PLANT GENOMICS AND PROTEOMICS

CHRISTOPHER A. CULLIS

WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

Plant Genomics and Proteomics

PLANT GENOMICS AND PROTEOMICS

CHRISTOPHER A. CULLIS

WILEY-LISS

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:
Cullis, Christopher A., 1945–
Plant genomics and proteomics / Christopher A. Cullis.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-37314-1
1. Plant genomes. 2. Plant proteomics. I. Title.
QK981.C85 2004
572.8′62—dc21
2003013088

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments

This book would not have been possible without the contributions of two individuals. First, I would like to thank my wife Margaret, whose efforts in reading the drafts and suggesting clarifications were invaluable. Any obscure or erroneous passages are certainly not her responsibility; she probably just could not get me to change my mind. Second, I would like to thank my son Oliver, with whom I shared the first attempts at writing a book and who contributed with comments on the clarity of early drafts.

Introduction

What possible rationale is there for developing a genomics text that is focused on only the plant kingdom? Clearly, there are major differences between plants and animals in many of their fundamental characteristics. Plants are usually unable to move, they can be extremely long lived, and they are generally autotrophic and so need only minerals, light, water, and air to grow. Thus the genome must encode the enzymes that support the whole range of necessary metabolic processes including photosynthesis, respiration, intermediary metabolism, mineral acquisition, and the synthesis of fatty acids, lipids, amino acids, nucleotides, and cofactors, many of which are acquired by animals through their diet. At a technological level genomics studies, which take a global view of the genomic information and how it is used to define the form and function of an organism, have a common thread that can be applied to almost any system. However, plants have processes of particular interest and pose specific problems that cannot be investigated in any one simple model and often even need to be investigated in a particular plant species. Plant genomics builds on centuries of observations and experiments for many plant processes. Because of this history, much of the experimental detail and observations span very diverse plant material, rather than all being available in a convenient single model organism. Thus algae may be appropriate models for photosynthesis and provide useful pointers as to which genes are involved but, conversely, cannot be useful for understanding, for example, how stresses in the roots might affect the same photosynthetic processes in a plant growing under drought or saline conditions. The genomics approaches to plant biology will result in an enhanced knowledge of gene structure, function, and variability in plants. The application of this new knowledge will lead to new methods of improving crop production, which are necessary to meet the challenge of sustaining our food supply in the future.

One of the particularly relevant differences, for this text, between plants and other groups of organisms is the large range of nuclear DNA contents

(genome sizes) that occur in the plant kingdom, even between closely related species. Therefore, it is harder to define the nature of a typical plant genome because the contribution of additional DNA may have phenotypic effects independent of the actual sequences of DNA present, for example, the role of nuclear DNA content in the annual versus perennial life cycle. An added complication is that rounds of polyploidization followed by a restructuring of a polyploid genome have frequently occurred during evolution. The restructuring of the genome has usually resulted in a loss of some of the additional DNA derived from the original polyploid event. Therefore, the detailed characterization of a number of plant genomes, rather than a single model or small number of models, will be important in developing an understanding of the functional and evolutionary constraints on genome size in plants. Despite this enormous variation in DNA content per cell, it is generally accepted that most plants have about the same number of genes and a similar genetic blueprint controlling growth and development.

As indicated in the opening paragraph, the wealth of data for many processes, such as cell wall synthesis, photosynthesis and disease resistance, has been generated by investigating the most amenable systems for understanding that particular process. However, many of these models are not well characterized in other respects and have relatively few genomics resources, such as sequence data and extensive mutant collections, associated with them. Therefore, the information derived from each of these systems will have to be confirmed in a well characterized model plant to understand the molecular integration and coordination of development for many of the intertwined pathways. This may not be possible in the best-characterized systems of each of the individual elements. *Zinnia* provides an excellent model to study the differentiation of tracheary elements because isolated mesophyll cells can be synchronously induced to form these elements *in vitro*. Therefore, this synchrony permits the establishment and chronology of the molecular and biochemical events associated with the differentiation of the cells to a specific fate and the identification of the genes involved in the differentiation of xylem. However, *Zinnia* does not have the experimental infrastructure to allow extensive genomic investigations into other important processes. Therefore, the detailed knowledge acquired would need to be integrated in another more fully described model plant, although the knowledge would have been difficult to identify without resource to this specialized experimental system. Therefore, the accumulation of genomic information will be necessary across the plant kingdom, with an integrated synthesis perhaps finally occurring only in a few model species. The relevant approaches will include the development of detailed molecular descriptions of the myriad of plant pathways for many plant species in order to unravel the secrets of how plants grow, develop, reproduce, and interact with their environments.

The publication of the *Arabidopsis* and rice genomic sequences has

facilitated the comparison between plants and animals at the sequence level. Not surprisingly, perhaps, the initial comparisons have shown that some processes, such as transport across membranes and DNA recombination and repair processes, appear to be conserved across the kingdoms whereas others are greatly diverged. Many novel genes have been found in the plant genomes so far characterized, which was expected considering the wide range of functions that occur in plants but are absent from animals and microbes.

The easy access to plant genome sequences and all of the other genomics tools, such as tagged mutant collections, microarrays, and proteomics techniques, has fundamentally changed the way in which plant science can be done. Old problems that appeared to be intractable can now be tackled with renewed vigor and enthusiasm. One example is the Floral Genome Project (http://128.118.180.140/fgp/home.html) tackling what Darwin referred to as "The abominable mystery," namely, the origin of flowering plants, that has gone unanswered for more than a century. More than just answering this question, though, the origin and diversification of the flower is a fundamental problem in plant biology. The structure of flowers has major evolutionary and economic impacts because of their importance in plant reproduction and agriculture.

The two different regions of the plant, the aerial portions (stems, leaves, and flowers) and the below-ground portions (roots), have received very different treatment as far as experimental investigations are concerned. The above-ground regions of the plant have clearly been more amenable to visual description and biochemical characterization. This is partly due to the difficulty in studying the roots. Not only are they normally in a nonsterile environment, beset with many microorganisms both beneficial and harmful, but they are also difficult to separate from the physical medium of the soil. As genomic tools continue to be developed it will become easier to delineate the contribution and characteristics of the associated microorganisms and the plant roots and so understand the interaction of the roots and the microenvironment in the soil. Of particular interest is the understanding of the beneficial interactions between the plant roots and microorganisms such as rhizobia and mycorrhizae, in contrast to the destructive interactions between the roots and pathogens.

The interface between the plant and pathogens is also important with respect to the aerial portions of a plant. The combination of an increased understanding of the pathogen's genome, as well as the responses that occur in both the pathogen and the host on infection, will open up new methods for controlling diseases in crops. The detailed understanding of the interplay between the plant and the pathogen should also enable the development and incorporation of more durable resistances to many of the destructive plant diseases, resulting in an increased security of the food supply worldwide. Therefore, these new interventions, supported by information from

genomics studies, will be important both for increasing yield and for reducing environmental hazards that may be associated with the current agronomic use of available fungicides and insecticides.

Light, as well as being the primary energy source for plants, also acts as a regulator of many developmental processes. Chlorophyll synthesis and the induction of many nucleus- and chloroplast-encoded genes are affected by both light quality and quantity. In this respect the close coupling of the nuclear and chloroplast genomes is another unique plant process. Many of the biochemical reactions of light responses have already been well documented, but the ability to recognize the genes that have been transferred from the organellar genomes to the nucleus may also shed light both on the coordinated control of these responses and on the evolutionary history, pressures, and constraints. Again, the input from the characterization of the genomes of algae and other microorganisms will greatly facilitate all such studies.

The synthesis of cell walls and their subsequent modification are clearly important processes in higher plants. The initial annotation of the *Arabidopsis* genome identified more than 420 genes that could tentatively be assigned roles in the pathways responsible for the synthesis and modification of cell wall polymers. The fact that many of these genes belong to families of structurally related enzymes is also an indication of the apparent gene redundancy in the plant genome. However, as will be discussed in this work, whether this redundancy is real, in the sense that one member of the family can effectively substitute for any of the other members, or whether this is only an apparent redundancy and the various genes reflect differences in substrate specificity or developmental stage at which they function, is still to be determined.

Plants synthesize a dazzling array of secondary metabolites. More than a hundred thousand of these are made across all species. The exact nature and function of most of these metabolites still await understanding. The combination of information from sequencing, expression profiling, and metabolic profiling will help to define the relationship between the genes involved, their expression, and the synthesis of these metabolites. The understanding of which member of a gene family is expressed in a particular tissue, and the specific reaction in which it is involved, will also shed light on the level of redundancy of gene functions for the synthesis of many of these compounds.

Many of the processes that are known to regulate or control development in animals including the modulation of chromatin structure, the cascades of transcription factors, and cell-to-cell communications, will also be expected to regulate plant development. However, the initial analysis of the *Arabidopsis* genome sequence indicates that plants and animals have not evolved by elaborating the same general process since separation from the last common ancestor. For example, although plants and animals have

comparable processes of pattern formation and the underlying genes appear to be similar, the actual mechanisms of getting to the end points of development are different. Once again, this reinforces the need to look specifically at the plant processes in order to understand how plants function.

One of the important ways in which the whole genome approach has changed plant biology is that international cooperation in many of the major projects is both necessary and important. The funding required for large-scale genomic sequencing makes it more important than ever to avoid unnecessary duplication. Thus the international coordination of both the *Arabidopsis* and the rice genome projects has ensured their completion with the minimal overlap of expenditure from the various international members, while still generating the appropriate scientific infrastructure and, in some cases, being responsible for the development of additional human and technological resources. These collaborations, both international as well as national, have improved the infrastructure for the science as well as moving knowledge forward at an ever-increasing rate.

The other important aspect of these genomics investigations is that the results are generally being widely disseminated, especially through Internet resources. Therefore, the constituency that is able to use these results to build detailed knowledge in specialist areas is ever widening. The structure of the informatics resources and the tools to query them must be compatible with the wide range of expertise of the interested parties. For individual investigators to be able to access and interrogate the results of major resource generators, such as sequencing projects, mutant collections, and the like, the data and resources must be made available. The availability of these resources is not just limited to the time that they are being actively generated but also after these projects are completed. Therefore, the archiving of biological and informatics resources to ensure their continued availability is vital, considering the investment that is being made in their generation.

The application of all this knowledge to the improvement of crops is not without controversy. The ability to manipulate plants for specific purposes with the introduction of new genetic material, that may or may not be of plant origin, is viewed with varying degrees of concern across the world. It is undoubtedly true that all of this new information can be useful in the development of new varieties by traditional breeding, but it will also have an input in developing totally novel strategies, including the use of plants to produce new raw materials. It will be important that the benefits of such engineered resources are spread across society and throughout the world to benefit both developed and developing countries, or they will never be generally accepted.

The primary aim of this text is to introduce the reader to the range of molecular techniques that can be applied to the investigation of unique and interesting facets of plant growth, development, and responses to the environment. The rapid progress made in this area has clearly been as a result

of increased funding in both the private and public sectors. The public sector efforts in the USA have been stimulated and supported by the National Plant Genome Initiative formally organized in 1997, along with major investments worldwide. This kind of support will be necessary for years to come to manipulate crop plants for improved productivity and ensure food security. The end result of all this investment should be a quicker introduction of new crop varieties in response to particular needs. The understanding of disease resistance, for example, and the development of new approaches to this problem are expected to reduce the time for new resistant varieties to be developed compared with the conventional introgression of new resistance genes from wild relatives. The combination of resources and technology that are currently available makes this an incredibly exciting time to be involved in plant genomics.

CHAPTER

1

THE STRUCTURE OF PLANT GENOMES

There is probably no one example that can be considered as the typical plant genome. They come in an amazing variety of shapes and sizes if one considers that the packaging into chromosomes is a form of shape. This variety can exist even within a family, with the result that plants are much more variable than any other group of organisms as far as these nuclear characteristics are concerned. In this chapter we consider how variable the DNA quantity can be, the variety of chromosome structures, and how all this variability in DNA quantity and packaging arose. These factors impinge on the design, feasibility, and interpretation of genomics studies.

DNA VARIATION—QUANTITY

The characteristic nuclear DNA value in a plant is generally expressed as the amount contained in the nucleus of a gamete (the 1C value), irrespective of whether the plant is a normal diploid or a polyploid (either recent or ancient). The use of a standard tissue is important because the nuclear DNA content can vary among tissues with some, for example the cotyledons of peas, having cells that have undergone many rounds of endoreduplication (Cullis and Davies, 1975). Nuclear DNA values have been reported in two different ways, either as a mass of DNA in picograms per 1C nucleus or as the number of megabase pairs of DNA per 1C nucleus. The relationship between these two ways is relatively easy to estimate because 1 pg of DNA is approximately equal to 1000 Mbp (the actual conversion is 1 pg ≡ 980 Mbp).

Plant Genomics and Proteomics, by Christopher A. Cullis
ISBN 0-471-37314-1

This 1C value for the amount of DNA in a plant nucleus can vary enormously. For example, one of the smallest genomes belongs to *Arabidopsis thaliana*, with 125 Mbp, whereas the largest reported to date belongs to *Fritillaria assyriaca*, with 124,852 Mbp, equivalent to 127.4 pg. This represents a 1000-fold difference in size between the largest and smallest genomes characterized so far. Some representatives that span these extremes are included in Table 1.1 and are taken from the database maintained by the Royal Botanic Gardens, Kew (http://www.rbgkew.org.uk/cval/homepage.html).

However, this range may not represent the true limits because DNA values have been estimated in representatives of only about 32% of angiosperm families (but only representing about 1% of angiosperm species), 16% of gymnosperm species, and less than 1% of pteridophytes and bryophytes. This variation occurs not only between genera but also within a genus. One example is the genus *Rosa*, in which there is a more than 11-fold variation in genome size. The fact that this range in DNA content is not associated with variation in the basic number of genes required for growth and development has led to its being referred to as the C-value paradox.

Genome size is an important biodiversity character that can also have practical implications. One example is that the genome size seems to constrain life cycle possibilities, in that all of those plants that have above a certain DNA content are obligate perennials (Bennett, 1972). Another example is that species with large amounts of DNA (>20 pg per 1C) can be problematic when studying genetic diversity with standard amplified fragment length polymorphism (AFLP) techniques such as have been encountered with *Cypripedium calceolus* (1C = 32.4 pg) and *Pinus pinaster*

TABLE 1.1. Selected DNA Values

Genus	Species	1C pg
Cardamine	*amara*	0.06
Arabidopsis	*thaliana*	0.125
Rosa	*wichuraiana*	0.13
Luzula	*pilosa*	0.28
Oryza	*sativa*	0.5
Rosa	*moyesii*	1.45
Gnetum	*ula*	2.25
Zea	*mays*	2.73
Nicotiana	*tobaccum*	5.85
Ginkgo	*biloba*	9.95
Allium	*sativum*	16.23
Pinus	*ponderosa*	24.2
Fritillaria	*assyriaca*	127.4

From http://www.rbgkew.org.uk/cval/homepage.html

(1C = 24 pg) (cited in Bennett et al., 2000). On the other hand, a very small DNA content has been a major factor in determining the early candidates for genome sequencing. Consequently, *Arabidopsis thaliana* (a dicot) was the first plant chosen for genome sequencing, partly because it had one of the smallest C values known for angiosperms. Rice was the second genome sequenced and was the first monocot chosen because it had the smallest C value among the world's major cereal crops, even though it did not have the smallest genome in the grasses. This distinction currently goes to the diploid *Brachypodium distachyon*, which has a 1C value of 0.25–0.3 pg, whereas the rice genome is nearly twice this size (Bennett et al., 2000).

The determination of the genome sequence of *Arabidopsis* gives some indication of what the minimum genome size for a higher plant is likely to be. The extensive duplication that was found in the *A. thaliana* genome could well have been the result of polyploidy earlier in the evolutionary history of this plant. Thus the number of genes necessary and sufficient to determine a functional higher plant is likely to be somewhat less than 25,000, the current estimate for *A. thaliana*. Additional DNA will need to be associated with these genes to ensure appropriate chromosome function by defining the centromeres and telomeres. Therefore, the most stripped-down plant genome is unlikely to be much below 0.1 Gb, because in addition to the 25,000 genes, DNA associated with centromeres and telomeres that ensure chromosome stability and segregation at cell division will also have to be included. However, a great deal more information is still required before a conclusion that this minimal number will be sufficient to ensure the full range of functions that can be performed by plants.

As will be seen below the actual amount of DNA that is associated with various structures within the genome can vary. However, it is not just in this context that it is important to know the C value. DNA amounts have been shown to correlate with various plant life histories, the geographic distribution of crop plants, plant phenology, biomass, and sensitivity of growth to environmental variables such as temperature and frost. The C value may also be a predictor of the responses of vegetation to man-made catastrophes such as nuclear incidents. It has been shown that plants with a higher DNA content and particular chromosome structures are more resistant to radiation damage (Grime, 1986).

Chromosome Variation

Chromosome number and size are very variable. The stonecrop, *Sedum suaveolens*, has the highest chromosome number (2n of about 640), whereas the lowest chromosome number is that of *Haplopappus gracilis* (2n = 4). Ferns also have extremely high values. An increase in the number of chromosomes is usually associated with a reduction in chromosome size. The actual

structure of a chromosome can also vary, with most species having the usual chromosome structure of a single centromere. However, some plants have holocentric chromosomes where kinetochore activity (regions that attach to the spindle at mitosis and meiosis) is present at a number of places all along the chromosome.

In the genus *Luzula*, which has holocentric chromosomes, the chromosome number can also vary widely, with *L. pilosa* having 66 chromosomes and *L. elegans* having 6 as the diploid number (Figure 1.1a, b). As can be seen in the figure, the size of a chromosome in these two species is very different. The quantity of DNA in each chromosome is also very different; *L. elegans* has 3 chromosomes in which to package the 1446 Mbp of DNA in the 1C nucleus, whereas in *L. pilosa* 33 chromosomes are available for only 270 Mbp of DNA in the 1C nucleus. Each of the *L. elegans* chromosomes is of similar size and contains an average of 482 Mbp of DNA, whereas each *L. pilosa* chromosome only packages about 8 Mbp of DNA. Therefore, within this genus, a single chromosome of one species (*L. elegans*) contains an amount of DNA equivalent to that present in the complete rice genome, whereas the other (*L. pilosa*) has chromosomes that are each the size of an average microbial genome.

The arrangement of kinetochore activity all along the chromosome has consequences for meiosis, including a restriction of the reduction division to the second division of meiosis rather than the first, as is the case in most plants. It also restricts the regions that can recombine and so may have other consequences for the plants that must be considered in relation to function and evolution of the genome. However, it does mean that almost any chromosome fragment will have a kinetochore and so be maintained through cell division. Therefore, fragmentation of the chromosomes will not be lethal and can generate different chromosome numbers. The organization of the genome into this type of package leads to extreme resistance to radiation damage. Figure 1.2 shows mitosis from a callus cell of *L. elegans*. Although the plants were grown from irradiated seeds they showed no apparent phenotypic abnormalities. In fact, plants are very tolerant of chromosome aberrations, with ploidy changes being very frequent. This property can be utilized in generating material that is targeted to understanding of particular regions of the genome, for example, the production of wheat addition and deletion lines that have been important resources in the effort to unravel the enormous wheat genome (Sears, 1954) and for isolating single maize chromosomes (Kynast et al., 2001).

As mentioned above for the genus *Luzula*, the chromosomes can vary greatly both in size and number. Situations also exist in which there is relatively little difference in the chromosome number but there are very large differences in the chromosome sizes. Within the legumes this has been extensively characterized. For example, both *Vicia faba* and *Lotus tenuis* have a chromosome number of 6, whereas the lengths of these

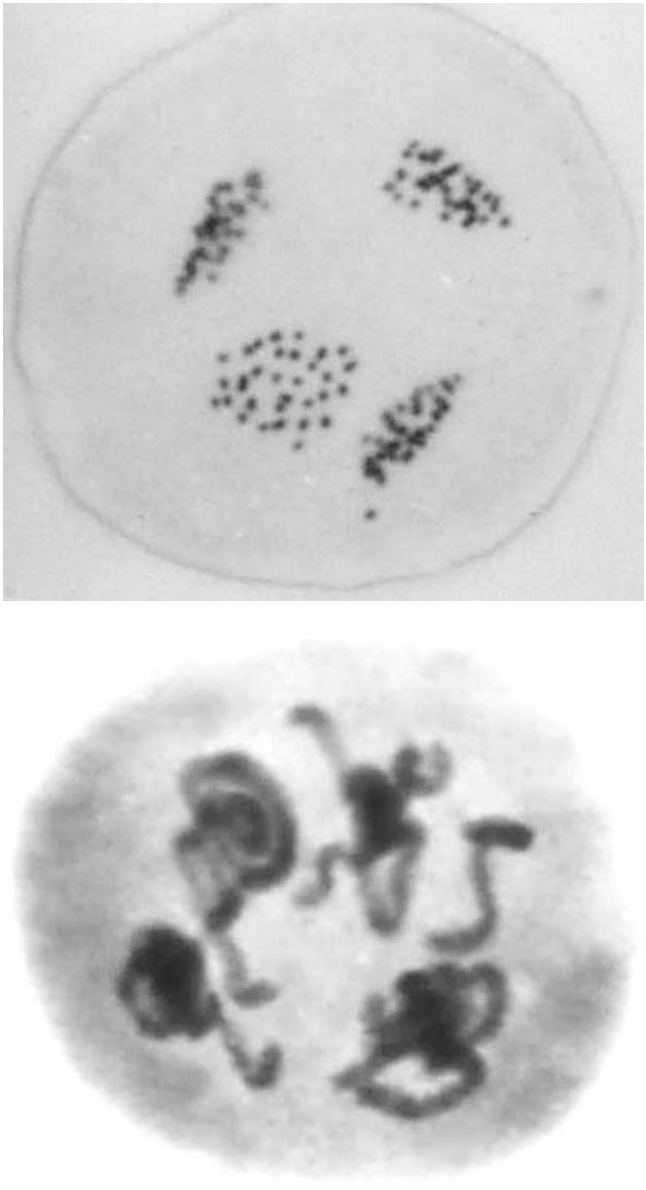

FIGURE 1.1. Metaphase of meiosis II in *L. pilosa* (a) and *L. elegans* (b). (Photographs by Dr. G. Creissen.)

FIGURE 1.2. Mitotic metaphase in *L. elegans* callus derived from seed irradiated with 80 krad. At least 3 centric fragments are visible in addition to the 6 chromosomes. (Photograph by Dr. B. Bowen.)

TABLE 1.2. Chromosome Number, Chromosome Length, and DNA Content of Two Legumes

Species	Haploid set of chromosomes (n)	Average length of chromosomes (μm)	Nuclear DNA content (pg)
Lotus tenuis	6	1.8	0.48
Vicia faba	6	14.8	13.33

From http://www.biologie.uni-hamburg.de/b-online/e37/37c.htm

chromosomes only partly reflect the differing DNA contents in these two species (Table 1.2, Figure 1.3), with the DNA per unit length differing over threefold (0.044 pg/μm in *Lotus* and 0.15 pg/μm in *Vicia*) (from http://www.biologie.uni-hamburg.de/b-online/e37/37c.htm).

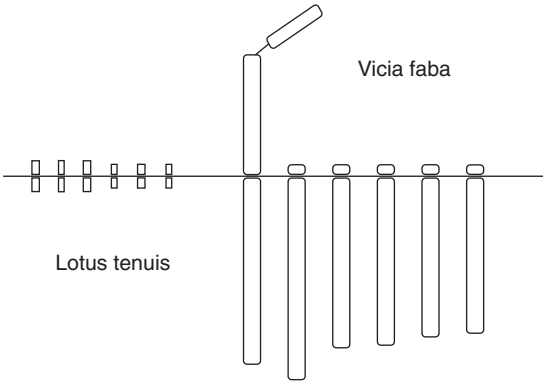

FIGURE 1.3. Chromosome sizes in *Lotus* and *Vicia*.
(From http://www.biologie.uni-hamburg.de/b-online/e37/37c.htm.)

ORIGIN OF DNA VARIATION

The sequences in the genome are generally classified with respect to the number of times they are represented. The three main classes to which they are assigned, low copy, moderately repetitive, or highly repetitive, have somewhat arbitrary cutoffs, with both copy number and function playing a part in the classification. These three classes and some of their characteristics are:

- Low-copy-number or unique sequences that probably represent the genes
- Moderately repetitive sequences, many of which may be members of transposable element families that are distributed around the genome
- Highly repetitive sequences, many of which are arranged in tandem arrays

The arrangement of these sequences with respect to one another has functional consequences for the plant.

LOW-COPY SEQUENCES

The two complete genome sequences from *Arabidopsis thaliana* and rice are from genomes that vary nearly fourfold in size, so the estimates of gene number from these two sequences will go some way toward establishing how the gene number might change with genome size. The initial estimates from the rice genome sequence (Goff et al., 2002) are that rice has about twice the number of genes that are found in *Arabidopsis*. As gene finding programs