Genomics, DNA chips and a revolution in plant biology

S. C. Maheshwari, Nirmala Maheshwari and S. K. Sopory

International Centre for Genetic Engineering and Biotechnology, B.O. Box 10504, Aruna Asaf Ali Road, New Delhi 110 067, India

The article gives an overview of current research on plant genomics. We begin with a brief account of the major techniques developed for the Human Genome Project such as preparation of RFLP maps, making of cosmid and YAC (later also BAC and PAC) libraries, construction of contigs and finally sequencing such as by use of M13 vectors and use of modern machines and advanced computers. We then introduce to the reader the more recent work with *Arabidopsis* and rice genomes utilizing the various basic techniques. The article ends with a brief account of the new DNA microchip technology, whereby one can simultaneously monitor expression of hundreds and thousands of genes. Microchips thus are the hub of a new revolution in functional genomics.

GENOMICS is a rapidly emerging area of research, which came into existence in the closing years of the last century and promises to become a dominant theme of intellectual activity in the coming decades, revolutionizing our understanding of biology as never before[1–3]. Plant genomics is a part of this larger field and embraces the study of whole genomes of plants, their physical and molecular organization, evolution and functions of the myriads of constituent genes. Included in it are also studies on genome-wide or global patterns of gene expression, that have recently become possible by the new DNA chip technology, allowing visualization of activity of hundreds and thousands of genes simultaneously[4] (for more information and a general view, see refs 5–8). The sequencing of entire genomes is in fact propelling the development of many other novel technologies and the shaping of a new discipline of bioinformatics. Genomes of eukaryotes comprise thousands of genes and millions of base pairs – storing the output of massive data and information not only on genes themselves, but relating also to their expression, and later retrieving and analysing this information requires new computational tools, even new mathematical techniques, not envisioned before.

The advent of genomics is a consequence of the discovery of the elegant and simple procedure of sequencing of nucleic acids pioneered by Sanger and coworkers in UK and by Maxam and Gilbert in USA in 1977 (refs 9, 10). But it is not so much this technical ability *per se* which has aroused world-wide interest in genomics – rather the discipline has assumed tremendous importance today because deciphering the sequence of bases of DNA would unlock the whole blueprint of the development of an organism. Appropriately, therefore, the impact of research on genomics, of which the Human Genome Project (HGP) is a prime example, has been compared to the discovery and consolidation of the Periodic Table in chemistry[1]. The study of various genomes also holds the key to understanding the origin and evolution of plants and animals – as put so aptly by Lander and Weinberg[2], it is a journey to the centre of biology. But genomics is not merely of fundamental or academic interest; it indeed will be of vital importance for the agriculture of tomorrow[11–14] as it is proving already for medicine and human welfare. Only a few plant species stand between prosperity, on the one hand, and hunger and starvation on the other. Clearly, if we understand the genomes of crops like rice, wheat, maize, beans and potato, we can ensure a better future with the capability of surer, more perfect and precision genetic manipulation for enhanced yield and survival under adverse conditions.

Historical

The HGP launched in 1988, with the biggest infusion of money (the total project is expected to cost over 3 billion dollars, equivalent to about 3500 crores rupees) thus far in any biological project, is the first major endeavour to characterize a higher genome and came into being because of the foresight and vision of pioneers like Watson and Sanger[15–17]. However, going back a little further, the HGP and other related projects undertaken (and some completed in the meantime, e.g. on *E. coli*[18], yeast[19–21] and the nematode *Caenorhabditis*[22]) are themselves an extension of the classic studies by Morgan, Sturtevant and colleagues on mapping genes by linkage analysis[23,24]. The first ever genetic map prepared was that of *Drosophila* in 1919 (ref. 23). Turning to plants, the pioneering studies of Emerson and his

*For correspondence. (e-mail: maheshwarisc@hotmail.com)

associates like Stadler, Rhoades, McClintock and others in the forties led to the establishment of the linkage map of maize[24]. Yet, it is the molecular biology revolution heralded by the discovery of the double helix in the fifties[25], leading to studies of how genes function and then proceeding to cloning and more importantly sequencing of genes in the seventies, that resulted in the arrival of the genomic biology on the global scene.

However, genome-wide sequencing itself could not have proceeded without extensive automation. High-through-put (HTP) is the keyword in the context of genomics and it is investigators like Hood and Hunkapiller who in the mid-eighties prepared the ground for the modern large-scale sequencing by innovation of methods such as use of fluorescent primers or fluorescent dideoxynucleotide terminators, laser excitation of bands of newly synthesized DNA, sensing of their colours by photomultipliers and finally automatic output of sequence data through computers attached to sequencers[26,27].

Basic techniques

Problems with eukaryotic genomes

It is the grand goal set in the HGP which spawned the development of additional new methodologies required for sequencing whole genomes. Work on the yeast *Sachharomyces cereviseae*[19–21] and *Caenorhabditis elegans*[22], which was undertaken as pilot projects to enable evaluation of the new procedures at the behest of Watson, Sanger, Brenner and others, showed the way eukaryotic genomes could be conquered. To briefly review the fundamentals, certainly base sequencing is the essence of any genome sequencing project – Sanger's dideoxy chain terminator method is now preferred over that of Maxam and Gilbert. Given the fact that the technique is in use in hundreds of laboratories, such work may at first sight appear very simple, both conceptually and operationally. But the actual execution of any genome project for multicellular organisms is rather complex because of the great mass of DNA in them[28–30]. In humans, we have about 3.2 billion base pairs of DNA and probably 50,000–100,000 genes. Many plants have comparable or larger genomes – maize has about the same amount of DNA as humans and wheat has about five times more than us. Even *Arabidopsis*, known to possess the smallest genome among flowering plants, has 120–140 million base pairs of DNA and around 20,000–25,000 genes. In sharp contrast, for sequencing, at a time, in single lane and in a single gel in the electrophoresis apparatus, one can handle DNA fragments that are only 1–2 kb pairs long (actual sequences may be read of only 200–500 bases). The problem of sequencing eukaryotic genomes is exacerbated by the fact that if we were to digest the DNA of a higher organism to fragments of such a small size – right from the start – millions of fragments would result, which would make our task not only totally unmanageable, but would leave us no easy way of knowing their exact addresses in the genome.

The first steps – Construction of linkage maps with RFLP or other markers

For reasons explained above, in higher organisms all genome projects have had to follow, at least thus far, a somewhat orderly approach for the construction of genome libraries and their analysis[29,30]. The establishment of a classical 'genetic' linkage map has been a prerequisite for providing a rough road map to the order of at least some genes. These maps, however, suffer from the disadvantage that neither the 'visible' markers (e.g. flower colour) represent molecular markers nor is their position precise since they are based on recombination frequencies, the distances being measured in centimorgans (cM). From a genetic linkage map, thus, one has to proceed to the construction of a 'physical' map with landmarks along the whole length of the DNA molecule, that comprises each chromosome and with distances measured in base pairs. In this regard, a major advance occurred in 1980 when the concept of molecular marker maps was developed by Botstein and coworkers[31,32], which relied on restriction fragment length polymorphism (RFLP) seen commonly between the genomes of allied taxa, species or races, and its detection by Southern blotting. Even though RFLP markers are also positioned initially by linkage analysis and thus only approximately, they represent real molecular markers that can be actually hybridized to cloned DNA fragments and thus let the work on construction of an accurate physical map to begin. An essential step in genome analysis – after digestion of DNA into smaller fragments, whether by restriction endonucleases or by shearing – is cloning in a vector. Cloning enables multiplication of the DNA fragments for further handling or distribution to different laboratories, and an extensive collection of these clones constitutes a genomic library. The RFLP markers (which in earlier years largely represented random genomic DNA clones, but increasingly now comprise cDNAs or genes already identified) are then hybridized to the vectors containing the DNA fragments that are to be sequenced, enabling the positioning of the various cloned fragments along the chromosomes in the cytogenetic map.

YAC libraries

Among other significant advances that occurred during this period was the development of cosmid vectors[33],

which allowed 40–50 kb pairs long DNA fragments to be cloned, instead of those only a few to around 20 kb pairs long that was earlier possible in the ordinary bacterial cloning vectors such as pBR322, pUC or the lambda phage. However, even cosmid vectors were not quite appropriate for the task of sequencing the human genome and it is the development of YACs (yeast artificial chromosomes) in the late eighties which allowed longer, up to 1000 kb, fragments to be cloned that gave a real impetus to genome studies[34]. Not only cloning of such large fragments reduced the number of potential gaps in the assembly of contigs, but the entire task was made more manageable – although one still dealt with thousands of YAC colonies requiring special procedures to store, further replicate and analyse the clones. Work on the human genome gives an idea of the scale of effort required – even though 7500 YAC clones with an average insert size of ~400 kb DNA can contain the entire genome, actually as many as 60–70,000 clones have had to be screened by a thousand or more probes. This is because some redundancy in preparing libraries is essential to ensure that overlapping clones are available for the entire genome. To prepare libraries, after the colonies are individually picked from petri dishes, they are first transferred to 96-well microtitre plates[17,28]. But since 96-well plates are rather inconvenient to replicate, manage and distribute, the colonies are gridded on special nylon filters by a 96-prong spotting device operated by a robot which enables repetitive stamping on large nylon filters for preparation of replicates. The filters too can be made at very high density and the number of spots varied from 576 to 9200 per 22 cm^2 membrane (this is achieved by slightly offsetting the delivery of the new set of drops).

Screening libraries, ordering clones and constructing contigs

The nylon membrane can be made to rest on an agar culture medium from which cells can absorb nutrients and further multiply. But it can also be dried, and this greatly facilitates the screening process since the filters can be bathed by a DNA probe such as by a RFLP marker (the cells are lysed *in situ* and yeast DNA denatured before hand). Any spot that hybridizes can be detected by autoradiography and the corresponding yeast clone located in the microtitre plate[17]. However, RFLP markers are few and far between and often there is only one marker per YAC clone. So when the PCR technique became available, to order the clones more precisely (and have an idea of the overlap relationship of each clone), in the early nineties, the concept of marking them by STS (sequence tagged *s*ites) was developed[35,36]. STS represent a landmark of a unique short ~ 300 bp DNA sequence that can be had from any region of the cloned DNA, the critical feature being that it should be possible for probes to be generated by PCR by defining a pair of sites for annealing primers. To generate STS, only single-pass partial sequencing of clones is required and it is through such probes that an investigator can determine the overlap relationship between clones, e.g. whether a clone under test, is identical, shorter or longer than a previously tested clone or represents a new clone altogether.

In the early days of the HGP, one worried about organizing an expensive central store for all the YAC libraries, the stability of clones as also physical transportation of probes. The advent of STS made these concerns superfluous since information about STS needs to exist only in a network of computer databases, which makes it possible to mark YAC clones in any laboratory across the world without having to exchange any probe physically. Analogous to STSs are the ESTs, the so called *e*xpressed *s*equence tags, which came soon after and increasingly represent an important parallel way of not only generating markers for identifying clones, but of finally annotating the sequences and reaching the goal of total sequencing of genome. Work on ESTs was pioneered by Venter and colleagues[37] in USA who saw their many advantages over conventional STSs. Like STSs, ESTs represent single-pass sequences, but *unlike* STSs – which are derived from genomic clones – ESTs represent short sequences of cDNAs from the 3′ or 5′ ends. Not only are EST obtained far more easily, they are very useful because the cDNAs from which they are derived represent genes that are actually expressed and ESTs, therefore, greatly facilitate the task of annotation of the final sequence.

When ordered overlapping clones are available, contigs are constructed which represent stretches of contiguous sequence-ready clones[38]. As work picks up, the number of contigs increases for some time, but later decreases and ultimately one should ideally be left with only one contig per chromosome. But often, this is not possible on the basis of overlapping clones alone and one has to resort to gap closure by chromosome walking for which special probes have to be generated from clones at the ends of contigs, by either plasmid rescue or by inverse polymerase chain reaction. These probes are then employed to identify new YAC clones to bridge the gaps. However, in certain studies earlier, cosmid-based contigs were constructed first and one used a larger YAC clone to bridge the gaps[39].

Sequencing

After a minimal set of overlapping YAC clones has been identified representing a *tiling path*, the DNA in each clone is further fragmented either randomly or by restriction digestion and then sub-cloned in cosmids as

per standard practice. Overlapping cosmid clones are identified by fingerprinting which is done by restriction-digesting the clones, filtering out common bands of vectors with the aid of computers and then comparing the digestion patterns to find contiguous clones (25–50% similarity is indicative of an overlap). The cosmid clones are then ready for further breakage – this is done randomly either by shearing or by sonication – into 1–2 kb fragments and sub-cloning into pUC plasmids or M13 phage. Finally, the DNA from each sub-clone is extracted, reactions set-up for *in vitro* DNA synthesis as necessary in the Sanger method employing the M13 phage, and sequenced[40].

New vectors BACs and PACs and the shot-gun approach

In the early days of the HGP, one placed great reliance on YAC clones and many YAC libraries were constructed all over the world with a total redundancy of 60 genomic equivalents. However, since YAC clones were often found to be unstable or chimeric (incorporating more than one DNA fragment), efforts were already well underway by the mid-nineties to develop alternate vectors like BACs or PACs[41–43]. Satisfactory BAC vectors have since been developed and, of late, DNA fragments (~ 100 kb long) are being directly cloned in BACs. Overlapping BACs can be found by fingerprinting – one can then proceed directly to sub-cloning and sequencing. This is a major improvement since, unlike YACs, which can be multiplied only in yeast cells, BACs are maintained in bacteria, simplifying all pre-sequencing steps. What is more, the arrival of BACs has been followed by the proposal in recent years to do away with the initial ordering of BACs by RFLP or other genetic markers. In fact, with the aid of computers and special software, Venter and associates have found that provided a BAC library is prepared at a redundancy of 10, the major sequencing effort can largely be restricted to BAC ends alone (the authors call them STCs or sequence tagged connectors) which provide adequate information for determining overlaps[44]. In this strategy, while the ends of every BAC are sequenced, the number of BACs that need to be sequenced fully is much smaller and is determined later. To begin construction of a contig, one starts with a fully sequenced 'seed' BAC. The database search for BAC-ends may, now, reveal overlaps with about 30 other BACs. And having located the two most important extreme-end BACs, one sequences them fully, continuously expanding search for new BACs for full sequencing at either end, until one tackles an entire chromosome and then the whole genome! This random shot-gun method does require more extensive sequencing and powerful computers, but it has been claimed that in the end it more than compensates for the time and money that go in a conventional sequencing procedure, where considerable effort has to go just in constructing linkage maps of molecular markers.

Sequencing the *Arabidopsis* genome

It is but natural that the HGP should have aroused worldwide interest among plant biologists to undertake similar ventures in plants. Special credit goes to three American groups, namely of Meyerowitz (at Pasadena), Somerville (earlier at East Lansing, now at Palo Alto) and Goodman (earlier at Penn State, now at Harvard) for taking the first steps leading to the organization of a $12.7 million (Rs 500 crores) multinational project on sequencing of the *Arabidopsis* genome[45–47] and which later has come to be known as the AGI (the *Arabidopsis* Genome Initiative). Indeed, the project came into existence close on the heels of the HGP in 1990. Earlier, the groundwork of genome mapping studies had been laid in the laboratories of Redei[48] (at Columbus, Ohio) and Koornneef[49] (Wageningen) who made the first classical linkage map.

RFLP maps and work on genomic libraries

The first RFLP map of *Arabidopsis* was prepared by the group of Meyerowitz[50] and then a second one by that of Goodman[51]. The two RFLP maps were later integrated[52] by a collaborative effort between these two laboratories and also of Koornneef and the number of markers greatly increased in recent years by Whittier, Dean and co-workers[53] working at Tsukuba in Japan and the John Innes Institute at Norwich in the UK. A point to note is that the early RFLP maps were derived from F3 individuals following a cross between two ecotype races and the original seed stock had been exhausted. The Recombinant Inbred (RI) lines, which were later developed by Reiter *et al.*[54] and Lister and Dean[55] and derived by repeated selfing of single-seed descent lines, provided a permanent stock of RFLP marker lines to which more markers could be added. By 1996 the distance between RFLP markers had been reduced to > 1.0 cM which represents a physical spacing of about ~ 200 kb (and suitable for chromosome walking). Apart from RFLP markers, which are seen only by Southern analysis, in recent years, many PCR-based markers have been added to the *Arabidopsis* genetic map, namely RAPDS, CAPS, AFLPs, SSPs and MS (microsatellite markers). In certain cases, RFLP markers have been converted to PCR-based markers[56]. As to YAC libraries, among the first to be made were those by Ward and Jen[57] of CIBA-GEIGY (USA) and the Somerville group[58]. A French group[59] subsequently made the CIC large insert (~ 400 kb) library which has been of con-

siderable use to later workers. For reasons explained earlier, however, these libraries are now being gradually replaced by BAC and P1 phage-based libraries. The TAMU-BAC library was made by Rod Wing and co-workers[60] at Texas A&M University and another, the IGF-BAC library has been made more recently by Altmann, Lehrach and co-workers at the Institut für Genbiologische Forschung and the Max Planck Institutes for Molecular Biology and for Molecular Genetics in Berlin and Golm[61]. Workers in Japan have been active in constructing P1-based libraries[62].

Physical maps and sequencing of chromosomes 2 and 4

An international collaborative group comprising some 25 laboratories spread over USA, Europe and Japan has undertaken the sequencing project under the AGI, with various chromosomes assigned to different groups. As in other genome sequencing projects, work on even the same chromosome has often to be assigned to more than one laboratory. Dean and co-workers constructed the first physical map for chromosome 4 followed by that for chromosome 2, which together have led the way for sequencing of the entire *Arabidopsis* genome (see ref. 63 for a review). Initially, the conventional 'clone-by-clone' approach was the dominant approach[37]. However, after Venter and co-workers advocated the benefits of the shot-gun approach for sequencing the human genome, more and more laboratories are now adopting it for sequencing genomes also of plants. Since ordered libraries of YAC clones had already been available, currently a 'hybrid' strategy is being followed in which the new BAC clones are being hybridized to YACs already aligned to chromosomes, helping BAC-based physical maps to be made with speed[64]. From this point on, a random shot-gun approach is being employed for assembling complete sequences. Extensive use has also been made of cDNA libraries and ESTs in recent years.

A high point in this field has been the completion of sequencing of chromosome 2 by the American group[65] and chromosome 4 by the European group[66] at the end of 1999 (Figure 1; for an excellent commentary see ref. 67). Close to 8,000 genes (thought to code for proteins) have been found in these chromosomes. To sequence nearly 40 Mb of DNA in the two chromosomes and clear a path through the jungle, nearly 400 BAC clones have been employed (including also a few P1 clones). For chromosome 4, additional 56 cosmid clones were employed. And to give an idea of the enormity of the task, a total of 5,90,165 reactions for sequencing have had to be carried out for chromosome 2 alone. Progress is now so rapid (with 200 genes being added every month to the database) that by the time this issue is in print the sequence of the whole genome would become available about 4 years ahead of the original schedule, no doubt due to further automation and vastly improved procedures. In fact, Davis and co-workers at the Stanford DNA Sequencing and Technology Development Center have been working on designing and fabrication of robots that handle as many as 10,000 samples per day from libraries of M13 plaques – allowing sequencing of 1 Mb of DNA, at a cost of $0.29 (Rs 13–14) per base[68].

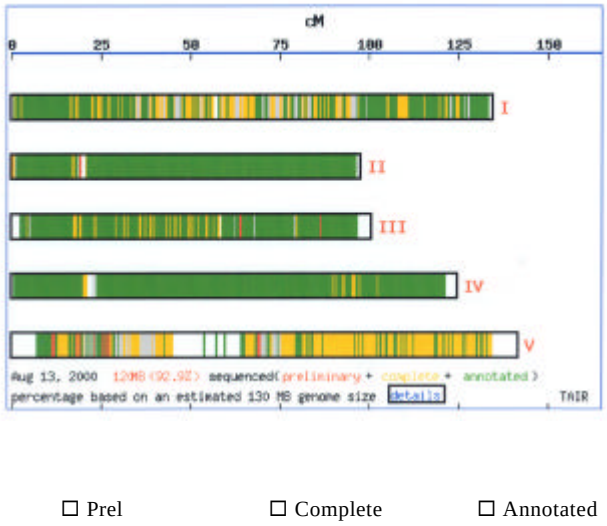

Figure 1. Progress in sequencing of the *Arabidopsis thaliana* genome as accessed in August 2000 via internet through the Arabidopsis Information Resource Site (TAIR).

Rice genome project

Realizing that nearly half of the world's population depends on rice for their daily diet and the obvious significance of this crop, the Japanese started a Rice Genome Programme (RGP) in 1991 (refs 69, 70), with keen foresight and wisdom, and for which they deserve the admiration of the scientific community all over the world. In 1998, a 10-year duration, second phase of the RGP was launched with the objective of complete sequencing of the genome of this plant and, of course, initiate work also on functional analysis of genes. Towards this aim, in fact, one International Rice Genome Sequencing Programme was initiated and, as of now, 10 countries, including Canada, China, France, Japan, Korea, Taiwan, Thailand, UK, USA and India are actively involved in the project. A new database known as the INE (an abbreviation for Integrated Rice Genome Explorer) has also been established – pronounced I-ne, it also refers to the rice plant in the Japanese language[71]. The rice genome is estimated to comprise ~430 Mb DNA, which is about four times the amount of DNA in *Arabidopsis*.

RFLP linkage maps and ESTs

At the time of writing this article, none of the rice chromosomes had been completely sequenced. But a very large part of the genome had been cloned in YACs and first generation physical maps of the entire set of chromosomes, from 1 to 12, already published between 1996 and 1997 with a genome coverage of about 78%. Probably, the total number of genes in rice is not much larger than in *Arabidopsis* and most of them should have homology between the two plants. But serious difficulties in sequencing rice genome exist because of repeated sequences which make up much of its bulk.

As with *Arabidopsis*, the initial phase of the rice project has been taken up largely with the development of molecular marker maps. The first linkage map of 135 RFLP markers was published in 1988 by McCouch *et al.*[72], as a result of a collaboration between IRRI and Cornell University and employing a cross between an *indica* and a *javanica* variety. The same group[73] later brought out a saturated map of 726 markers (mainly RFLP), this time employing a back-cross between *O. sativa* and an african wild rice, *O. longistaminata*, with an average distance of about ~2 cM between markers. However, in recent years, the Japanese have done far more extensive work under the auspices of the RGP[74]. The latest of their maps[75], one of the densest known of any plant yet, is largely derived from a cross between the *japonica* variety Nipponbare and the *indica* variety Kaslath, and comprises 2275 markers with an average distance of only ~ 0.5 cM between markers and corresponding to a physical distance of ~150 kb. Many other types of markers, for example RAPD – based on PCR technique, have also been made. Thus, similar to *Arabidopsis*, the early work began with random genomic clones, but in recent years cDNAs and ESTs have supplied a large number of markers[76,77]. At this point, nearly 40,000 ESTs had been catalogued and of these *ca* 4000 integrated in the linkage map.

Physical maps and progress towards sequencing

At the time when we began writing this article, it appeared that complete sequencing of rice genome might still take a few years, although considerable progress had been made in preparing not only contig maps employing YACs, but also BACs (for a detailed review of work until 1997, see refs 78 and 79). Nevertheless, while this article was in the final phase of preparation, a sensation had been created by a press release of Monsanto (now a division of Pharmacia Corporation) and commentaries in journals, thereafter, that the rice genome has already been largely sequenced[80]! The work has mainly been done at Washington University, Seattle by Hood, Mahairas and associates. Since all this work was so far secret, it has certainly left a sense of unease among the participants of the RGP. But it should help the world generally and enable precise and complete sequencing done much earlier than originally scheduled. On its part, the Japanese Government has recently invested more money in the project, redoubling its effort.

Interim lessons from plant genome projects

The complete sequencing of the *Arabidopsis* genome is scheduled to be completed by the end of the year 2000. While the Monsanto group may also have large tracts of rice DNA sequenced, we may have to wait, however, for a few years to obtain a completed and annotated sequence. Only when both these plants are completely sequenced, one can have a proper idea of the basic principles of genomic organization of plants. The larger questions, e.g. whether there is any grand design relating to the distribution of genes and various classes of repeated DNA (whenever present) within a genome, cannot be answered as of now. However, some interim lessons can be had, and specially so, from the complete analysis of chromosomes 2 and 4 of *Arabidopsis*. Several points are noteworthy: (1) The genome of *Arabidopsis* is very compact – on an average, there is a gene every 4–5 kb, the approximate length of a coding region is 2–2.5 kb; (2) The telomere and centromere regions are full of repeated DNA families; (3) Several hundred transposons and pseudogenes are present in the centromeric regions; (4) There is extensive duplication of genes, with many genes occurring as multigene families such as those coding for rubisco, cab and Ca^{+}-binding proteins; (5) Sometimes entire stretches of DNA and genes are duplicated between chromosomes, e.g. chromosome 4 has a 37-gene sequence duplicated in chromosome 5; (6) Interestingly, sometimes, there are clusters of genes that are commonly regulated, for example, by abscisic acid, as in chromosome 2; (7) Approximately 20% genes have signal sequences that target products into organelles, i.e. chloroplasts or mitochondria. A rather surprising finding is that many chloroplast and mitochondrial genes, coded normally by organellar DNA, also exist in the nuclear genome. Indeed, almost the entire mitochondrial genome of *Arabidopsis* is represented also on chromosome 4 at the centromeric region.

Another significant outcome of the *Arabidopsis* genome project is that, for the first time, it has become possible to have a rough estimate of the number of genes in various functional groups, now that 50–60% of the ORFs and genes can be assigned some role on the basis of homology in the BLAST and other databases. That as much as nearly 20% of the genes involved in

coding for energy-related functions mentioned above is striking but understandable, because plants are autotrophic. A noteworthy point is that another large fraction, of approximately 20–30% of the total number of genes, is involved in information transfer, signal transduction and defence. There are also many surprising revelations from the comparisons of ORF sequences with those in databases. For example, a homologue of nodulation gene – so far thought to be present only in legumes – is present even in *Arabidopsis*. Also, the *BRCA* (breast cancer) gene, thus far known only in animals, appears to have a homologue also in plants. Apparently, despite the diversity in phenotypes, the basic genetic endowment is very similar not only among plants but also between animals and plants.

Functional genomics

While it is satisfying for us to learn that in *Arabidopsis* 50–60% of the genes can be assigned some function on the basis of homology search in databases, clearly the function of the remaining 40–50% will remain unknown even after the entire genome has been sequenced. By the end of 2000, thus, we will have thousands of unknown genes to contend with and the same may be true in the case of rice as sequence data start pouring in. Thus, already, interest in many laboratories is shifting from structural to functional genomics and many strategies are being pursued towards this aim[81]. To find the function of a new gene, one can, for example, overexpress it, which can lead to accentuation of a particular phenotypic effect. One can also abolish expression by antisense DNA or by activating transposons to 'knock-out' genes[82,83]. Of late, a considerable amount of work has been done on the insertion and mobilization of the well-known transposons of maize, namely the Ac-Ds, Emu-Spm and Mu systems in model plants such as tobacco or *Arabidopsis* and to develop special strategies to identify transposon-tagged lines in large populations. Earlier, T-DNA mutagenesis had also been employed to silence genes. Transposons too, are inserted through *Agrobacterium*-mediated T-DNA transformation. However, the special advantage of transposon-based mutagenesis is that transposons can not only multiply and further move in a genome, but they can also be lost, which can aid in the confirmation of the role assigned tentatively to a gene of interest.

Function of genes can also be deduced by studying the correlation between levels of expression in a whole plant or particular organs, tissues or cells and response of plants to imposition of special environmental conditions like exposure to light, drought, heat or salt stress and this is where the new DNA chips come in. This is discussed below in some detail.

DNA chips

The great use of gene expression studies in determining the function of a gene needs no further explanation. But a serious limitation of the current technique of monitoring gene expression, namely via Northern blots, is that we can analyse expression of only one gene at a given time. But now we are poised for a new revolution, because DNA chips will allow simultaneous and *en masse* analysis of hundreds and thousands of genes[5–8,84]. In fact, it has become possible to analyse expression of all the genes of an organism such as yeast on a single chip[85]. Doubtless, it will be possible at the end of the year to do the same with *Arabidopsis* and later with rice and have an idea of entire sets of genes that may become active or inactive under a chosen experimental condition in particular tissues. This technology has been heralded variously as a 'lab-on-a-chip', 'array-of-hope' or a 'revolution on a square cm' and should give important insight not only into the number of genes involved in a developmental transition, for example, in mitosis-to-meiosis transition or resistance to a pathogen or tolerance to desiccation, but also into their function.

Fabrication of chips

To explain the basic principle of fabrication of DNA chips, it is worth reminding ourselves that for any analysis of expression, a DNA gene probe is an absolute necessity. In the ordinary Northern analysis, various types of RNA molecules remain immobilized in the gel, whereas the DNA molecules constituting the probe are initially free in a solution. Excess probe (representing DNA molecules that do not hybridize) is later washed away. In the chips, the opposite is true – the gene probes (cDNA or unique deoxyribonucleotide oligomers) remain fixed on a glass or nylon membrane support and it is the sample RNA or the cDNA which is then poured over the probe area for hybridization. Before doing so, however, the RNA or cDNA in the sample is tagged with a fluorescent dye so that laser excitation can reveal the particular genes for which transcripts have been made. Hybridization at particular sites on a gene chip is detected by scanning each spot for fluorescence.

For preparing the probes on solid supports, there are two basic procedures (Figure 2). In one procedure shown on right in Figure 2, originally developed by Davis, Brown and their colleagues at Stanford and now commercialized by Synteni/Incyte Companies in USA, cDNAs representing various genes, already isolated or the PCR products corresponding to them, serve as probes (for an example of pioneering studies with genes of humans, see ref. 86). With improved procedures, a micro-arraying robot fitted with a 16-pronged pipetting

head now allows delivery of drops as small as 1 µl, on an ordinary, but specially pre-coated microscopic glass slide. The distance between two adjacent spots can be as little as 100–150 microns, so that about 10,000 cDNAs or PCR products can be arranged in a 1.25 cm^2 area. Already an entire yeast genome has been represented on a single chip[6–8] and further refinements are being made continuously so that about 100,000 spots can be micro-arrayed, which will give us the astounding power of accomodating the entire human genome! Large-scale production of chips is being done by laying out multiple glass slides on a platform and instructing the robot to go through repetitive micro-arraying cycles.

In the alternative technology, for example, pioneered by Fodor, Lockhart and coworkers[87,88] and being commercialized by Affymetrix Company, also in USA, oligomers up to 20-mers are synthesized *in situ* in a predetermined area on the chip itself, employing a combination of organic chemistry and photolithography techniques (see Figure 2 left). Since there is no delivery of liquid drops and no problem of spreading, many more probes can be had and already techniques have been refined to the extent that as many as 400,000 different oligomer-probes can be synthesized in a 1.2 cm^2 area. At each spot, oligomers are made in a way that they represent a unique region of a particular gene.

Two-colour hybridization scheme and detection of signals

Now, to some further explanation of the procedure for detection of signals. The fluorescent labels most commonly employed to enable visualization of RNA or cDNA samples or PCR-derived copies are biotin or avidin. By using a red fluor for a control sample and a green fluor for the test sample, one can obtain a differential display of gene expression. The two samples are mixed 1:1 before the start of hybridization. Spots representing genes, whose expression is identical under both experimental and control conditions, are seen as yellow. However, genes which are over- or under-expressed reveal themselves by a range of colours varying from red to green at other spots. Fluorescence analysis and quantitation of the level of expression is made possible by newly developed fluorescence scanners in which each of the 10,000 or more spots are excited by two laser beams (one for each flour) and the output detected by photomulitpliers or CCD cameras, after light from each spot has passed through a confocal microscope-like assembly built with the scanner. The output can be digitized in a form suitable for quantitative assessment by a computer attached to the scanner. Standardization of techniques has reached a level of sophistication such that micro-arraying, hybridization as well as analysis by scanners can be completed within just one day or even an afternoon!

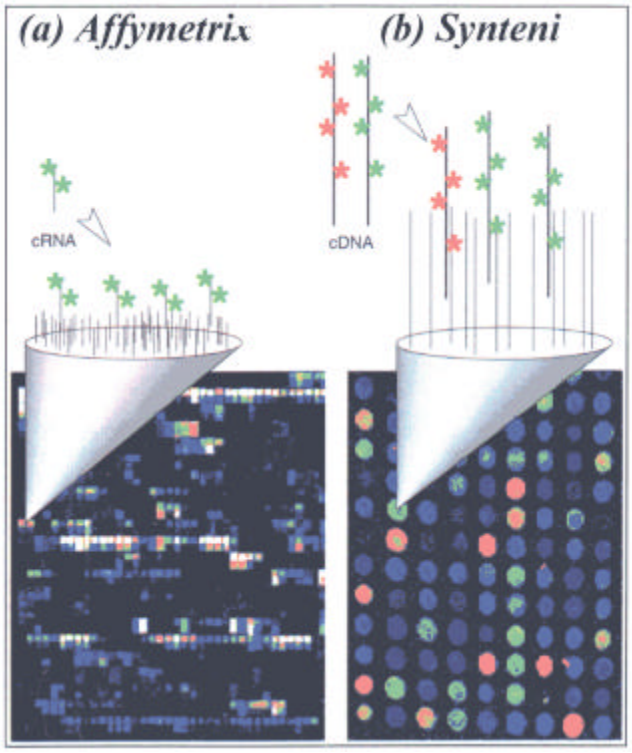

Figure 2. Diagrammatic representation of the types of DNA chips 'microarrays' that are currently available. On the left is a synthetic chip marketed by Affymetrix Company where short oligos representing unique sequences of various genes of an organism are synthesized *in situ* by a combination of photolithography and organic chemistry techniques. On the right is another kind of DNA chip, for example, one pioneered by Synteny/Incyte Company where cDNAs representing various genes of an organism are spotted on glass sides or another appropriate support by a microarrayer. Both types of chips are employed to monitor gene expression by labeling the mRNAs with fluorescent tags and allowing them to hybridize to specific spots. Hybridization is monitored by a scanner attached to a computer (adapted from Gerhold, D. *et al.*, *TIBS*, May 1999, Elsevier Science, with permission).

Not unexpectedly, there is considerable interest among plant biologists in research with DNA microchips[89–91]. The real power and use of DNA-chip technology for plants awaits, however, the complete sequencing of *Arabidopsis*, rice or other genomes and the availability of a complete set of gene probes. The practice of plant biology may then dramatically change with output of new data on an unprecedented scale. It is likely that very soon a typical researcher, for example, a graduate student, instead of working on a single gene for the period of his doctoral work, will be analysing entire clusters of hundreds and thousands of genes. That is the kind of 'scale-up' the new technology promises.

Conclusions and looking into the future

The progress of biology has often been critically dependent on introduction of novel techniques and meth-

odology. The electron microscope resulted in a major transformation of the biological scene as had also happened with the discovery and use of isotopes. Recently, cloning and sequencing techniques have brought forth the new revolution in both biology and biotechnology. Looking into the future, it seems certain that the DNA chips accomodating entire plant genomes, e.g. *Arabidopsis* and rice, would have a major impact in plant biology in the same manner as human DNA chips are already beginning to have in medicine. The chips should significantly aid in unravelling the functions of many new genes that are being discovered. Of course, parallely, several new projects are under way towards the same goal. Besides insertional mutagenesis mediated by T-DNA or transposons mentioned earlier, pioneering studies have already come up for the direct analysis of the gene products such as proteins by MALDI-TOF, which promises to be a very powerful approach for unravelling the functions of new genes[92,93]. In fact, a new area of 'proteomics' is already developing rapidly. It does not seem far-fetched that in another decade we may see the peaking of proteomics and already have an idea of the nature and function of most of the important genes and their protein products.

What next? After all the genes are discovered will that be the end of molecular biology? This question seems difficult to answer. Certainly, there is a finite number of genes and proteins in any organism and the essential genes and proteins, basic for survival, will all be known. In a conceptual sense thus, one phase of biology may indeed be over. Of course, biotechnologists may continue to tinker around the metabolic pathways for decades to come and we are likely to witness genetic engineering of crops and other plants of utility on an unprecedented scale.

However, to return to basic sciences, man will continue to be interested in his own origin, the origin of plants, the origin of a genome, the mode of diversification of plants and their evolution. One exciting finding of the limited genome sequencing efforts which has come up is that of synteny[94]. The grasses that include rice, wheat and maize, all seem to have an astonishing degree of colinearity of genes, meaning thereby the existence of an ancestral prototype genome – the diversity in genome size is largely a consequence of increase of repeated DNA. Similarly, limited mapping efforts in *Brassica* show that the basic genome organization is rather similar to that in *Arabidopsis*. Indeed, there is some degree of homology even between *Arabidopsis* and rice. Clearly, questions like how the first chromosome was organized, how genomes evolved and diversified in various phyla of dicots and monocots, and what may be the direction of evolution in future will need to be addressed and become dominant themes of future research activity. An interesting area of research will concern the genomes of chloroplasts and mitochondria. There is excellent evidence that they represent bacterial organisms that became endosymbionts and many of their genes have been moving to the nucleus of the host genome. But will more genes move to the nucleus such that as independent entities their importance may further decline? These and other questions will increasingly engage our attention in the coming years. From the Linnaean and Darwinian era of taxonomy and evolution, we moved into molecular biology in the last half of the twentieth century. But, with sequencing of several genomes being completed (a project of great relevance completed recently is the complete sequence of the alga *Synechocystis*[95]), we seem to be coming a full circle again and re-entering an era of taxonomy and evolution with far more sophisticated tools and knowledge and which will give much deeper insight than was ever possible before.

Note added in proof: The complete sequence of *Arabidopsis* genome has now been published in the December issue of *Nature*.

1. Lander, E. S., *Science*, 1996, **274**, 536–539.
2. Lander, E. S. and Weinberg, R. A., *Science*, 2000, **287**, 1777–1782.
3. Brent, R., *Cell*, 2000, **100**, 169–185.
4. Meinke, D. and Tanksley, S., *Curr. Opin. Plant Biol*., 2000, **3**, 95–142.
5. DeRisi, J. L., Iyer, V. R. and Brown, P. O., *Science*, 1997, **278**, 680–686.
6. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. and Herskowitz, I., *Science*, 1998, **282**, 699–705.
7. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D., *Proc. Natl. Acad. Sci. USA*, 1999, **95**, 14863–14868.
8. Ferea, T. and Brown, P. O., *Curr. Opin. Genet. Dev*., 1999, **9**, 715–722.
9. Sanger, F., Nicklen, S. and Coulson, A. R., *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 5463–5467.
10. Maxam, A. M. and Gilbert, W., *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 560–564.
11. Pereira, A., *Biotechnol. Dev. Monit*, 1999, **40**, 2–7.
12. Rounsley, S. and Briggs, S., *Curr. Opin. Plant Biol*., 1999, **2**, 81–82.
13. Somerville, C. and Somerville, S., *Science*, 1999, **285**, 380–383.
14. Nap, J. P. and Pereira, A., *Mol. Breed*., 1999, **5**, 481–483.
15. Watson, J. D., *Science*, 1990, **248**, 44–49.
16. Cantor, C. R., *Science*, 1990, **248**, 49–51.
17. Cooper, N. G., *The Human Genome Project – Deciphering the Blueprint of Heredity*, University Science Books, Mill Valley, California, 1994, pp. 1–360.
18. Blattner, F. R. *et al.*, *Science*, 1997, **277**, 1453–1462.
19. Dujon, B., *Trends Genet*., 1996, **12**, 263–270.
20. Oliver, S. G. *et al*., *Nature*, 1992, **357**, 38–46.
21. Goffeau, A. *et al*., *Science*, 1996, **274**, 546–563.
22. Wilson, R. K., *Trends Genet*., 1999, **15**, 51–58.
23. Sturtevant, A. H., *J. Exp. Zool*., 1913, **14**, 43.
24. Sturtevant, A. H., *A History of Genetics*, Harper and Row, New York, 1965.
25. Watson, J. D. and Crick, F. H. C., *Nature*, 1953, **171**, 737–738.
26. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. and Hood, L. E., *Nucleic Acids Res*., 1985, **13**, 2399–2412.
27. Smith, L. M. *et al.*, *Nature*, 1986, **321**, 674–679.

28. Bentley, D. R. in *Genomics* (eds Dixon, G. K., Copping, L. G. and Livingstone, D.), Bios Scientific, London, 1998.
29. Brown, T. A., *Genomes*, John Wiley (Asia), Singapore, 1999.
30. Griffiths, A. J. F., Gelbart, M., Miller, J. H. and Lewontin, R. C., *Modern Genetic Analysis*, W. H. Freeman, New York, 1999.
31. Botstein, D., White, R. L., Scolnick, M. and Davis, R. W., *Am. J. Human Genet*., 1980, **32**, 314–331.
32. Botstein, D. and Lander, E. S., *Genetics*, 1989, **121**, 185–189.
33. Collins, J. and Hohn, B., *Proc. Natl. Acad. Sci. USA*, 1978, **75**, 4242–4246.
34. Burke, D. T., Clarke, G. F. and Olson, M. V., *Science*, 1987, **236**, 806–812.
35. Olson, M. V., Hood, L., Cantor, C. and Botstein, D., *Science*, 1989, **245**, 1434–1435.
36. Green, E. D. and Olson, M. V., *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 1213–1217.
37. Adams, M. D. *et al*., *Science*, 1991, **252**, 1651–1656.
38. Coulson, A., Sulston, S., Brenner, S. and Karn, J., *Proc. Natl. Acad. Sci. USA*, 1986, **83**, 7821–7825.
39. Coulson, A. M., Waterston, J., Kiff, J., Sulston, J. and Kohara, Y., *Nature*, 1988, **335**, 184–186.
40. Watson, J. D., Gilman, M., Witkowski, J. and Zoller, M., *Recombinant DNA*, W. H. Freeman, New York, 1992, pp. 1–626.
41. Sternberg, N., *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 103–107.
42. Shizuya, H. *et al*., *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 8794–8799.
43. Ioannou, P. A. *et al*., *Nat. Genet*, 1994, **6**, 84–89.
44. Venter, J. C., Smith, H. O. and Hood, L., *Nature*, 1996, **381**, 384–386.
45. Meyerowitz, E. and Somerville, C. (eds), *Arabidopsis*, Cold Spring Harbor Laboratory Press, 1994.
46. Meinke, D. W. *et al.*, Progress Report Six, National Science Foundation, Arlington, USA, 1997.
47. Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. and Koornneef, M., *Science*, 1998, **282**, 662–682.
48. Redei, G. P., *Bibliogr. Genet.*, 1970, **20**, 1–27.
49. Koornneef, M. *et al*., *J. Hered.*, 1983, **74**, 265.
50. Hwang, I. *et al*., *Plant J*., 1991, **1**, 367–374.
51. Nam, H. G. *et al.*, *Plant Cell*, 1989, **1**, 699–715.
52. Hauge, P. M. *et al.*, *Plant J*., 1993, **3**, 745–754.
53. Liu, Y. G. *et al*., *Plant J.*, 1996, **10**, 733–736.
54. Reiter, R. S., Williams, J. G. K., Feldman, K. A., Rafalski, A., Tingey, S. V. and Scolnik, P. A., *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 1477–1481.
55. Lister, C. and Dean, C., *Plant J*., 1993, **4**, 745–750.
56. Jarvis, P., Lister, C., Sabo, V. and Dean, C., *Plant Mol. Biol*., 1994, **24**, 685–687.
57. Ward, E. R. and Jen, G. C., *Plant Mol. Biol*., 1990, **14**, 561.
58. Somerville, C. and Grill, E., *Gen. Genet*., 1991, **226**, 454–460.
59. Creusot, F. *et al*., *Plant J*., 1995, **8**, 763–770.
60. Wing, R., Choi, S., Creelman, R. A. and Mullet, J. E., *Plant Mol. Biol*., 1995, **13**, 124–128.
61. Mozo, T., Fischer, S., Meiewert, S., Lehrach, H. and Altmann, T., *Plant J*., 1998, **16**, 377–384.
62. Liu, Y-G., Mitsukawa, N., Vazquez-Telf and Whittier, R. F., *Plant J*., 1995, **7**, 351–358.
63. Bevan, M., *Bioassays*, 1999, **21**, 110–120.
64. Mozo, T. *et al*., *Nat. Genet*., 1999, **22**, 271–275.
65. Lin, X. *et al*., *Nature*, 1999, **402**, 761–769.
66. Mayer, K. *et al*., *Nature*, 1999, **402**, 769–777.
67. Meyerowitz, E., *Nature*, 1999, **402**, 731–732.
68. Marzialli, A., Willis, T. D., Federspill, N. A. and Davis, R. W., *Genome Res*., 1999, **9**, 451–462.
69. Sasaki, T., in *Molecular Biology of Rice* (ed. Shimamoto, K.), Springer-Verlag, Tokyo, 1999, pp. 3–16.
70. Sasaki, T. and Burr, B., *Curr. Opin. Plant Biol*., 2000, **3**, 138–141.
71. Sakata, K. *et al*., *Nucleic Acids Res*., 2000, **28**, 97–101.
72. McCouch, S. R. *et al*., *Theor. Appl. Genet*., 1988, **76**, 315–329.
73. Causse, M. A. *et al*., *Genetics*, 1994, **138**, 11251–11274.
74. Kurata *et al*., *Nat. Genet*., 1994, **8**, 365–372.
75. Harushima, Y. *et al*., *Genetics*, 1998, **149**, 1–16.
76. Sasaki, T. *et al*., *Plant J*., 1994, **6**, 615–624.
77. Yamamoto, K. and Sasaki, T., *Plant Mol. Biol*., 1997, **35**, 135–144.
78. Zhang, H. B. and Wing, R. A., *Plant Mol. Biol*., 1997, **35**, 115–127.
79. Kurata, N., Umehara, Y., Tanoue, H. and Sasaki, T., *Plant Mol. Biol*., 1997, **35**, 101–113.
80. Pennisi, E., *Science*, 2000, **288**, 239–240.
81. Bouchez, D. and Höfte, H., *Plant Physiol*., 1998, **118**, 725–732.
82. Martienssen, R. A., *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 2021–2026.
83. Sundaresan, V., *Trends Plant Sci*., 1996, **1**, 184–190.
84. Schena, M. (ed.), *DNA Microarrays: A Practical Approach*, Oxford University Press, Oxford, 1999, pp. 1–210.
85. DeRisi, J. L., Iyer, V. R. and Brown, P. O., *Science*, 1997, **278**, 680–686.
86. Heller, R. A. *et al*., *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 2150–2155.
87. Chee, M. *et al*., *Science*, 1996, **274**, 610–611.
88. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J., *Nat. Genet. Suppl.*, 1999, **21**, 20–24.
89. Richmond, T. and Somerville, S., *Curr. Opin. Plant Biol*., 2000, **3**, 108–116.
90. Lemieux, B., Aharoni, A. and Schena, M., *Mol. Breed*., 1998, **4**, 277–289.
91. Kehoe, D., Villand, P. and Somerville, S., *Trends Plant Sci*., 1999, **4**, 38–41.
92. Blackstock, W. P. and Weir, M. P., *Tibtech*., 1999, **17**, 121–127.
93. Smith, Harry B., *Plant Cell*, 2000, **12**, 303–304.
94. Devos, K. M. and Gale, M. D., *Plant Mol. Biol*., 1997, **35**, 3–15.
95. Kaneko *et al*., *DNA Res*., 1996, **3**, 109–136.

ACKNOWLEDGEMENTS. We express our sincere gratitude to Drs S.K. Mukherjee and Suresh Nair at the ICGEB, and Professors A.K. Tyagi and J.P. Khurana, Department of Plant Molecular Biology, Delhi University South Campus, for many helpful suggestions to improve the manuscript. Thanks are due also to Mrs Shashi Mehta and Mr Satyendra Patwal, Plant Molecular Biology Department, Delhi University South Campus, for their invaluable help in finalising this manuscript.