley is ten times larger than that of rice, although they contain approximately the same number of genes.

Thus, one may assume that the development of plant genomics will first of all depend on the detection of structural analogies and genetic relationships between separate loci and chromosomes belonging to different genomes, and will be based on the knowledge of total structures of the Arabidopsis and rice genomes. In this context it is worthwhile to cite B. Barr (Brookhaven, USA), who mentioned that "sequencing of maize and wheat genomes would be simply foolish" [38]. However, it cannot be excluded that finally a plant with a large genome, such as a diploid cereal (for example, barley), will be added to the "reference" organisms.

In this field, which can be called comparative mapping of separate loci and functional linkage groups, most important are and will be the modern chromosome technologies. Knowledge of the nearly complete primary structure of the Arabidopsis genome opens new, sometimes hardly predicted prospects for comparative genomics of plants as well as other organisms including man. Thus, about 100 human "twin genes" were found in the Arabidopsis genome, including those responsible for such severe diseases as cystic fibrosis and breast cancer [50]. Quite interesting are also the results of comparison of groups of genes of Arabidopsis and other eukaryotes. Thus, the housekeeping genes are evolutionarily conserved in the yeast, Arabidopsis, and human genomes and have similar structure. At the same time, the genes characteristic only of multicellular organisms are different in man and Arabidopsis. Probably, evolution of genes responsible for "multicellularity" followed different ways in plants and animals [50].

CHROMOSOME TECHNOLOGIES IN PLANT GENOMICS

Chromosome technologies in plant genomics can be used in two directions:

(1) genome cloning and sequencing;

(2) genome studies on the basis of comparative genomics.

Although these directions are only conditionally separated, they require certain skills in identification of individual chromosomes, characterization of the plant karyotype, including its natural variability, isolation of individual chromosomes and their loci for cloning, and in locating cloned genes and anonymous nucleotide sequences on chromosomes.

Methods for studying individual plant chromosomes have been successfully developed during the last 25 years. They are mainly based on the C- and N-banding techniques yielding similar patterns. These techniques made possible description (with full identification of individual chromosomes) of karyotypes of main cereal crops (wheat, rye, barley, etc.) and some other plants, both economically important and model objects [58–67]. As a result, important information was obtained concerning the origin of species, their relationship, interspecific and varietal polymorphism, etc.

However, C-banding has significant limitations, being applicable only to rather large chromosomes rich in constitutive heterochromatin. It is also important that DNA retained in C-stained chromosomes cannot be used for further molecular-biological investigation, first of all, for cloning. For this reason, successful use of chromosome technologies in plant genomics requires their marked development and

improvement. The detailed knowledge of plant chromosomes selected for sequencing or study using the approaches of comparative genomics is necessary along with the development of a technology for banding that would allow the use of DNA for further cloning and creation of the chromosome-, locus-, and band-specific DNA libraries. Finally, the most important will be the improvement of *in situ* hybridization (FISH) techniques and solution of the problem of repetitive nucleotide sequences.

The main approaches used in genomics of man and some animals to obtain the chromosome- and locus-specific DNA libraries failed while they were applied in plant genomics. Up to the present time, there is no real success in obtaining pure fractions of individual plant chromosomes by flow sorting. In spite of some achievements, there are serious limitations in application of microdissection to plant chromosomes.

Microdissection is based on cutting out under the microscope an individual chromosome or its part, cloning the obtained material, and further creation of the chromosome- and locus-specific markers or "fishing out" the required clones from the genomic DNA libraries of a given organism [68, 69]. The material is obtained from chromosome spreads either by cutting out a necessary chromosome with a micromanipulator or by burning out with a laser the unnecessary part of the preparation [62], and then the necessary material is amplified and cloned.

In this way, chromosome-, locus, and band-specific DNA libraries are obtained. In particular, this approach was used to create the barley chromosome-specific markers that are successfully used for physical mapping of chromosomes of this species [37]. Laser microdissection is widely used in studies of rice genome, where this method was the basic one in obtaining chromosome-specific DNA libraries [43]. Among other economically-important plants whose genomes are successfully studied by microdissection, rye, barley, and sugar beet should be mentioned [70].

However, application of this approach to other plant chromosomes is connected with serious difficulties. Many species have very small chromosomes that are not identified or hardly identified even with the modern banding methods.

It was already noted that the main methods of plant chromosome analysis such as C- and N-banding make chromosomal DNA unsuitable for further cloning. It is necessary to develop a banding technique that would allow further cloning of obtained DNA. Especially interesting is the G-like staining that does not destroy DNA and makes possible identification of individual chromosomes or their loci [64, 71] as with human chromosomes.

DNA obtained by microdissection and amplification (mainly by DOT-PCR) contains mainly repetitive sequences of different composition and size, which are first of all used as probes for FISH. However, successful attempts to use microdissected material for chromosome painting [46], isolation of coding sequences [47], and genomic cloning [70] have been described. Two latter approaches are based both on "fishing out" sequences of interest from the total DNA libraries using chromosome-specific probes [47] and on creating plasmid DNA libraries with the material obtained by microdissection [70].

The use of FISH in plant genomics deserves special consideration. The above-mentioned abundance of repetitive sequences in plant genomes (up to 99%) makes difficult the application of this technique for mapping individual plant genes [40]. Certain complications arise also in comparative genomic hybridization (FISH-CGH) [71, 72]. At the same time, the method is successfully used not only in human genomics but in comparative analysis of human and animal genomes (ZOO-FISH). Thus, the use of fluorescence-labeled gibbon genomic DNA provides a new type of banding for human chromosomes. It can be assumed that such cross hybridization (plants × animals including man) may be useful in comparative plant genomics.

"Classical" cytodiagnostic technologies, first of chromosome banding, are and will be of extreme importance for comparative and evolutionary plant genomics. They are relatively cheap and quick in evaluation of intra- and interspecies variability, in studies of complex allopolyploid genomes such as tetraploid and hexaploid wheat, a wheat–rye hybrid triticale, and others, in analysis of evolution at chromosomal level, in investigation of formation of synthetic genomes and introgression of alien genetic material, and in revealing genetic relationships between individual chromosomes of different species [58–67]. No doubt, studying the plant karyotype by classical cytogenetic approaches, reinforced with rapidly developing methods of molecular biology and computer technologies of image analysis, will remain for a long time and maybe forever the most important approach to studies of the plant genome [40, 73].

These technologies are especially important for assessing such genome characteristics like the karyotype stability and variability at the level of separate organisms as well as for population, variety, and species. This is very important for allopolyploid species (wheat, oat, triticale). Finally, it is difficult to imagine how to evaluate the number and spectrum of chromosomal rearrangements without banding, and just this criterion seems to have prospects for monitoring the environment by the state of the plant genome.

ACKNOWLEDGMENTS

The work was supported by the Russian Foundation for Basic Research (projects nos. 99-04-48832, 00-04-49036, 00-04-81086, 01-04-48716) and the Russian State Program on Support of Leading Scientific Schools (project no. 00-15-97833)

REFERENCES

1. Aparicio, S., *Nature Genet.*, 1998, vol. 18, pp. 301–303.
2. Burris, J., Cook-Deegan, R., and Alberts, B., *Nature Genet.*, 1998, vol. 20, pp. 333–335.
3. Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., *et al.*, *Nature*, 1999, vol. 402, pp. 489–502.
4. Littele, P., *Nature*, 1999, vol. 402, pp. 467–468.
5. Phillips, R.L. and Freeling, M., *Proc. Natl. Acad. Sci. USA*, 1998, vol. 95, pp. 1969–2032.
6. Timbertake, W.E., *Nature Biotechnol.*, 1998, vol. 16, pp. 116–117.
7. Pen, A., *Science*, 1998, vol. 282, pp. 652–654.
8. James, R., *Nature Biotechnol.*, 2000, vol. 20, pp. 153–155.
9. Bennetzen, J.L., *Trends Genet.*, 1999, vol. 15, pp. 85–87.
10. Winkler, H., *Verbreitung und Versache der Parthenegenesis im Pflanzen und Tierre*, Jena, 1920.
11. Riger, R. and Michaelis, A., *A Dictionary in Genetics and Cytogenetics*, Moscow: Kolos, 1967, p. 81.
12. Kartel', N.A., Makeeva, E.N., and Mezenko, A.M., *Genetika. Entsiklopedicheskii slovar'* (Genetics. Encyclopaedic Dictionary). Minsk: Tekhnologiya, 1999, p. 90.
13. Kiselev, L.L., *Ontogenez*, 1999, vol. 28, pp. 233–236.
14. *Methods in Genome Analysis in Plants*, Jauhar, P.P., Ed., Boca Raton: CRC, 1996.
15. Baev, A.A., Venkstern, T.V., and Mirzabekov, A.D., *Mol. Biol.*, 1967, vol. 1, pp. 754–765.
16. Walbot, W., *Nature*, 2000, vol. 408, pp. 794–795.
17. Zhuravleva, G.A., Mironova, L.I., and Inge-Vechtomov, S.G., *Mol. Biol.*, 2000, vol. 34, pp. 560–571.
18. The *C. elegans* Sequence Consortium, *Science*, 1998, vol. 282, pp. 2012–2018.
19. Meyerowitz, E.M., *Nature*, 1999, vol. 402, pp. 731–732.
20. Lin, X., Kaul, S., Rounsley, S., *et al.*, *Nature*, 1999, vol. 402, pp. 761–767.
21. Mayer, K., Schuller, C., Wambutt, R., *et al.*, *Nature*, 1999, vol. 402, pp. 769–777.
22. The Arabidopsis Genome Initiative, *Nature*, 2000, vol. 408, pp. 796–813.
23. Theologis, A., Ecker, J.R., Palm, C.J., *et al.*, *Nature*, 2000, vol. 408, pp. 816–820.
24. Salanoubat, M., Lemcke, K., Rieger, M., *et al.*, *Nature*, 2000, vol. 408, pp. 820–822.
25. Tabata, S., *Nature*, 2000, vol. 408, pp. 823–826.
26. Ventor, J.C., Adams, M.D., Sutton, G.G., *et al.*, *Science*, 1998, vol. 280, pp. 1541–1543.
27. Tatevosov, S., *Kommersant Den'gi*, 2000, no. 20 (273), pp. 45–50.
28. Adams, M.D., Celniker, S.E., Holt, R.A., *et al.*, *Science*, 2000, vol. 287, pp. 2185–2195.
29. Hodgkin, J., *Nature*, 2000, vol. 404, pp. 442–443.
30. Butler, D., *Nature*, 1999, vol. 401, pp. 729–730.
31. Kornberg, T.B. and Krasnov, M.A., *Science*, 2000, vol. 287, p. 2218.
32. Rubin, G.M. and Lewis, E.B., *Science*, 2000, vol. 287, pp. 2216–2218.
33. Flavell, R.B. and Smith, D.B., *Heredity*, 1976, vol. 37, pp. 231–252.
34. Flavell, R.B., Rimpau, J., and Smith, D.B., *Chromosoma*, 1977, vol. 63, pp. 205–222.
35. Flavell, R.B., O'Dell, M., and Smith, D., *Heredity*, 1979, vol. 42, pp. 309–322.
36. Bennett, M.D., *Proc. Natl. Acad. Sci. USA*, 1998, vol. 95, pp. 2011–2016.
37. Bevan, M. and Murphy, G., *Trends Genet.*, 1999, vol. 15, pp. 211–214.
38. Adams, D., *Nature*, 2000, vol. 408, pp. 792–793.
39. Bennetzen, J.L., *Proc. Natl. Acad. Sci. USA*, 1998, vol. 95, pp. 1975–1978.
40. Jiang, J. and Gill, B.S., *Genome*, 1994, vol. 37, pp. 717–725.
41. Tanksley, S.D., Ganal, M.W., and Martin, G.R., *Trends Genet.*, 1995, vol. 11, pp. 63–68.
42. Gale, M.D. and Devos, K.M., *Proc. Natl. Acad. Sci. USA*, 1998, vol. 95, pp. 1971–1974.
43. Busch, W., Martin, R., Herrmann, R.G., and Hohmann, U., *Genome*, 1995, vol. 38, pp. 1082–1090.
44. Mao, Y.W., Liang, S.Y., Song, W.Q., *et al.*, *Cell Res.*, 1998, vol. 8, pp. 285–293.
45. Zhou, Y., Hu, Z., Dang, B., *et al.*, *Chromosoma*, 1999, vol. 108, pp. 250–255.
46. Shibata, F., Hizume, M., and Kuroki, Y., *Chromosoma*, 1999, vol. 108, pp. 266–270.
47. Hernould, M., Glimelius, K., Veuskens, J., *et al.*, *Plant J.*, 1997, vol. 12, pp. 703–709.
48. Moore, G., Gale, M.D., Kurata, N., and Flavel, R., *Bio-Technology*, 1993, vol. 11, pp. 584–589.
49. Messing, J. and Llaca, V., *Proc. Natl. Acad Sci. USA*, 1998, vol. 95, pp. 2017–2020.
50. Penisi, E., *Science*, 2000, vol. 290, pp. 2220–2221.
51. Sommerville, C. and Dangl, J., *Science*, 2000, vol. 290, pp. 2077–2078.
52. Vision, T.J., Brown, D.G., and Tanksley, S.D., *Science*, 2000, vol. 290, pp. 2114–2117.
53. Sasaki, T., *Proc. Natl. Acad Sci. USA*, 1998, vol. 95, pp. 2027–2028.
54. Dennis, C., *Nature*, 2000, vol. 408, p. 791.
55. Kato, K., Miura, H., and Sawada, S., *Genome*, 1999, vol. 42, pp. 204–209.
56. Ratber, M., *Nature Biotechnol.*, 1998, vol. 16, pp. 810–811.
57. Moore, G., Foote, T., Helentjaris, T., *et al.*, *Trends Genet.*, 1995, vol. 11, pp. 81–82.
58. Linde-Laursen, I., *Cytology and Cytogenetics of Hordeum vulgare and Some Alien Species Using Chromosome Banding Technique*, Riso: R-529, 1985.

59. Muravenko, O.V., Badaev, N.S., Badaeva, E.D., *et al.*, *Dokl. Akad. Nauk SSSR*, 1986, vol. 288, pp. 724–727.
60. Badaeva, E.D., Sozinova, L.F., Badaev, N.S., *et al.*, *Cereal Res. Commun.*, 1990, vol. 18, pp. 273–281.
61. Badaeva, E.D., Budashkina, E.B., Badaev, N.S., *et al.*, *Theor. Appl. Genet.*, 1991, vol. 82, pp. 227–232.
62. Gill, B.S., Friebe, B., and Endo, T.R., *Genome*, 1991, vol. 34, pp. 830–839.
63. Friebe, B. and Gill, B.S., *Methods in Genome Analysis in Plants*, Jauhar, P.P., Ed., Boca Raton: CRC Press, 1996, pp. 39–60.
64. Muravenko, O.V., Fedotov, A.R., Punina, E.O., *et al.*, *Genome*, 1998, vol. 41, pp. 616–625.
65. Badaeva, E.D., Friebe, B., and Gill, B.S., *Genome*, 1996, vol. 39, pp. 293–306.
66. Badaeva, E.D., Friebe, B., and Gill, B.S., *Genome*, 1996, vol. 39, pp. 1150–1158.
67. Samatadze, T.E., Muravenko, O.V., Klimakhin, G.I., and Zelenin, A.V., *Genetika*, 1997, vol. 33, pp. 130–132.
68. Kao, F.T., *Methods in Genome Analysis in Plants*, Jauhar, P.P., Ed., Boca Raton: CRC, 1996, pp. 329–343.
69. Rodova, M.A., Zoshchuk, S.A., Barskii, V.E., *et al.*, *Genetika*, 1995, vol. 31, pp. 1016–1020.
70. Jung, C., Claussen, U., Horsthemke, B., *et al.*, *Plant Mol. Biol.*, 1992, vol. 20, pp. 503–511.
71. Heslop-Harrison, J.S. and Schwarzacher, T., *Methods in Genome Analysis in Plants*, Jauhar, P.P., Ed., Boca Raton: CRC, 1996, pp. 163–180.
72. Mukai, Y., *Methods in Genome Analysis in Plants*, Jauhar, P.P., Ed., Boca Raton: CRC, 1996, pp. 181–194.
73. Gill, B.S. and Friebe, B., *Current Opin. Plant Biol.*, 1998, vol. 1, pp. 16–39.

Note Added in Proof

Serious events happened in structural genomics during the time that passed after submission of this manuscript. On February 15 and 16, 2001 *Nature* and *Science* published the results of the 12-years-long efforts of the world community on deciphering the human genome. They are presented in two big volumes and consist of scientific reports of numerous members of different research groups and various scientific comments.

The detailed analysis of results requires thorough consideration both of already published results and data of genomic databases, which will take much time and is far beyond the limits of this paper (for details, see [1–17]). Here we shall try to consider only how and to what extent these achievements will influence the strategy and fate of structural studies of plant genomes. However, this requires at least a bried analysis of recent results.

The work on structural investigation of the human genome was carried out by two research groups whose main directions were mentioned above: the International Consortium and a private American company Celera Genomics. These research groups are extraordinarily large: publications of Celera Genomics [1] include over 450 authors, whereas the complex of papers issued by the International Consortium in *Nature* includes more than 4000 authors [11–13].

Data on genome structure published by the Consortium were obtained on the basis of a "classical" approach of successive sequencing (clone-to-clone technology). In several vast publications there are data both on draft sequencing of human genome and studies of individual chromosomes [14–16]. However, it should be noted that the results of these chromosomal studies are much inferior in detail to those published before on human chromosomes 21 and 22.

While papers concerning chromosomes 21 and 22 (see above) contain the term "sequencing" in their titles, articles about chromosomes 12, 14, and Y deal with their physical mapping [14–16], and physical maps for sequencing are described for chromosomes 1, 6, 9, 10, 13, 20, and X [17].

The results presented by Celera Genomics are largely based on the total sequencing technology and were obtained during the last two and a half years. However, it is said in the abstract that the published data are the result of team work of the Celera and International Consortium groups. The detailed list of available data banks used in this work is given. A general conclusion is drawn that both genomic strategies led to very close results. It is underlined that at present the point is obtaining only a draft sequence of human genome rather than a substantially completed one, as stated concerning chromosomes 21 and 22.

No doubt, obtaining even a "draft" sequence of the total human genome is of great scientific interest.

In particular, it confirms the earlier hypothesis that the human genome contains about 30,000 genes [5]. It was shown that 1% of the genome is occupied by exons, 24% by introns, and 75% by intergene DNA. The distinctions between haploid genomes are on average 1 bp per 1250. The differences between the human genome and other sequenced genomes were thoroughly analyzed. The differences between the human genome and those of anthropoid apes do not exceed 1–2%.

At the same time, the published results make still more evident that the strategy of total sequencing cannot be applied directly to large and very large genomes without using the results of classical technology of consecutive sequencing. Obtaining a draft variant of the human genome structure took about

12 years, and by the time of beginning of these investigations there were much more data on some parameters (the number of revealed and mapped genes, development and successful use of cell hybrids, including the radiation ones, the technology of preparing pure fractions of individual chromosomes, etc.) as compared with the available information about the genomes of the most important crops such as wheat, maize, and barley.

Therefore, our conclusion that the structural genomics of plants with large and superlarge genomes will develop in the nearest future on the basis of the comparative principle (comparative genomics) has not been shaken in the least.

REFERENCES

1. Ventor, J.C., Adams, M.G., Myers, E.W., *et al.*, *Science*, 2001, vol. 291, pp. 1304–1351.
2. Pennisi, E., *Science*, 2001, vol. 291, pp. 1177–1180.
3. Roberts, L., *Science*, 2001, vol. 291, pp. 1182–1188.
4. Marshall, E., *Science*, 2001, vol. 291, pp. 1189–1193.
5. Claverie, J-M., *Science*, 2001, vol. 291, pp. 1255–1257.
6. Galas, D.J., *Science*, 2001, vol. 291, pp. 1257–1261.
7. Baltimore, D., *Nature*, 2001, vol. 409, pp. 814–816.
8. Olson, M., *Nature*, 2001, vol. 409, pp. 816–818.
9. Bork, P. and Copley, R., *Nature*, 2001, vol. 409, pp. 818–820.
10. Birney, E., Bateman, A., Clamp, M.E., and Hubbard, T.J., *Nature*, 2001, vol. 409, pp. 827–828.
11. The International Human Genome Consortium, *Nature*, 2001, vol. 409, pp. 860–921.
12. The International Human Genome Consortium, *Nature*, 2001, vol. 409, pp. 934–941.
13. Aach, J., Bulyk, M.L., Church, G.M., *et al.*, *Nature*, 2001, vol. 409, pp. 858–859.
14. Montgomery, K.T., Lee, E., Miller, A., *et al.*, *Nature*, 2001, vol. 409, pp. 945–946.
15. Bruls, T., Gyapay, G., and Petit, J.-L., *Nature*, 2001, vol. 409, pp. 947–948.
16. Tilford, C.A., Kuroda-Kawguchi, T., Skaletski, H., *et al.*, *Nature*, 2001, vol. 409, pp. 943–945.
17. Bentley, D.R., Deloukas, P., French, L., *et al.*, *Nature*, 2001, vol. 409, pp. 942–943.