Genetic Engineering 6

edited by
Peter W. J. Rigby

Genetic Engineering 6

Edited by

Peter W. J. Rigby

Laboratory of Eukaryotic Molecular Genetics,
National Institute for Medical Research,
The Ridgeway, Mill Hill, London

ACADEMIC PRESS · 1987

Harcourt Brace Jovanovich, Publishers
London · Orlando
San Diego · New York · Austin
Boston · Sydney · Tokyo · Toronto

ACADEMIC PRESS LIMITED
24/28 Oval Road,
London NW1

United States Edition published by
ACADEMIC PRESS INC.
Orlando, Florida 32887

British Library cataloguing in Publication Data

Genetic engineering.—1987
1. Genetic engineering—Periodicals
575.1 QH442

ISBN 0-12-270306-5

Printed in Northern Ireland at The Universities Press (Belfast) Ltd.

Preface

One of the greatest benefits to ensue from the development of techniques for manipulating genes has undoubtedly been the opening up of areas of biology in which studies at the molecular level were slow, difficult or even impossible. Perhaps nowhere has this been seen to greater effect than in the application of molecular genetic techniques to plant science and plant breeding. Bob Williamson and I, in prefaces to the two preceeding issues of this series, wrote that we hoped to see a volume devoted to plant molecular biology, and I am therefore very pleased that this aim has been achieved.

While plant science has made important contributions to the development of our understanding of the molecular biology of genes, the discovery of transposable elements being the obvious example, the application of the techniques of molecular genetics to plants has led to enormous advances in biological knowledge, to the mobilization of much that was already known and, with great rapidity, to the improvement of crops of agronomic importance. In much of this work plant scientists have exploited a very great advantage they have over other eukaryotic molecular biologists: the ability to regenerate fertile organisms from tissue culture cells.

In Chapter 1, Joachim Messing discusses the genes which encode the major storage proteins of plant seeds. These genes, particularly those which encode the zeins of maize, have served as an important model for studies of plant gene structure and function. Moreover, these proteins determine the amino acid content of seed, and our ability to manipulate the genes which encode them has many implications for the improvement of the nutritional value of crop plants.

A striking feature of plant cells is that their cytoplasms contain multiple autonomous genetic systems and studying them, particularly those of the mitochondrion and chloroplast, has been revolutionized by the application of cloning techniques. David Lonsdale reviews our knowledge of the cytoplasmic molecular genetics of plants, and points out how these systems have enormous general value for the study of nuclear–organelle interactions. Moreover, cytoplasmic characters can affect features of agronomic importance and so their manipulation will also contribute to the development of improved crop varieties.

In the final chapter, Conrad Lichtenstein and Sheryl Fuller describe the vectors that have been developed for the genetic engineering of plants. These systems are very powerful and, when they are combined with the ability to regenerate plants from cultured cells, they offer the prospect not only of fascinating science but economic benefit as well.

I hope, firstly, that this volume will provide a valuable compendium of information for those already active in the field; secondly, that it will provide a primer for plant scientists who wish to apply the techniques of molecular biology to their work; and thirdly, and perhaps most importantly, that it will enable those who work with mammalian systems to gain an insight into the enormous excitement in, and the value of, plant molecular biology.

I wish to thank the authors for the enthusiasm and effort that they have brought to this volume and the staff of Academic Press for all of their help.

London, July 1987 *Peter W. J. Rigby*

Contents

The genes encoding seed storage proteins in higher plants

Joachim Messing

The molecular biology and genetic manipulation of the cytoplasm of higher plants

David M. Lonsdale

Vectors for the genetic engineering of plants

Conrad P. Lichtenstein and Sheryl L. Fuller

The genes encoding seed storage proteins in higher plants

JOACHIM MESSING

Waksman Institute, Rutgers, The State University,
Piscataway, New Jersey 08845, USA

GENETIC ENGINEERING Vol. 6
ISBN 0-12-270306-5

I Introduction

Although we do not yet understand the mechanism of the regulation of gene expression, more and more information is accumulating about the general features of the organization of genes in prokaryotic and eukaryotic organisms. In particular, the use of DNA sequencing methods, of site-directed mutagenesis (for review, see Messing, 1983a), and of foot printing methods (Church and Gilbert, 1984) have advanced our knowledge of the *cis*-acting sequences that are recognized by transcription factors (Dynan and Tjian, 1985).

These lines of study have far-reaching implications, not only for medical care, but also for the world food supply. Although the yield of crops per acre has risen substantially because of the development of superior seed material due to the breeding efforts in research institutions and seed companies, the introduction of highly efficient farm machinery and the use of chemicals as fertilizers and herbicides, the exploration of the genetic potential of the plant kingdom for our benefit has hardly begun. One example that shows us how the control of gene action has a direct agronomic impact is derived from the cost of feed in poultry and hog production.

These animals need a balanced composition of amino acids in their nutrients. Although the corn kernel is high in carbohydrates and protein, the latter is low in tryptophan and lysine, two essential amino acids. Therefore, a diet is used as feed that compensates for the lack of these amino acids by adding soybean meal to the ration. However, both corn and soybean are low in methionine, and the diet has to be supplemented with free methionine. The latter supplement is the more costly since it is produced by fermentation. On the other hand, the proportions of amino acids in the animal feed are not determined by the free amino acid pools, but are largely dependent on the amino acid composition of the major proteins in the seed of these plants. The hydrolysis of these proteins in the gut is the major route of amino acid production for the nutrition of these animals. Therefore the nutritional quality of animal feed is determined by the control of the expression of the genes encoding the major seed proteins, and not by that of the amino acid biosynthetic pathways.

These major seed proteins are a protein family—also called storage proteins—the expression of which is highly regulated during seed development. Because of this regulation and the availability of a range of variants that modulate the synthesis of the various subclasses of this protein family, the corresponding genes represent an interesting model for nuclear gene structure and function in higher plants. Since they have been one of the first plant multigene families to be studied at the DNA level, they are also one of the most thoroughly investigated. Although more and more is being learned about other nuclear genes that are not only developmentally controlled but induced by other stimuli like stress and light, it is now clear that the storage protein genes have played an important role in the development of plant molecular biology and they represent a suitable focus for discussing our present knowledge of plant gene structure and function.

The example above has shown us a direct link between the study of gene expression in plants and the importance of manipulating the control of gene expression for agronomic use. Because of the universality of some of the principles in genetic switches, observations made in one system may provide insight into others. From the analysis of the 18S ribosomal DNA sequence of maize (Messing *et al.*, 1983) it appears that in many regions of the ribosomal RNA plants share more homology to animal organisms than to simpler eukaryotic organisms like yeast. Therefore, the analysis of plant genomes should be of value not only in terms of such practical consequences as crop improvement, but also in increasing our understanding of the biology of eukaryotic organisms.

II The plant genome

The size of the plant genome is quite variable (Bennet and Smith, 1976). For example, maize has 10 chromosomes and a genome size of about 5×10^9 bp, while *Arabidopsis*, a plant of the mustard family, has only 5 chromosomes and a genome which is about 70-fold smaller. Although it is unclear what role the size of the genome plays in gene expression, one reason for the difference seems to be the lower content of repetitive DNA; moreover, the genes are organized in smaller multigene families than in plants with a more complex genome (for review, see Sommerville *et al.*, 1985).

Whatever the physiological role of the variation in genome size within the plant kingdom, there are practical implications for the study of gene structure. A smaller genome can be handled more conveniently for the isolation of genes and their mutant derivatives. An heterologous probe can be used to screen firstly a genomic library of *Arabidopsis*, and the *Arabidopsis* probe can then be used to screen a more complex plant genomic library, usually from a plant of agronomic value. Furthermore, the genomic organization in

Arabidopsis can be compared to the organization of more complex genomes and much will be learned from these differences. The latter point may be more important than the convenience of a smaller genomic library, since the larger genomic library does not have to be complete in order to isolate nuclear genes. For example, while *Arabidopsis* seems to have only a few copies of storage protein genes, maize appears to have more than a hundred copies of these genes. However, it will be desirable to utilize the complexity of the larger genomes and to obtain genomic clones of more than one member of a multigene family, since natural allelic variations may occur more frequently in the complex systems. These allelic variations are important resources of material that can be used to study the regulation of gene expression.

When I screened the GenBank database for nucleotide sequences from higher plants several years ago, they represented only a fraction of the mammalian database or of the total database (Messing, 1984). In the meantime, the information on plant sequences has increased rapidly and the gap between sequences that have been entered into the central data depository and those that have been published is growing. If we consider the information that is available, a great deal of attention has been given to chloroplast biology because of the uniqueness of photosynthesis in plants versus animals. Since the genome organization of the chloroplast is different from that of the nucleus, it contributes much less to the understanding of nuclear genes. On the other hand, a few nuclear genes have been isolated that play a role in plastid development and function. The most notable of these are the genes which encode the small subunit of ribulose bisphosphate carboxylase that is imported from the cytoplasm into the chloroplast by the interaction of a transit peptide sequence with a receptor in the chloroplast membrane (Berry-Lowe *et al.*, 1982; Coruzzi *et al.*, 1983, 1984; Mazur and Chui, 1985), and the glycosyl transferase that is imported into amyloplasts (Shure *et al.*, 1983; Klösgen *et al.*, 1986). Studies of the very complex plant mitochondrial DNA also allow us to predict that many mitochondrial proteins, like the ATP synthase (Boutry and Chua, 1985), are encoded by the nuclear genome (Lonsdale, this volume).

However, many more plant nuclear genes that are important to seed development have been studied than have those that are involved in nuclear–organelle interactions (Heidecker and Messing, 1986). Therefore, much more information about plant gene structure is available for this class of gene. Before we investigate the information derived from their structure, a short discussion of seed development will be useful.

III Seed development

Although seed development and morphology vary significantly within the plant kingdom, it may be worthwhile for the subsequent discussion to focus on the reproduction of maize as an example. Maize is not only one of our most important food crops, but also one of the genetically best studied plants. Its usefulness as a genetic model organism will become particularly clear from its form of seed development.

The male and female reproductive parts are both found on the same plant. Therefore, not only reciprocal crosses but also self-pollination can be carried out. The male reproductive part consists of the tassels that grow on the top of the plant; anthers, that contain the pollen grain, develop from the tassels. The female reproductive part consists of the ears that develop at a lower point on the stem or stalk. Each ear develops hundreds of ovaries and from the cell wall of the ovary grows a long style, the silk, with numerous protruding hairs. Before we discuss the actual pollination and fertilization events, let us consider the developmental stages prior to pollination. The florets on an immature ear are arranged in the same order as the kernels of the mature ear. Therefore, the ear can be compared to a Petri dish that contains hundreds of colonies. Each floret will develop into a new progeny with a distinct genetic character. Since some plants develop side stalks, or tillers, a single plant may yield three to four ears, and since ears can bear up to a thousand kernels, one cross can yield a few thousand progeny. Therefore, even rare genetic events (10^{-7}) can be screened for within one growing period (three months). When a seed is dried and kept under dry and cool conditions, each of these progeny can be conserved for many years. It is already clear that there are many experiments that cannot be done in animal systems (size of offspring, dormancy, gene dosage, somatic embryogenesis) which are readily performed with maize and that because of its phylogenetic position, maize will be a better model system for the human genome than fungi.

In the immature ear with its hundreds of ovaries, each ovary contains an ovule and each ovule has a cell that will develop into the embryo sac. This diploid progenitor cell of the ovule will undergo meiosis, producing four haploid spores of which only one survives. This spore undergoes further divisions and will give rise to the embryo sac surrounded by the ovule. The embryo sac or megagametophyte has quite a specific organization of haploid cells. However, there are two important cells in the embryo sac that we will focus on. For more details on these developmental stages, a detailed description can be read elsewhere (Randolph, 1936). One of the haploid cells of the embryo sac is the egg cell that develops at the end of the embryo sac where the pollen tube will enter. The other cell is the largest cell in the embryo sac with two haploid nuclei, which is also known as the central cell. The reason that we focus on these two cells will become clear when we discuss the fertilization event itself. Similar meiotic divisions take place in the anther of the tassel and also give rise to four haploid spores. Division within the spores produces two very small cells with very condensed chromatin in their nuclei. These are the two sperm cells of the pollen grain which also contains the cell with the working nucleus. At maturity pollen is released through the filament of the anther and blown by the wind to reach either another plant or to fall on the silk of the same plant.

To avoid self-pollination the ears are usually covered with a bag. For crosses, pollen is collected with a bag and sprayed over the ear of the recipient. When pollen is caught by the hairs of an ear, the ovules of which contain eggs that are ready for fertilization, a tube extends from the pollen grain through the silk to the ovule where it deposits the two sperm cells into the embryo sac. One of these sperm cells will fuse with the egg cell and give rise to the embryo, while the other fuses with the central cell. The central cell now contains three haploid nuclei. First the sperm nucleus will fuse with one of the nuclei of the central cell and then this diploid nucleus will fuse with the other haploid nucleus to form a triploid nucleus in the central cell.

We see that pollination leads to a double fertilization event in which the egg cell develops into the diploid zygote and the central cell into the triploid endosperm. In summary, the mature kernel will be composed of three tissues of different origin. The outer cell layer of a mature kernel, or the pericarp, develops from the diploid ovule cells that surround the embryo sac and is therefore of maternal origin. The inner two tissues develop from the embryo sac and are the result of the fertilization events. The endosperm tissue is triploid and receives two genome equivalents from the mother and one genome equivalent from the father. The embryo has one genome equivalent from each

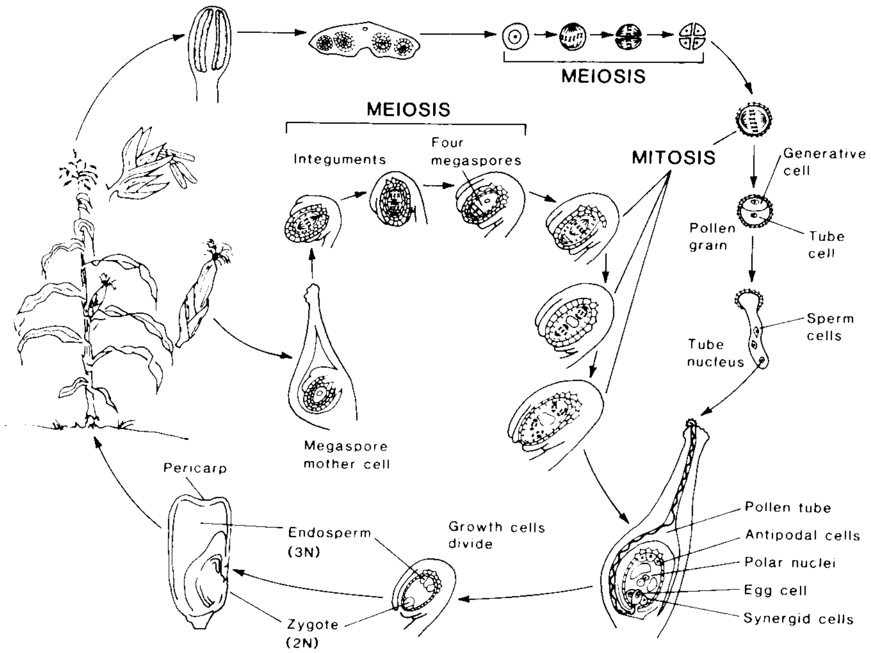

Figure 1 Schematic representation of the reproductive development of the maize plant. Explanation is given in the text under seed development.

parent and gives rise to the progeny plant. Therefore, kernel analysis can provide information on two subsequent generations and on gene dosage. A summary of the reproductive development of the maize plant is shown in Fig. 1.

Of the three tissues of the mature kernel, it is the endosperm in particular that has been studied genetically in great detail in maize. It is the tissue in which energy is channeled from the site of photosynthesis and deposited in the form of proteins and carbohydrates. Embryos can be separated from the endosperm and will germinate alone on nutrients; however, natural germination is dependent on the nutrients in the endosperm. These nutrients are one of the major staple food sources in world agriculture. Therefore, endosperm represents an interesting tissue with which to study gene expression and the molecular parameters of food yield.

IV Storage proteins, the major seed proteins

A Introduction

After fertilization the primary endosperm nucleus will undergo mitotic division without cell division. However, each triploid nucleus is arranged in a highly ordered fashion. The progeny nuclei are placed in planes and the central cell has between 128 and 256 nuclei five days after pollination (DAP). Cell wall deposition then commences. Mitotic divisions of the mononucleated cells continue in this early phase of development. These divisions, however, occur more in the periphery than in the centre of the endosperm at later times. Since cells are fixed in their position, the time at which a phenotypic variation occurs during development can be traced in the cell pattern (Steffensen, 1968). This has been particularly useful for the study of the timing of unstable mutations that are caused by transposition events in the waxy locus (McClintock, 1964, 1978) which encodes a bound starch granule, UDP-glucose starch glycosyl transferase, and controls the accumulation of amylose during endosperm development (Nelson *et al.*, 1965; Echt and Schwartz, 1981). Mitotic divisions cease 12–16 DAP and cell differentiation commences. No further cell divisions occur, although DNA synthesis continues in many cells giving rise to highly polyploid nuclei (Kowles and Phillips, 1985). Starch and storage protein deposition are characteristic of the differentiation process; cells become progressively filled with starch granules and protein bodies. Starch granules accumulate within amyloplasts which are derived from proplastids (Badenhuizen, 1969; see Lonsdale, this volume), while protein bodies are formed from enlargements of the rough endoplasmic reticulum (Khoo and Wolf, 1970). A more detailed description of endosperm differentiation has been given elsewhere (Soave and Salamini, 1984).

The protein bodies described above contain the major fraction of the seed proteins (Burr and Burr, 1976; Burr *et al.*, 1978) which have been classified by their physical properties. Osborne (1924) named the alcohol-soluble fraction prolamins, that soluble in neutral pH and high salt globulins, that soluble in neutral pH and low salt albumins, and that soluble in high pH glutelins. The alcohol-soluble fraction is the largest and contains about 60% of the total protein in the mature kernel. The proteins in this fraction are called the prolamins because of their high content of proline and amide nitrogen. Their synthesis starts about 12 DAP and continues linearly until about 40 DAP (Ingle *et al.*, 1965; Viotti *et al.*, 1975; Marks *et al.*, 1985a,b). Since they dominate the kernel protein, and since the free amino acid pools in the kernel are very small, their primary structure strongly influences the amino acid composition of the kernel and therefore the nutritional quality of maize (Messing, 1984). The prolamin fraction is also called zein to distinguish it from the same group of proteins in other species, for example the hordeins from barley (*Hordeum vulgare*). The zeins are probably the best characterized group of storage proteins and I will use them as an example of storage proteins in general.

B The zeins are a protein family

One significant property of all storage proteins is their existence as families of proteins rather than as a single protein species (Wilson, 1983). Furthermore, these families of proteins are coordinately synthesized during seed development. Thus they share a common architecture in their primary structure and common signals for gene expression (Messing *et al.*, 1983). Even with our modern insight into these protein families at the molecular level, it seems that the original classification by protein fractionation techniques is still valid. Therefore, I will define the zeins as the hydrophobic proteins that are coordinately expressed during maize endosperm development, are secreted into protein bodies, are hydrophobic and which predominantly contain proline and amino acids with amide nitrogen. Nevertheless, the prolamin content and the solubility in ethanol of zeins varies among different subclasses and subfamilies over a large range. All zeins sequenced so far contain 27–40% prolamin (Kirihara and Messing, in prep.). Their solubility in ethanol differs primarily by the absence or presence of reducing agents in the extraction procedure which can be explained by the 3–7 times higher content of cysteine in the primary structure of those that require the reducing agent. In addition, a fractionation of some subfamilies by solvents with different alcohol concentration is also possible (Das and Messing, in prep.). This differential extraction in the absence or presence of reducing agents can be used to divide the zein protein family into two classes, z1 and z2 (Heidecker and Messing, 1986). This division may be useful, not so much in terms of the fractionation technique, but because it indicates the presence of dominant amino acid residues other than glutamine, the main amide nitrogen donor. After the extraction of z1 zeins, other z1 zeins are still extracted with the z2