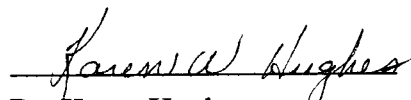


To the Graduate Council:

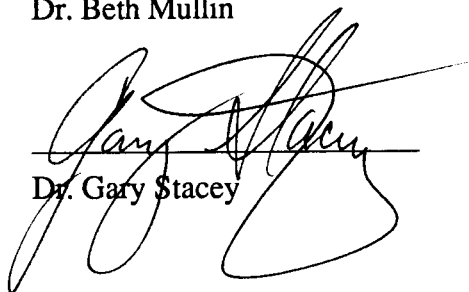
I am submitting herewith a dissertation written by Roel Peter Funke entitled "Development of high molecular weight DNA methods in soybean for positional cloning of genes known only by phenotype". I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.


Dr. Peter M. Gresshoff, Major Professor

We have read this dissertation and recommend its acceptance:


Dr. Karen Hughes


Dr. Beth Mullin


Dr. Gary Stacey

Accepted for the Council:


Associate Vice Chancellor
and Dean of The Graduate School

**DEVELOPMENT OF HIGH MOLECULAR WEIGHT DNA METHODS IN
SOYBEAN FOR POSITIONAL CLONING OF GENES KNOWN ONLY BY
PHENOTYPE**

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Roel Peter Funke

December 1994

ACKNOWLEDGEMENTS

There are many people who contributed to the good atmosphere in which this work was done. They are too many to name, but they include the fluctuating population of the PMG lab and Janice Crockett. Their friendship and support is appreciated very much. In particular, I want to thank Peter Gresshoff for his enthusiasm and for the education I received in his laboratory. Alexander Kolchinsky was directly involved in much of the work, and I cannot exaggerate how important his friendship and collaboration were to me. Debbie Landau-Ellis will inherit the high molecular weight DNA torch in the laboratory, and I want to wish her good luck and thank her for her constant good humor and cooperation. Finally, I want to thank my good friend Jacque McLaughlin for the hours I spent on her computer, for inspiring the way I made my figures, and for filling in the colors over the past year.

"I can, in cartoon, conceive how this codex might be read, how merely speaking its words creates three-dimensional globules whose folds make each a minute chemical computer. I grant that one enormous concatenated clause of A's, T's, C's, and G's is the plan for hemoglobin and another, every bit as inanimate, for insulin.

But I lack the critical keystone in the arch he was after: I cannot see how form emerges from the same mechanism. When to make bone, when pancreas, scale, hair, skin, or bowel? How large should the heart get? How to start it pumping? How indicate a heart at all?

...Pattern-juggling pattern actually makes life. From brittle and spiny to sleek and silver: an impossible spreading text of four letters... Eel on one hand, elm on the other: two combinations on the same letter set."

Richard Powers

-The Gold Bug Variations

ABSTRACT

The purpose of this research was to develop high molecular weight DNA methods in soybean for map-based cloning of genes known only by their phenotype. Soybean will be a source of genes controlling nodulation and symbiotic nitrogen fixation. In particular, the work was started with the intention of cloning the *nts* gene of soybean, which is involved in autoregulation of nodule number in the symbiosis with *Bradyrhizobium japonicum*.

Procedures for the preparation and analysis of high molecular weight DNA were used for physical mapping of a region in the soybean genome containing duplicated sequences using pulsed field gel electrophoresis. The duplicate soybean genome caused difficulties in relating bands on pulsed field gel Southern to linkage groups on the genetic map, and these were resolved by using two-dimensional gel electrophoresis. An estimate of the relationship between physical and genetic distance in a cluster of markers was produced.

10% of a haploid soybean genome was cloned in yeast artificial chromosomes ranging in size from 50 to 900 kilobases, with an average of 175 kb. These clones were shown to be stable and useful for a variety of genome analysis applications, including rescue of end clones, cloning of repetitive sequences, and fluorescent in situ hybridization.

TABLE OF CONTENTS

CHAPTER	PAGE
1. THE USE OF INDUCED MUTANTS TO DISSECT MOLECULAR MECHANISMS OF NODULATION	1
2. ARGUMENT FOR POSITIONAL CLONING AS A STRATEGY FOR THE ISOLATION OF THE <i>NTS</i> GENE	9
3. EXAMPLES OF PLANT GENES CLONED ON THE BASIS OF GENETIC MAP POSITION	13
4. PROBLEMS AND PROSPECTS FOR POSITIONAL CLONING IN SOYBEAN	20
5. PROCEDURES FOR THE ANALYSIS OF LARGE DNA MOLECULES FROM SOYBEAN USING PULSED FIELD GEL ELECTROPHORESIS	33
6. PHYSICAL MAPPING OF A REGION IN THE SOYBEAN (<i>GLYCINE MAX</i>) GENOME CONTAINING DUPLICATED SEQUENCES	58
7. PLANT YEAST ARTIFICIAL CHROMOSOME LIBRARIES AND THEIR USE: STATUS AND SOME STRATEGIC CONSIDERATIONS	75
8. CONSTRUCTION AND CHARACTERIZATION OF A PARTIAL SOYBEAN YEAST ARTIFICIAL CHROMOSOME LIBRARY	84
9. CONCLUSIONS AND PERSPECTIVES	121
REFERENCES	126
APPENDICES	140
APPENDIX 1. GLOSSARY AND LIST OF ABBREVIATIONS	141
APPENDIX 2. DISSECTING MOLECULAR MECHANISMS OF NODULATION: TAKING A LEAF FROM <i>ARABIDOPSIS</i> .	144
APPENDIX 3. METHODS FOR YAC LIBRARY CONSTRUCTION OF SOYBEAN	149

APPENDIX 4. HIGH EDTA CONCENTRATIONS CAUSE ENTRAPMENT OF SMALL MOLECULES IN THE COMPRESSION ZONE OF PULSED FIELD GELS, RESULTING IN SMALLER THAN EXPECTED INSERTSIZES IN YACS PREPARED FROM SIZE SELECTED DNA	172
APPENDIX 5. PULSED FIELD GEL ELECTROPHORESIS	178
VITA	197

LIST OF FIGURES

FIGURE	PAGE
4.1. Mapping of the nts locus between two RFLP markers	26
5.1. Scheme of the BioRad CHEF apparatus	36
5.2. Preparation of high molecular weight DNA in plugs	38
5.3. Optimization of conditions for depurination and UV nicking	45
5.4. Pulsed field gel run under conditions to resolve fragments up to 2 Mb	47
5.5. Pulsed field gel run under conditions to resolve fragments up to 400 kb	48
5.6. Southern blots illustrating variable degrees of methylation	51
5.7. Southern blot demonstrating the effect of DNA concentration	54
5.8. Southern blots hybridized with the markers in the pA-36 cluster	55
6.1. Schematic of the procedure for two-dimensional electrophoresis	60
6.2. Linkage groups H and B of the RFLP map of soybean	61
6.3. pA-36 polymorphism	62
6.4. Southern blots of pulsed field gels	65
6.5. Two-dimensional electrophoretic analysis of marker pA-36	66
6.6. Two-dimensional electrophoretic analysis of pA-89	68
6.7. A physical map of the cluster of markers in linkage group H	71
7.1. Soybean YACs produced without size selection of insert DNA	77
8.1. YAC cloning vector	86
8.2. Partial digestion of soybean DNA	90
8.3. Small YACs produced in spite of stringent size selection	92
8.4. Integrity of large DNA fragments during size selection	94
8.5. Summary of the results of two YAC transformation experiments	97
8.6. Current status of the soybean YAC library	99
8.7. Effect of PFGE size selection on the average size of recombinant YACs	100

8.8. Amplification of pooled DNA from 96-well plates	102
8.9. Amplification of rows, columns, and individual wells	105
8.10. Screening of filter replicas of the soybean YAC library	108
8.11. Isolation of repetitive elements from YACs	111
8.12. Fluorescent In Situ Hybridization (FISH) of YAC A12	113
8.13. Analysis of the stability of soybean YACs	115
8.14. "Vectorette" amplification of the ends of YACs	118
8.15. Sequence of the ends of YAC YBE129	119
A3.1. Pre-electrophoresis and partial digestion of soybean agarose DNA plugs	152
A3.2. Preparative gel	156
A3.3. YAC4 vector preparation	157
A3.4. Scheme of the <i>Eco</i> RI cloning site of YAC4	170
A4.1. Soybean YACs from compression zones	175
A4.2. The effect of EDTA concentration in DNA plugs	176
A5.1. Scheme of the gelbox for CHEF	181
A5.2. Digestion of agarose embedded soybean DNA	190
A5.3. Efficiency of pre-electrophoresis	192
A5.4. Two-dimensional electrophoresis of soybean DNA	193
A5.5. Preparative electrophoresis of soybean YACs	196

CHAPTER 1

THE USE OF INDUCED MUTANTS TO DISSECT MOLECULAR MECHANISMS OF NODULATION

INTRODUCTION

A compelling reason to isolate and study legumes with altered symbiotic performance is that they offer new experimental conditionalities with which to dissect the molecular and physiological mechanisms of nodulation. Better understanding of this complex developmental sequence may provide a basis for modifying or improving the symbiosis. Natural variants have been isolated by screening large numbers of cultivars and geographical accessions, usually with the purpose of finding genotypes with improved performance characteristics such as pH tolerance, higher yield, or "nitrate resistance" (Vincent, 1980). Alternatively, variants may be produced by mutagenesis programs. Mutagenesis rarely improves a plant, but the disruption of specific genes can produce plants with specific and informative failures in the developmental sequence leading up to the mature symbiosis. Also, the genetic basis for natural variability in symbiotic performance may be complex and shrouded by many other genomic differences between cultivars. Mutagenesis has the advantage that it can be performed in particular genotypes chosen for the availability of molecular maps and other prior experimental advantages. Finally, mutagenesis usually produces only one or a few changes in the genome that can, in theory at least, be tracked down more easily and related to the observed phenotype.

The use of natural and induced variation for specific traits has been shown to be useful in reconstructing and analyzing complex developmental and signal transduction pathways in the model plant *Arabidopsis thaliana* (e.g., the light response, reviewed by Chasan, 1994). Genetic analysis of variants can be used to estimate how many genes are involved in a developmental program or environmental response. It can show the dominance relations between different alleles, and double mutants can be used to determine the order in which genes act in a cascade. If the genes can be cloned on the basis of their genetic map location, *in situ* hybridization analysis can be used to determine their time and location of expression. Sequence analysis can suggest a function on the basis of homology to previously cloned and characterized genes. Map-based cloning of genes known only by their phenotype via natural or induced mutations, known as positional cloning (Wicking and Williamson, 1991), has been used successfully in the model plant *Arabidopsis* and in tomato (see Chapter 3).

OVERVIEW OF THE SOYBEAN-*Bradyrhizobium japonicum* SYMBIOSIS

The stages of development of the nitrogen-fixing symbiosis between legumes and soil-borne bacteria of the genera *Rhizobium* and *Bradyrhizobium* have been characterized and reviewed extensively (Brewin, 1993; Caetano-Anollés and Gresshoff, 1991). In the case of the symbiosis between *B. japonicum* and soybean, the bacterium is attracted by flavonoids secreted by soybean roots in the soil and releases richly decorated lipo-oligosaccharides that induce the early stages of the nodulation response (Sanjuan *et al.*, 1992). The ensuing developmental sequence has been divided into discrete, codified phenotypic stages (Vincent, 1980). Although the genomes of both the bacteria and the host contribute to the development of the root nodule, it is essentially a plant-derived organ. The host determines the time and location of nodule initiation, as well as nodule morphology, and regulates the total number and activity of nodules.

The bacterial nodulation signal induces curling of hairs in a susceptible region behind the growing root tip, and the bacteria enter the root where the walls of these hairs are less thick. Soybean forms determinate nodules, in which cell divisions are initiated in the subepidermal layer of the root cortex. Centers of active cell division do not persist past the early stage, and cell extension is necessary for the development of the functional nodule (Gresshoff, 1993). Ensheathed in an infection thread that keeps them isolated from the cytoplasm of the cells in the root, bacteria penetrate successive cell layers of the cortex. Infection thread branches permeate the nodule primordia, and bacteria are released into the host cells. In infected cells the bacteria are enclosed in a peribacteroid membrane (or symbiosome, Roth *et al.*, 1988) where they continue to divide until they mature into nitrogen-fixing bacteroids. Exchange of signals and nutrients takes place across the plant derived peribacteroid membrane. The plant supplies fixed carbon to the nodules primarily in the form of sucrose. Nodule metabolism generates organic acids such as malate and succinate, which are carried across the peribacteroid membrane by a dicarboxylate transporter (Udvardi *et al.*, 1988). Photosynthates are used for the growth and maintenance of nodule tissue, and the energy consuming reactions of N_2 reduction by the bacteroids and assimilation of NH_4^+ in the host cell cytoplasm. Carbon skeletons derived from photoassimilates are also necessary for the synthesis of ureides (in the case of soybean and other tropical legumes) for the export of N from the nodule to the rest of the plant. Fixed nitrogen is transported in the xylem to the shoot of the plant, where the leaves and reproductive organs degrade it to NH_4^+ which is utilized for the synthesis of amino acids, secondary products, and macromolecules including nucleic acids and proteins (Schubert, 1986).

NITRATE INHIBITION OF NODULATION

While low levels of nitrate in the growth medium, in the range 0.5-2 mM, stimulate the growth of legumes and contribute to an increase in nodule mass per plant, higher levels of combined nitrogen have a negative effect on nodulation (Carroll *et al.*, 1985).

The inhibitory effects of combined nitrogen on nodulation are exerted at different levels (Streeter, 1988). These effects have been observed and studied mostly in response to nitrate, since this is the form of combined N that is normally available to the plant. First, the infection process itself is inhibited. In soybean, very early events in the infection process are affected, including root hair deformation, attachment of bacteria to the root, and infection thread formation (Carroll and Matthews, 1990).

Second, nodule mass per plant is negatively affected by the presence of nitrate in the soil at levels exceeding about 4 mM (Carroll *et al.*, 1985).

Finally, nitrogenase activity per unit mass of nodule is diminished by the addition of combined nitrogen (Carroll *et al.*, 1985). This effect of nitrate has received the most attention, because it is more amenable to experimental manipulation and the acetylene reduction assay provides a relatively unambiguous measure of the effect of different treatments on the level of nitrogenase activity.

The mechanisms of nitrate inhibition remain unknown, but it seems likely that different effects are responsible for the inhibition of infection, on the one hand, and for the damping of nodule activity on the other (Streeter, 1988).

A number of paradigms are in use to explain the effect of nitrate on nodule growth and activity (for reviews, see Streeter, 1988, and Carroll and Matthews, 1988). An early theory which sought to establish an external effect of combined nitrogen on nodulation was that the Rhizobia reduced nitrate to nitrite, which in turn catalytically destroyed IAA. Since IAA was thought to be necessary for infection, experiments showing that the addition of nitrate to *Rhizobium* cultures destroyed IAA produced by the bacteria were interpreted as supporting the importance of these reactions in the local, external inhibition of root nodulation by nitrate (Tanner and Anderson, 1963). This idea has not been pursued recently, and the role of bacterial nitrate metabolism in the inhibition of early nodulation events has generally been discounted relative to the importance of internal effects in the legume host (Carroll and Matthews, 1990).

The most popular of the paradigms is the carbohydrate deprivation theory, according to which the assimilation of available nitrate consumes reducing equivalents to the point that not enough carbohydrate is available for nodule growth and activity. Direct tests of this and other models are often difficult and inconclusive, because it is difficult to control all the conditions of an experiment. For example, an experiment showing that feeding with sugars could alleviate nitrate inhibition was considered to provide evidence in favor of the carbohydrate deprivation theory, until it was shown that the sugars also inhibited nitrate uptake and metabolism (Streeter, 1988).

Induced mutants have been used to test the regulatory effects of host nitrate metabolism on nodule activity. Experiments with a nitrate reductase-deficient mutant of *Pisum sativum* showed that nodule nitrogenase activity was less affected in the mutant than in the wild type when root systems were exposed to 5 mM nitrate. Moreover, nodule starch reserves were less depleted in the mutant, which correlated with the fact that a smaller fraction of the nitrate it absorbed from the growth medium was reduced. This evidence was consistent with the carbohydrate deprivation mechanism of nitrate inhibition of nodule function (Walsh and Carroll, 1992).

According to the nitrite toxicity theory, this product of nitrate assimilation is responsible for poisoning nitrogenase enzyme and for the oxidation of leghemoglobin to its ferric (inactive) form. This is a late effect of nitrate on nodule activity. While nitrite is in fact toxic to living organisms and is known to be a potent inhibitor of nitrogenase, controversy surrounds the question of whether enough nitrate reduction takes place in the bacteroids or in the cytosol of infected cells to elevate nitrite levels to the point where nitrogenase is affected (Streeter, 1988).

SYSTEMIC REGULATION OF NODULE NUMBER AND NODULE FORMATION

Systemic regulation of nodule development is manifested in a number of ways (for a review, see e.g., Gresshoff and Caetano-Anollés, 1992). In soybean, nodules first develop in the susceptible zone a certain distance behind the growing root tip, where there are emergent, developing root hairs. As the nodules mature, they suppress the emergence of more nodules in younger regions of the root, so that nodulation is largely restricted to the crown of the tap root. Surgical removal of the mature nodules permits the emergence of more nodules in the susceptible region (Caetano-Anollés *et al.*, 1991). The fact that nodules in the root crown restrict nodulation of young regions basipetally through feedback

regulation was also confirmed by the observation that lower regions of the root system are competent to be nodulated if inoculation is delayed (Pierce and Bauer, 1983). In soybean, this autoregulatory response is exerted through the arrest of nodule growth just prior to emergence (Calvert *et al.*, 1984; Mathews *et al.*, 1989). When arrested infections are released by excision of mature nodules, they develop into functional nodules containing the original bacterial inoculum (Caetano-Anollés *et al.*, 1991).

Another manifestation of systemic autoregulation comes from split-root experiments, in which nodulation in one half of a mature root system inhibits nodulation on the other half when an inoculum is applied to it (Kosslak and Bohlool, 1984; Olsson *et al.*, 1989). The average number of nodules across the entire root system remains fairly constant for a particular genotype, suggesting that there is careful homeostatic control of total nodule number and activity (Caetano-Anollés *et al.*, 1991).

As well as a systemic feedback system, there is evidence that short distance interactions have a role in restricting nodule development. When nodules were removed from the primary root, new nodules emerged not only in the younger regions of the root, but also in those regions from which the nodules were removed (Caetano-Anollés *et al.*, 1991). This suggests that there is a local suppressive field around nodules in which younger infections are arrested.

IMPORTANCE OF NITRATE INHIBITION AND AUTOREGULATION OF NODULE DEVELOPMENT

There has been a longstanding interest in legumes capable of sustaining a nitrate-tolerant symbiosis for a number of reasons (Carroll and Matthews, 1990; Streeter, 1988). First, when legumes are planted in rotation with other crops, nitrate inhibition of nodulation increases the overall need for fertilizer and leaves less residual nitrogen in the soil. Also, when legumes are intercropped with nonlegumes or used as green manure, they compete with neighboring nonsymbiotic plants for combined nitrogen. Finally, if other barriers to yield increase can be overcome, it might become economically feasible to apply fertilizer to nitrate-tolerant legumes in the field to further boost productivity. Thus, increasing the tolerance of the symbiosis to nitrate and understanding the mechanism of nitrate inhibition is an important goal for the improvement of agricultural productivity.

Similar arguments can be applied for legumes altered in the autoregulatory response. The availability of legumes with relaxed systemic regulation resulting in a greater number of active nodules could increase residual nitrogen in the soil.

THE GENOME OF THE PLANT HOST CONTROLS BOTH AUTOREGULATION AND NITRATE SENSITIVITY

Many different studies have led to the conclusion that the genome of the legume host controls nitrate sensitivity during the infection process. For example, attempts to isolate bacteria that could determine a nitrate-tolerant symbiosis have not been successful, and inoculation of nitrate tolerant soybean mutants with different bacterial strains showed that there was no effect of the bacterial genotype and that the host plant was responsible for the observed nodulation phenotype (Carroll *et al*, 1985, Gremaud and Harper, 1989). Furthermore, in alfalfa plants that nodulate in the absence of *Rhizobium*, nodulation was inhibited by nitrate in the growth medium (Truchet *et al*, 1989).

These and other studies also indicate that the nodule is largely a plant-derived organ and that systemic control of nodulation is part of a developmental program that is autonomous to the legume host. Spontaneous nodules in alfalfa emerge in the region of the root that is normally susceptible to infection by *Rhizobium*, and nodule excision experiments showed that these nodules inhibit the appearance of nodules on the lateral roots, suggesting that the presence of bacteria is also not necessary for systemic regulation of nodule number (Caetano-Anollés *et al.*, 1990).

The controlling importance of the genome of the host justifies a strategy of trying to elucidate the mechanisms of nitrate tolerance and autoregulation using plants that show natural or induced variation in these characters.

GERMPLASM FOR THE STUDY OF AUTOREGULATION AND NITRATE INHIBITION

In soybean a number of natural variants have been isolated with superior nodulation and nitrogen fixation characteristics in high nitrate soils (reviewed by Carroll and Mathews, 1988). In particular, some accessions of Korean origin were identified that had 17 times more nodules, 15 times greater nodule mass, and higher rates of nitrogen fixation than cultivar Bragg in the presence of nitrate (Herridge and Betts, 1985). The physiological basis for this nitrate resistance is not clear and could not be attributed to differences in nitrate reductase activity between Bragg and the Korean genotypes. The genetic basis for the phenotype of these lines is unknown, and the prospects for isolating the genes involved by positional cloning seem remote.

Nodulation mutants have been produced in a number of legumes using ethyl methyl sulfate

(EMS) mutagenesis (reviewed by Caetano-Anollés and Gresshoff, 1991). In soybean, several non-nodulating mutants have been characterized, as well as supernodulating mutants that have increased nitrogen fixation ability in the presence of nitrate (Carroll *et al.*, 1985). The mutant phenotypes were shown to be controlled by single recessive genes. In double mutants the non-nodulation phenotype is epistatic to supernodulation (Mathews *et al.*, 1990).

In the absence of nitrate the supernodulation mutant nts382 develops up to 40 times more nodules than the wild type parent Bragg. Nodulation is faster in the mutant (Mathews *et al.*, 1989). However, total nitrogenase activity per plant is the same, showing that there is decreased specific nitrogenase activity in the mutant (Carroll *et al.*, 1985).

Nodulation in the mutants is not restricted to the crown of the tap root, but rather a high proportion of infections along the tap root develop into mature nodules (Mathews *et al.*, 1989). This shows that feedback suppression of nodulation has been lost in the mutants. Other mechanisms to control nodulation still operate in the mutants, and they have been useful in defining additional ways in which the plant controls the timing and frequency of success of infection by the bacterium. For example, while systemic autoregulation of nodulation does not function in the nts mutants, fully developed nodules still exert a local suppressive effect (Caetano-Anollés *et al.*, 1991).

The mutants have also allowed studies of the influence of different plant organs on autoregulation. Grafting experiments by Delves *et al.* (1986) showed that the shoot controls the supernodulation phenotype of the mutant. Thus the shoots of the mutant induced an nts phenotype on wild-type roots, and in reciprocal grafts mutant roots did not express the nts character under the control of a wild-type shoot. The same results were obtained for other independently isolated nts mutants of soybean (Gremaud and Harper, 1989). These results suggest that the shoot portion of the plant produces an autoregulatory signal that is involved in both nitrate inhibition and systemic regulation of nodule number (Delves *et al.*, 1986; for models of systemic control of nodulation see Caetano-Anollés and Gresshoff, 1991; Gresshoff and Caetano-Anollés, 1992). In different supernodulating and hypernodulating pea mutants isolated by EMS mutagenesis, increased nodulation was determined by the root system in one case, whereas grafting experiments with the other mutant showed that the signals for the hypernodulating character originated from the shoot (Rosendahl *et al.*, 1989; Duc and Messenger, 1989). These pea and soybean mutants may represent different genetic lesions in the same pathway and it would be valuable to study them in conjunction, for example to see if they complement one another in intra- and interspecific grafts.

When a number of independent nitrate tolerant supernodulating soybean lines were intercrossed, no complementation was observed. This suggests that the same gene was affected in each case. Thus, either this gene is particularly labile to EMS mutagenesis, or it is the only gene that can mutate to disrupt the plant's internal autoregulation system (Delves *et al.*, 1988). Alternatively, other genes involved in autoregulation and nitrate tolerance may be duplicated or critical to another aspect of the plant development. Mutations in genes that are redundant would be expected to have no phenotype, or perhaps only a conditional phenotype, while mutations in essential genes would be lethal. In either case, they would not be detected in mutagenesis screens.

Nts mutants are primarily defective in endogenous autoregulation. Day *et al.* (1989) showed that when nodule number and mass in nts382 are decreased to wild-type levels by lowering the bacterial inoculum, 'pseudo' wild-type nodules are obtained. The nodules showed wild-type morphology and specific nitrogenase activity, and were sensitive to nitrate shock. Thus the reduced nodule mass and specific nitrogenase activity of the mutant, as well as the apparent insensitivity of nodule function to nitrate, are a pleiotropic effect of the mutation (Gresshoff *et al.*, 1988). Nodule development in the mutant is impervious to levels of nitrate that are inhibitory to wild-type plants, but the experiments with 'pseudo' wild-type nodules show that the mutation does not confer nitrate insensitivity to nodule function. This provides genetic confirmation of the observation that the effects of nitrate on nodule emergence and nodule function are exerted through different mechanisms.

In the course of selecting nodulation mutants with increased tolerance to soil nitrate, the nts condition has been found repeatedly to be coupled with supernodulation. This is true for soybean as well as for other legumes (e.g. Duc and Messenger, 1989; Akao and Kouchi, 1992; Gremaud and Harper, 1989). Increased nodulation is associated with partial nitrate resistance even in natural isolates (Herridge and Betts, 1985). It has been suggested that the effect of nitrate on symbiotic development is very subtle and is simply overwhelmed in so-called nts mutants by the unrestrained nodule formation that occurs in the absence of any internal control of nodule number (Streeter, 1988). It seems likely though that the inhibitory effect of nitrate on nodule initiation and development is due to an interaction between nitrate and a shoot derived autoregulation signal (Day *et al.*, 1989).

Isolation of the *nts* gene by positional cloning would represent one entry point into dissecting the molecular basis of autoregulation, and could disclose the biological reason why supernodulation leads to nitrate tolerant nodule development.

CHAPTER 2

ARGUMENT FOR POSITIONAL CLONING AS A STRATEGY FOR THE ISOLATION OF THE *NTS* GENE

INTRODUCTION

Systemic communication in plants is poorly understood. For some phenomena, such as flowering and responses to wounding or pathogen attack, specific messengers such as methyl jasmonate, systemin, and salicylate have been isolated using enrichment procedures and identified as the active principles that elicit transcription of characteristic proteins in target tissues (Staswick, 1992; Pearce *et al.*, 1991; Raskin, 1992). For example, salicylate was discovered to have a role in the regulation of flowering when fractions of honeydew from aphids feeding on reproductive plants were tested for their ability to induce flowering in plants that were kept in a noninductive light cycle. Similarly, salicylate was discovered to be the "calorigen" that causes heat production in male flowers of voodoo lily by purification and enrichment of an extract from flower primordia. Salicylate has also been identified as a systemic signal in the induction of pathogenesis-related proteins (Yalpani *et al.*, 1991). A receptor for salicylate has been isolated and the corresponding cDNA has been isolated from an expression library using an antibody specific for the binding protein (Chen *et al.*, 1993). Systemin was isolated from tomato leaves as a factor that induces proteinase inhibitor activity, and is able to be transported from wound sites to distal tissues of the plant (Pearce *et al.*, 1991).

ATTEMPTS TO CHARACTERIZE THE SHOOT DERIVED INHIBITOR

Since the shoot derived inhibitor of nodulation has been proposed to be transported to the roots in the vascular stream of the plant (Caetano-Anollés and Gresshoff, 1991), attempts were made to isolate the active principle as a fraction from phloem exudate of wild type plants. However, comparison of phloem exudates from wild type and mutant plants by high performance liquid chromatography showed no reproducible differences and yielded no components from the wild-type exudate that, when applied to mutant plants, could restore a normal nodulation phenotype. Since the extent and composition of compounds that are transported symplastically can vary according to physiological conditions and are very hard to manipulate experimentally, a straightforward search for candidate messenger molecules is not always a promising approach to this problem.

In the absence of other explanations and based on the possibility that hormones might have a role in autoregulation as they do in many other plant developmental processes, attempts were also made to demonstrate the involvement of IAA, GA₃, and ABA in autoregulation. Different treatments were shown to have an effect on nodulation which depended on the nitrogen status of the plant, and some differences in the response of wild-type and mutant plants were observed (Gresshoff and Caetano-Anollés, 1992). These kinds of experiments can support the hypothesis that plant hormones are involved in autoregulation, but they are phenomenological and cannot prove that these known hormones actually mediate the physiological action of *nts*.

Finally, attempts were made to identify polypeptide differences between the leaves of wild type and mutant plants. This strategy was based on the assumption that the absence of the shoot derived inhibitor in the mutants is caused by the transcriptional failure of a gene controlling nodulation. mRNA from wild type and supernodulating plants was translated *in vitro* and the products were analyzed by 2d gel electrophoresis. It was observed that some proteins fluctuated quantitatively with inoculation and development, but no differences correlated with the supernodulation phenotype could be identified (Sayavedra-Soto *et al.*, 1994).

In vitro translation and 2d gel electrophoresis can reveal only major differences and gross transcriptional deficits. As a matter of fact the mutation may not involve a transcriptional failure that would be reflected at the polypeptide level. For example, when the *Pto* resistance gene of tomato was isolated by positional cloning, the product of the gene that conferred resistance was found to be expressed to the same extent in resistant and susceptible plants (Martin *et al.*, 1993).

The majority of mutations caused by EMS are base substitutions, although there are precedents for EMS-induced insertions and deletions. In plant systems, one EMS-induced deletion has been reported, a 30-bp lesion at the maize *waxy* locus (Okagaki *et al.*, 1991). This deletion was virtually undetectable by Southern blot analysis. On the other hand, EMS-induced mutations in the ETR1 ethylene response gene of *Arabidopsis*, which was recently cloned by chromosome walking, were determined to be single nucleotide substitutions resulting in missense mutations (Chang *et al.*, 1993). In this case, no differences were observed in the size or quantity of gene transcripts from wild-type and mutant plants.

There are additional approaches to the cloning of genes which have not been applied to the

nts mutation. Subtractive hybridization has been used to clone genes by hybridizing total genomic DNA from mutant lines to sheared DNA from wild-type plants (Gibson and Somerville, 1993). Subtraction by this procedure should leave a remainder of fragments corresponding to a region of the genome that is missing in the mutant line. As mentioned, because of the mode of action of EMS, it is unlikely that such a sizable deletion exists in *nts382*. Subtractive hybridization could also be used at the cDNA level to isolate genes that might not be expressed in the mutant. This is a more stringent assay for transcriptional differences than a straightforward comparison of *in vitro* translation products, but the same conceptual objections can be raised against it.

In differential display, subsets of mRNA from different sources are compared by PCR amplification (Liang and Pardee, 1992). One primer is anchored to the polyadenylate tail of the mRNAs, and another short arbitrary primer attaches to the transcripts at different distances from the first primer. After reverse transcription and amplification, products resolved on a sequencing gel may distinguish between mRNAs from the different sources.

Another method was recently described which can be used to clone the sequences which differ between two complex genomes. Representational display analysis (RDA) is a procedure for the amplification and purification of restriction fragments that are present in one genome but not in another (Lisitsyn *et al.*, 1993). This approach is attractive because it does not require any knowledge of the time and tissue specific expression pattern of the target gene, and could in principle be used to compare the genomes of wild-type and mutant plants. However, in cases where the method was used successfully, the differences between the genomes were still relatively large, including insertions of viral genomes as single copies in human DNA (modelled by adding λ DNA to human DNA in the correct proportion), rearrangements, and deletions. The method involves digesting the DNA with relatively infrequent cutting restriction endonucleases. Eventually only a subset of the genomes is compared with each endonuclease, and although different endonucleases can be used, the probability of detecting a new restriction fragment produced or lost as a result of a single base pair change or small deletion in a complex genome must be considered to be small.

In summary, although some of these approaches could be worthwhile, they all suffer from similar limitations. They are still unlikely to identify single base pair changes, and the probability of success is smaller if transcripts of the gene containing the mutation accumulate to the same level in wild-type and mutant plants. To some extent, all these methods rely on prior knowledge and are limited by the sensitivity of the techniques used.

The failure to detect a difference does not prove that there is not one. Although the use of leaves is justified by the demonstration that the supernodulation phenotype is conditioned by the shoot, this is not a very strict delimitation of the site of expression of the *nts* gene. If only one or two leaves at a certain stage of development are involved, or only certain parts of those leaves, a very severe dilution of the differences between wild-type and mutant plants could be contributing to the failure to find them. Finally, any differences that are detected could be functionally downstream of the primary lesion.

Since the product of *nts* is unknown and most methods for gene isolation require prior knowledge of sequence or expression pattern which are not available for *nts*, another method for cloning the gene has to be used. Two methods for cloning genes about which nothing is known except for their phenotype and inheritance pattern are insertional mutagenesis and positional cloning (see Kolchinsky *et al.*, 1994). Positional cloning requires the availability of genetic or molecular markers that are linked to the gene, a condition that has been met for *nts*.

POSITIONAL CLONING

Since the *nts* gene is known only by its phenotype and genetics and no clues exist about its nature or the precise site of expression, it is a classical candidate for isolation by positional cloning. The idea of positional cloning is based on the availability of detailed genetic maps and molecular markers from the DNA area surrounding the gene, the ability to handle and analyse large DNA fragments by pulsed field gel electrophoresis, and on the creation of libraries of large DNA fragments cloned in yeast artificial chromosomes (YACs) or other suitable vectors (Paterson and Wing, 1993; Gibson and Somerville, 1993). When molecular markers tightly linked to the gene of interest are found, a contig of overlapping clones has to be built. The search for the gene is based on physical mapping of rare recombinants and detection of the expressed sequences in the contig. Eventual confirmation would come with transformation of the mutant with the putative gene to restore the wild-type phenotype.

CHAPTER 3

EXAMPLES OF PLANT GENES CLONED ON THE BASIS OF GENETIC MAP POSITION

INTRODUCTION

In recent years, plant genes known only by their phenotype have been cloned from *Arabidopsis thaliana* and tomato (*Lycopersicon esculentum*). These plants share biological features that make them suitable for map-based cloning methods. The nuclear genomes are small and relatively devoid of highly repetitive interspersed sequences (Meyerowitz, 1987; Lapitan, 1992). This facilitated the early development of tools and resources that play an important role in positional cloning. The genetic maps of *Arabidopsis* and tomato, based mostly on RFLPs, have been sufficiently detailed that most genes should lie within a few hundred kilobases of the nearest marker (Chang *et al.*, 1988; Nam *et al.*, 1989; Tanksley *et al.*, 1992). This places them within the present capabilities of positional cloning. In addition, comprehensive genomic libraries containing large inserts are available for both species (Grill and Somerville, 1990; Ward and Jen, 1990; Martin *et al.*, 1992). In the case of *Arabidopsis* in particular, these resources are being consolidated rapidly. A physical map of the entire genome is under construction. The genetic maps of both species are being elaborated with the addition of RFLP and RAPD markers. These advances and others will simplify positional cloning in the future (Gibson and Somerville, 1992).

The purpose of this chapter is to summarize the cloning of a number of genes based solely on their position on a genetic map. These examples will give an idea of the scale of the task, and will illustrate the kinds of resources that are needed. In subsequent chapters they will be compared to the status in soybean with regard to the *nts* locus.

CLONING AND FINE MAPPING OF THE GENOMIC SEGMENT SPANNING THE GENE

The first demonstration of map-based cloning was the isolation of a gene controlling fatty acid desaturation of membrane and storage lipids (Arondel *et al.*, 1992). The *fad3* mutation was genetically mapped by scoring the fatty acid composition of seeds of 137 F₂ progeny from a cross between mutant and wild-type lines of *Arabidopsis*. This placed the gene 0.4 cM to one side of a pair of adjacent RFLP markers. Three YACs that hybridized to these markers were identified, and one more clone was added using an end-specific probe,

resulting in a contig that extended a total of 170 kb in the direction of *fad3* from the nearest linked marker. Based on the physical distance between the linked RFLP markers, it was estimated that 1 cM in this region was equivalent to no more than 105 kb, and it was judged that the YAC contig was large enough to encompass *fad3*. Thus, in this case, no additional fine mapping was needed once homologous YACs were isolated, and only one chromosome walking step was needed to reach the region of interest.

The cloning of the ethylene response gene *ETR1* from *Arabidopsis* required more extensive chromosome walking using cosmid libraries (Chang *et al.*, 1993). The gene was first localized to a 10 cM genetic interval between two morphological markers. Within this region, one RFLP marker was found that cosegregated completely with the mutant phenotype in 112 meioses. This marker was used to isolate two YACs that spanned approximately 230 kb. These YACs were used to assemble a fine structure contig from overlapping cosmids. Polymorphic DNA fragments from the chromosome walk were used to locate recombination breakpoints in F₂ progeny and further delimit the location of the mutant gene *etr1-1*. Fine mapping from two directions converged on a 47 kb segment that was flanked by markers that showed no recombination with the mutant phenotype. 75 F₂ plants were analyzed for recombination events to one side of the gene, and a population of 25 was used to detect recombinants as the gene was approached from the other direction.

Two groups reported the map-based isolation of an ABA response gene of *Arabidopsis* (Leung *et al.*, 1994; Meyer *et al.*, 1994). The process of identifying a segment of DNA known to contain the locus of interest was very similar to that outlined for the *fab* and *etr* loci. In both cases, flanking morphological markers were first identified that approximated the location of the gene to a 10 cM interval. RFLP markers in this interval were then analyzed for recombination with the *abil-1* mutation. Leung *et al.* used a marker that was 0.5 to 1 cM away from the locus to initiate a chromosome walk extending 500 kb towards *abil*. Meyer *et al.* identified a marker that cosegregated completely with the locus and assembled a YAC contig that extended about 250 kb to either side of this marker. Additional polymorphic markers were generated from subclones of the YACs in the contig, and mapping of these markers identified a maximal segment of about 150 kb that encompassed the gene.

The cloning of the *Pto* locus on chromosome 5 of tomato, which confers resistance to bacterial speck disease, is the only example of positional cloning in a plant species other than *Arabidopsis* (Martin *et al.*, 1993b). A lot of preliminary work was done to identify markers tightly linked to this gene. There are several reasons for this. First, the tomato

genome is larger than that of *Arabidopsis* (1000 Mb vs. 120 Mb), and the average spacing between the 1000 markers on the tomato RFLP map was about two times greater than in *Arabidopsis*. The density of markers on the *Arabidopsis* RFLP map (currently more than one every 400 kb) means that a gene of interest will lie within 100 kb of the nearest marker about half the time, and it is usually feasible to launch a chromosome walk from the nearest linked markers with reasonable confidence that the gene of interest will lie on YACs isolated using those markers (Gibson and Somerville, 1992). In tomato, the comparative sparsity of the markers (compared to *Arabidopsis* only, - relative to most other plants with larger and more complex genomes tomato is still considered a model plant well suited to map based cloning experiments) means that it may be necessary to enrich targeted regions of the genome with additional markers.

Second, the polymorphisms used to construct the tomato RFLP map were not always informative for mapping the *Pto* locus. The *Pto* resistance gene was introgressed into *L. esculentum* from a wild tomato species, *L. pimpinellifolium*. Thus susceptible and resistant near isogenic lines (NILs) of *L. esculentum* Rio Grande exist, with the resistant lines being polymorphic with respect to *L. esculentum* in the genomic interval surrounding the *Pto* locus. This polymorphism was exploited to identify molecular markers linked to the *Pto* gene (Martin *et al.*, 1991). However, the tomato RFLP map was based on restriction fragment length polymorphisms between *Lycopersicon esculentum* and *L. pennellii*. While the markers from chromosome 5 of the *L. esculentum*/*L. pennellii* map would be homologous and would almost certainly originate from the same region of the genome, they would not always reveal a polymorphism between *L. esculentum* and *L. pimpinellifolium* in that region. Thus a number of approaches were used to identify markers in the *Pto* region (Martin *et al.*, 1993a). First, markers on the tomato RFLP map from the general vicinity of *Pto* were assayed on the NILs to see if any of them revealed informative polymorphisms. Second, the NILs were surveyed with more than one hundred single primers to identify amplification products from the region of chromosome 5 containing *Pto*. Finally, restriction digests of the resistant and susceptible NILs were probed with pools of five random clones to identify polymorphisms (a strategy known as "multiprobing"; Young *et al.*, 1988).

Of 19 RFLP markers in the general *Pto* region, 9 yielded reliable polymorphisms when the resistant and susceptible NILs were surveyed with a number of restriction enzymes. Analysis of the NILs with single primers yielded 4 products that were present in one NIL but not the other. Multiprobing yielded a further 4 polymorphisms (Martin *et al.*, 1993a).

An F₂ population of 1200 plants from a cross between the NILs was developed to determine the order of these markers and to estimate the distances between them and the *Pto* gene. Since the *Pto* gene displays dominant action, susceptible F₂ progeny known to be homozygous recessives (*pto/pto*) were deemed to be most efficient with regard to detecting recombinants in the *Pto* region. Selection was facilitated by the fact that plants carrying one or two copies of the dominant *Pto* gene develop necrotic lesions in response to Fenthion, whereas homozygous recessives are insensitive to this insecticide. Of the 1200 F₂ progeny, 251 showed insensitivity. 81 showed recombination between two markers flanking a 20 cM region spanning the *Pto* locus, and these were analyzed with the previously identified informative markers. Linkage analysis revealed one marker that cosegregated completely with *Pto*. This was a marker from the original RFLP map.

Interestingly, of the four markers in the *Pto* region identified by arbitrary primer analysis of near-isogenic lines, the one closest to the gene was still 15 cM away (Martin *et al.*, 1993a).

The marker that cosegregated with *Pto* lay within 0.6 cM of the gene, taking into account the standard error involved (Martin *et al.*, 1993a). This marker was used to isolate a 400 kb YAC from the tomato YAC library (Martin *et al.*, 1992). The left and right ends of this YAC were isolated and placed on the linkage map of the *Pto* region. The left end fragment of the YAC mapped 1.8 cM away from the gene, whereas the right arm showed no recombination with *Pto* (Martin *et al.*, 1993b). It was important to determine whether the YAC in fact encompassed the gene of interest. To do this, 1300 plants from various F₂ and F₃ families and more than 50 cultivars were screened. One plant was isolated that was homozygous for the RFLP marker associated with the dominant *Pto* allele, but was also homozygous for the right arm polymorphism associated with the susceptible *pto* allele. Analysis of the progeny of this plant showed that they were all resistant to bacterial speck disease, proving that the plant was homozygous *Pto/Pto*, and that the gene was located just inside the right arm of the YAC that had been isolated using the linked RFLP marker.

ISOLATION OF CANDIDATE GENE SEQUENCES

In the cases described here, the cloning process was completed in different ways, although all converged on the final step of cloning candidate gene sequences into plant transformation vectors and testing them for complementation (see below).

For the *Arabidopsis* *ETR1* and *fad3* genes and the *Pto* gene from tomato, candidate gene sequences were identified by hybridizing all the DNA in the minimum genomic segment

spanning the gene to an appropriate cDNA library. This comprised only 47 kb subcloned into cosmids in the case of *ETR1*. The first cDNA clone obtained using a cosmid from the 47 kb interval was identified as the correct gene.

In the case of *fad3* and *Pto*, YACs known to contain the genes of interest (120 kb and 400 kb, respectively) were resolved by PFGE and isolated from low-melting agarose gels. The 120 kb YAC suspected of containing the *fad3* gene was hybridized to a *Brassica napus* cDNA library because a suitable developing seed library from *Arabidopsis* was not available, and a high degree of homology between corresponding genes from these two species was known to exist. Seventeen out of 2×10^5 cDNA clones tested hybridized at high stringency and the longest of these was confirmed to be *fad3* by functional complementation.

Martin *et al.* (1993b) probed about 10^6 clones of a tomato leaf cDNA library with the YAC that had been isolated from a preparative CHEF gel. Thirty promising plaques among 200 that hybridized were analyzed. Two of these cross-hybridized and one cosegregated with *Pto*. This clone detected numerous fragments when hybridized with blots of genomic DNA. It was found that all but one of these fragments were present on the YAC, suggesting that the cDNA clone represents a gene family that is clustered at the *Pto* locus. Genetic complementation using *Agrobacterium tumefaciens*-mediated transformation of a susceptible cultivar confirmed that this cDNA conferred resistance to bacterial speck disease.

In all these cases, a relatively large segment of DNA was scanned for expressed sequences, and a single candidate gene was winnowed from among a number of clones before genetic complementation was attempted. In the course of cloning the *ABI1* gene, two groups independently chose to reverse the order of these steps, testing subclones of a YAC contig for the ability to confer the ABA resistant phenotype. Several cycles of subcloning and retransformation converged on segments of a few kb which were then used to screen cDNA libraries (Leung *et al.*, 1994; Meyer *et al.*, 1994). Leung *et al.* used a 6.5 kb fragment to screen 2×10^6 plaques and recovered 25 clones, while Meyer *et al.* screened 3×10^5 plaques to obtain 7 hybridizing clones.

One reason for these different approaches may have been the relative abundance of the transcripts of the target genes. Arondel *et al.* (1992) reasoned that the *fad3* gene was probably expressed at moderately high levels, and thus used whole YACs to isolate moderately abundant cDNA clones. If, on the other hand, the target gene expressed at low

levels, this approach would probably yield mostly clones of more strongly expressed sequences in the neighbourhood of the target gene. Thus the choice of strategies would be based on some prior intuition about the pattern and level of expression of the target gene, as well as on what is discovered during the course of the work.

Other sophisticated methods exist for the identification of transcribed sequences (Kurnit, 1992). For example, the use of normalized cDNA libraries in which all genes are evenly represented simplifies screening for transcribed sequences by hybridization. In exon trapping, random DNA fragments from the chromosomal region containing the gene of interest are cloned into a vector that contains a splice donor site adjacent to the α subunit of galactosidase. Fragments containing a splice acceptor exon border result in RNA molecules that can be processed to excise the α subunit as if it were an intron, and following bacterial transformation those bacterial colonies that do not express β -galactosidase activity should contain "trapped" exons. Large scale sequencing can reveal the presence of introns and exons because of characteristic base pair compositions and periodicities. These and other state of the art methods have been pioneered as part of the human genome project, and may soon be applied to the more established and advanced model plant species.

COMPLEMENTATION BY TRANSFORMATION TO CONFIRM GENE IDENTITY

In all the examples discussed above, gene identity was confirmed by complementation analysis using *Agrobacterium*-mediated transformation. In the case of the *Arabidopsis* hormone response mutants *abil* and *etr1*, the mutant allele was dominant, and thus candidate sequences from mutant plants were tested for their ability to induce the mutant phenotype upon transformation into wild-type plants. In the case of the *Pto* gene, the dominant allele conferred resistance to bacterial speck disease, and the candidate gene from a resistant variety was transformed into a susceptible cultivar. The transformed plants exhibited resistant phenotypes.

Complementation is facilitated by the existence of an efficient transformation system. Such a system is composed of many different levels and disciplines, requiring appropriate vectors and an efficient DNA delivery system, as well as the ability, in most cases, to regenerate whole plants from transformed explants. In some cases, alternative or abbreviated strategies were used. For example, Arondel *et al.* (1992) exploited the fact that *fad3* has a large effect on the fatty composition of *Arabidopsis* roots and used *Agrobacterium tumefaciens* containing an Ri plasmid from *A. rhizogenes* to induce rooty

tumors on stem explants. The rooty tumors of transformed mutants were analysed for fatty acid composition.

Susceptibility to *Agrobacterium* and well-established tissue culture procedures were among the resources that enabled the success of map-based gene cloning in *Arabidopsis* and tomato. Nevertheless, the establishment of adequate high-frequency *Agrobacterium*-mediated transformation even in a model plant such as *Arabidopsis* took place gradually and required continuous refinements in plant tissue culture and transformation methods at least until the late 1980s (Koncz and Schell, 1992).

CHAPTER 4

PROBLEMS AND PROSPECTS FOR POSITIONAL CLONING IN SOYBEAN

INTRODUCTION

While soybean does not enjoy model plant status, it is a source for genes that control developmental processes, such as nodulation, that are unique to legumes. Thus for the isolation of genes such as *nts*, as well as other genes that have special application in soybean breeding, it is important to develop the necessary resources and technologies.

To evaluate the potential for positional cloning in soybean, it is instructive to compare its biological and experimental characteristics to those of the plants for which this strategy has been successful. Among the relevant biological characteristics are generation time, genome size and composition, and the ease with which discrete crosses can be performed. Certain other features such as seed size and the size of the mature plant become a consideration especially in comparison to *Arabidopsis*, whose minute seeds and relatively small adult size enable economies of space, and, on the other hand, allow for larger and more ambitious experiments. Important experimental resources include the availability of a detailed genetic map which can be applied to different crosses, the ability to isolate high molecular weight DNA for physical mapping and for the creation of a genomic library containing large inserts, and an efficient transformation and regeneration system.

Soybean does not have many of the attributes of a model plant. It is a diploid species, whose genome has the characteristics of an ancient tetraploid (Hymowitz and Singh, 1987; Funke *et al.*, 1993). Its twenty chromosomes are small and morphologically similar (Brar and Carter, 1993). In spite of this difficulty a cytological map of the pachytene chromosomes has been published (Singh and Hymowitz, 1988). The genetic linkage map is sparse compared to that of tomato, comprising about 56 loci that have been mapped to 19 linkage groups (Brar and Carter, 1993). Currently, 16 of the 20 possible trisomics of soybean have been identified, accounting for about 80% of the genome (Xu *et al.*, 1994). This should eventually permit the alignment of linkage groups with individual chromosomes. RFLP maps for soybean were developed comparatively late, but are expanding rapidly (Keim *et al.*, 1989; Lark *et al.*, 1993; see below). Part of the reason for the relatively slow development of soybean cytogenetics is that crosses are tedious to perform because of the small size of the flowers, and seed set is relatively low (Poehlman,

1979). While many wild soybean species contain valuable traits such as cold hardiness, disease resistance and tolerance to salt and heat, a large number of these genes involved are not available for introgression by breeding. This is because few intergeneric crosses are possible in soybean (Brar and Carter, 1993). Nor are many of these traits accessible for positional cloning starting from a wide cross using polymorphic parents.

A distinction should be made between the general problems associated with positional cloning in soybean relative to *Arabidopsis* and tomato, and the specific problems remaining with respect to the *nts* gene in soybean. Thus while the classical and molecular genetic resources for soybean may not be as advanced as for *Arabidopsis*, tomato, and a number of other crop species, the recent development of an RFLP map has allowed the location of the *nts* locus next to a cloned molecular marker (Landau-Ellis *et al.*, 1991; see below). The ability to analyze high molecular weight DNA from soybean has been demonstrated (Funke *et al.*, 1993). The remaining obstacles to cloning the gene are downstream of the first step of isolating tightly linked DNA markers, and include 1) physical mapping in the vicinity of the gene using additional linked markers, 2) the construction of a comprehensive genomic library containing large inserts, 3) the isolation of a clone spanning the *nts* gene, and, 4) the development of an efficient transformation and regeneration system for the mutant plants.

In this chapter some aspects of the soybean genome will be discussed. The mapping of the *nts* locus using the available RFLP linkage map and the search for additional linked markers will be reviewed. Different vectors for the cloning of large DNA fragments will be compared, and the status of transformation technology in soybean will be summarized.

GENOME SIZE AND COMPOSITION

The size and sequence organization of a genome and the degree to which it has been characterized shape the strategy that will be used in positional cloning. They are also predictors of the obstacles and rate of progress that can be expected.

There are about 1.07×10^9 bp in the haploid genome of soybean contained in 20 chromosomes (Blackhall *et al.*, 1991). In general, plant genomes contain much more DNA than seems to be required to encode all the genes and sequence specific functions, and there is a positive correlation between genome size and the proportion of repeated DNA sequences (Flavell, 1980; Lapitan, 1992). About 36% of the soybean genome has been estimated to consist of repetitive DNA, and single copy DNA is organized in stretches of about 1 kb interspersed with 0.3-0.4 kb repeated elements (Gurley *et al.*, 1979). The short

period interspersed pattern of single copy DNA seems to be a property of relatively large plant genomes. In the genomes of *Arabidopsis* and tomato, for example, single copy sequences are 120 kb in length on average (Lapitan, 1992). The short period alternation of single copy sequences with dispersed repeated DNA can be expected to add to the difficulty of chromosome walking in the soybean genome, and suggests that map based cloning should involve as few steps as possible.

Although they are relatively abundant in the genome, no high copy number dispersed repetitive DNA sequences from soybean have been cloned. Rhodes and Vodkin (1988) reported on the organization of *Tgm* transposable elements in soybean, but there are fewer than 50 of these elements in the genome and the elements differ from one another with respect to internal deletions and base pair substitutions. Some clustered repetitive sequences have been characterized and sequenced, including the ribosomal RNA genes and a 92 bp satellite (Gottlob-McHugh *et al.*, 1990; Kolchinsky and Gresshoff, 1994). The utility of dispersed repetitive sequences, such as TGR1 from tomato and *Cin1* in maize, is that they can be used to assay the representativeness of genomic libraries as they are being constructed.

Another salient feature of the soybean genome is that early in its evolution it was duplicated by tetraploidization (Hadley and Hymowitz, 1973). As a result, many molecular markers detect more than one band originating from homoeologous chromosomal regions on Southern blots (Funke *et al.*, 1993). Motifs of molecular markers and even restriction sites are sometimes conserved between related linkage groups, and it can be imagined that this could lead to misapprehensions about where in the genome one finds oneself during a mapping experiment. In other species with large polyploid genomes, such as wheat and other cereals, it has been suggested that the use of related diploid species with homologous genomes could simplify mapping and map based gene cloning (Lapitan, 1992; Kurata *et al.*, 1994).

MOLECULAR MAPS

An RFLP map of soybean is being developed at Iowa State University. *Pst*I plasmid clones polymorphic between *Glycine soja* (PI468.916) and *G. max* (A81-356022) constitute the USDA-ARS/ISU RFLP map of soybean (Keim *et al.*, 1989). Mapping distances between markers are based on recombinational analysis of 60-80 F₂ segregant plants. There are 23 linkage groups, which eventually will be aligned with the 20 chromosomes of soybean and integrated with the 19 linkage groups of the genetic map (see above).

This RFLP map currently contains about 500 loci and covers about 2900 centiMorgans (cM). Using a genome size of 10^9 bp, this means that there is one marker every 2000 kb. However, the markers on the map are probably derived mainly from euchromatic regions of the chromosomes, and therefore the average physical distance between them could be less than this. New markers are constantly being added to the map, and the markers are not distributed uniformly. Clusters of closely spaced markers are evident, as are markers that are separated from their nearest neighbors by many cM.

An important consideration for map based gene cloning is the number of kb per cM of map distance, because this gives some idea of the number of kb that must be traversed in search of candidate gene sequences starting from a marker linked to the gene of interest. Based on the overall recombinational size of the soybean genome and the DNA content, 1 cM represents an average of 500 kilobases of DNA. This value fluctuates throughout the genome, and the relevant conversion value must be determined in the immediate vicinity of the gene of interest.

Another genetic map of soybean is being developed using recombinant inbred lines (RILs) derived from a cross between the cultivars 'Minsoy' and 'Noir1' (Lark *et al.*, 1993). This map is being extended using RFLP, isozyme, morphological, DAF (Prabhu *et al.*, 1994), and biochemical markers. Previously cloned nodulin genes that show polymorphisms between the parental lines are also being placed on the map (F. Ghassemi, personal communication). Currently it encompasses about 1550 cM of the soybean genome in 31 linkage groups. While it is not as extensive as the USDA/ARS map, the use of recombinant inbred lines has the advantage that the segregants are virtually homozygous and in principle unlimited amounts of DNA can be obtained from them for dissemination to other laboratories.

Compared to the concerted effort in *Arabidopsis*, mapping in soybean suffers from the lack of a unified approach between laboratories. Thus in *Arabidopsis* certain pairs of ecotypes are used universally for mapping and the creation of genomic libraries, permitting direct comparison between laboratories of Southern blots and mapping results (Gibson and Somerville, 1992). In soybean, by contrast, the use of different cultivars by different laboratories leads to special problems which are exacerbated by the tetraploid nature of the genome (Funke *et al.*, 1993). Thus while RFLP probes are homologous between different cultivars, the bands on Southern blots are usually different in size and reveal different monomorphic and polymorphic bands. Thus it is often not possible to state whether the segregation pattern obtained with a certain probe in one laboratory describes the same

linkage group to which it was originally mapped.

A computer data base for soybean, called SOYBASE, is helping to redress some of these problems. As well as serving as a repository for information about the molecular biology of soybean, SOYBASE contains images of original Southern blots of markers on the publicly available genetic maps, allowing direct comparisons of band sizes and alleles. This will not eliminate the basic problems associated with repetitive DNA and heterogeneity of alleles between different cultivars, but should help to direct and rationalize mapping projects and extend the usefulness of the map to populations other than the parental lines that were used to construct the map.

COSEGREGATION OF *NTS* WITH A MOLECULAR MARKER

The location of the *nts* locus on the USDA-ARS/ISU RFLP map has been known for some time (Landau-Ellis *et al.*, 1991). To map the *nts* locus, crosses were made between *Glycine soja* PI468.397 and *G. max* (L.) Merrill line nts382. Seeds from a confirmed F₁ plant were sown and screened for their nodulation phenotype. Supernodulating plants, which were homozygous for the *nts* recessive allele, were visually selected from the segregating population and screened for cosegregation of the phenotype with an RFLP marker. The gene was first linked to the polymorphic copy of marker pA-36. pA-36 is duplicated in the soybean genome but appears only once on the RFLP map, the second allele detecting a monomorphic band in *Glycine soja* and *G. max*. *nts* was subsequently found to cosegregate with marker pA-132 from the USDA-ARS/ISU RFLP map, close to pA-36 on linkage group H.

Landau-Ellis *et al.* (1994) showed that the RFLP marker pA-132 contained three *Pst*I fragments that were unrelated and non contiguous in the soybean genome. Only one of them was linked to the *nts* locus. This 1.7 kb fragment was cloned separately from the other two and designated pUTG-132a.

The single copy marker pUTG-132a represents an anchor placing the *nts* gene in a single linkage group (Landau-Ellis *et al.*, 1991, 1994). All of 20 supernodulating plants were determined to be homozygous for the *G. max* RFLP pattern of marker pUTG-132a.

Sequencing of the pUTG-132a marker revealed an 832 bp deletion in *G. soja* (Kolchinsky *et al.*, submitted). Using this deletion, PCR primers were designed that were diagnostic for the pUTG-132a region from *G. max* or *G. soja*, thus permitting rapid screening of a larger number of segregants (Kolchinsky *et al.*, submitted). 171 F₂ plants showing the

supernodulating phenotype, representing 342 individual meiotic events, were screened with these primers. One recombination event was detected between the *nts* phenotype and the *G. max* amplification product. This locates the *nts* gene 1.1 cM away from pUTG132a. Mapping relative to the most distal marker on the same linkage group, pA-381, places the gene between that marker and pUTG-132a (see Figure 4.1; Gresshoff, 1993; Landau-Ellis, personal communication).

SEARCH FOR OTHER MARKERS LINKED TO THE SUPERNODULATION LOCUS

Different approaches can be used to enrich the area around pUTG-132a with additional markers. Since the RFLP map is continually growing, the possibility exists that new markers will be placed in the vicinity of the supernodulation locus. Mapping pUTG-132a onto the 'Minsoy' and 'Noir1' map has disclosed additional linked markers nearby, but these have not been mapped in the population segregating for the supernodulation character (Landau-Ellis, personal communication). Marker pA-381 has also been placed on the 'Minsoy' x 'Noir1' map, but it is on a different linkage group to pUTG-132a, perhaps because the segregating populations are too small to confirm continuity between two linkage groups that in fact lie on one chromosome. A large YAC or BAC clone spanning the pUTG-132a genomic region could also be a source of new polymorphic DNA fragments near the gene.

Rather than randomly enriching the whole genome with markers in the hope that some of the new markers will be located in the region of interest, methods have been developed that allow mapping of specifically targeted genomic regions (Michelmore *et al.*, 1991). The comparison DNA from near isogenic lines (NILs) can be used to reveal polymorphisms in specific regions. For experimental purposes, the wild-type plants and *nts382* have been used as NILs to search for markers tightly linked to the supernodulation locus (Caetano-Anollés *et al.*, 1993). Using single primer PCR amplification of endonuclease digested DNA, a number of amplification products were identified that are tightly linked to the *nts* phenotype in segregating F₂ plants. These polymorphic bands are being cloned and sequenced to convert them to SCARs (sequence characterized amplified regions; Kesseli *et al.*, 1993; J. Padilla, personal communication). SCARs overcome some of the weaknesses of randomly amplified markers and can be used efficiently in fine mapping using larger F₂ populations and in the identification of region-specific clones.

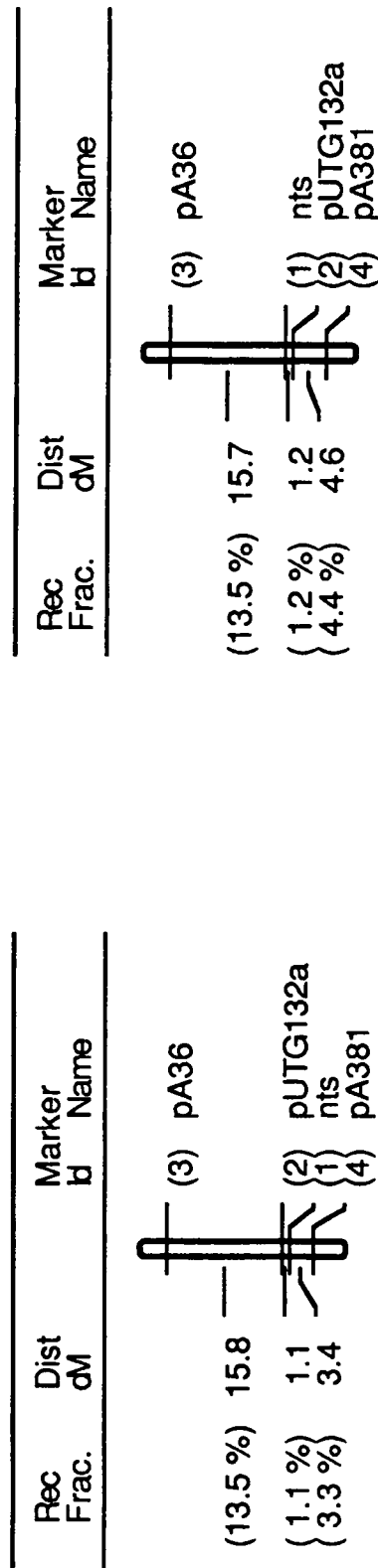


Figure 4.1. Mapping of the *nts* locus between two RFLP markers. The order of the markers and *nts* locus is based on the segregation data described in the text. The maps were prepared by D. Landau-Ellis and reproduced here with permission. The order on the left is 75 times more probable than that on the right. The segregation data obtained from crosses performed in Knoxville places marker pA-381 much closer to *nts* and pUTG-132a than was shown on the USDA RFLP map.

A second method utilizes pooled individuals from a population that is segregating for a trait or marker of interest (bulked segregant analysis or BSA; Michelmore *et al.*, 1991). Outside the targeted region, the DNA in the bulks should be a homogeneous mixture of the genomes of the parents in the cross. The bulks can contain as few as three or four individuals. This method has been used successfully to obtain region-specific markers linked to recessive resistance genes in lettuce (Kesseli *et al.*, 1993).

Different strategies can be employed with regard to the supernodulation locus. First, wild-type and supernodulating individuals from the F₂ population could be bulked and compared. Bulking by phenotype alone entails that some plants in the bulk with the normal nodulation phenotype will be heterozygous at the supernodulation locus, because the wild type allele is dominant to *nts*. This would reduce the efficiency of the procedure. Alternatively, bulks could be constructed that are homozygous for the *G. max* pUTG-132a RFLP pattern, on the one hand, and the *G. soja* pattern, on the other. Finally, intervals can be targeted by bulking individuals that are homozygous for *G. max* and *G. soja* markers flanking the supernodulation locus. This enlarges the region from which polymorphic markers can originate, although the risk of a double crossover in the region exists. Kolchinsky *et al.* (submitted) combined plants that were homozygous for the pUTG-132a marker as well as the two flanking markers pA-36 and pA-381. This represents a window of about 20 cM in which to detect potential *G. max/G. soja* polymorphisms. Bulk segregant analysis using 21 oligonucleotides was performed, and labelled amplification products were analyzed on sequencing gels. One polymorphism was found which was sequenced and converted to a SCAR. When F₂ plants were screened, the SCAR turned out to be only distantly linked, at 15.4 cM from pA-36 outside the cluster (Kolchinsky *et al.*, submitted).

PULSED FIELD GEL ELECTROPHORESIS AND PHYSICAL MAPPING

Physical mapping in the vicinity of the *nts* locus is necessary to determine the local ratio between genetic and physical distance, and to define a distinct DNA segment flanked by linked markers that is known to contain the gene.

Prerequisites for physical mapping include the ability to isolate high molecular weight chromosomal DNA in agarose blocks, digest the DNA with rare-cutting endonucleases, and resolve large DNA fragments by pulsed field gel electrophoresis (PFGE).

The application of PFGE to plant genomes had a delayed start compared to mammals

because of the lack of sufficiently saturated RFLP maps and the difficulty of isolating intact DNA from plant tissues. These obstacles have been rapidly overcome and a lot of physical mapping in plants is focused on specific regions of the genome containing genes known only by their phenotype, with the objective of determining whether the physical size of the region containing the gene will permit cloning of those genes by chromosome walking from neighboring markers. Examples include genes responsible for disease resistance and self-incompatibility (eg. Segal *et al.*, 1992, and Boyes and Nasrallah, 1993). Ganai *et al.* (1989) found that in the Tm-2a region of the tomato genome recombination was suppressed sevenfold compared with the expected average value, obtained by comparing the number of base pairs in the haploid genome to the size of the classical linkage map. This would make a chromosome walk in the region difficult due to the low density of markers and the large physical distances that would have to be cloned and traversed. By contrast, Segal *et al.* (1992) linked two markers separated by 4.1 cM near the I2 locus of tomato on a single 165 kb *MluI-NruI* restriction fragment. This implies a local ratio of physical to genetic distance ten-fold less than the average for the genome, and makes this resistance gene a suitable candidate for isolation by chromosome walking.

In other cases physical mapping yielded only the conclusion that genetically linked markers could not be located on the same large restriction fragment, permitting only the conclusion that the markers were separated by more than a minimum physical distance, and underscoring the need for the accumulation of more markers in targeted regions of the genome.

The development of these procedures for soybean is described in chapter 5, and in chapter 6 the mapping of a cluster of markers on the same linkage group as *nts* is described. Physical mapping in the immediate vicinity of the gene itself is contingent on the isolation of additional tightly linked DNA markers.

The recombinational distance between *nts* and marker pUTG-132a (1.1 cM) places it within several hundred kilobases of the marker pUTG-132a, assuming there are 500 kilobases of DNA per cM (Funke *et al.*, 1993). However, as was stated earlier, the conversion of recombinational distances to kilobases has to be confirmed by physical mapping of the region using pulsed field gel electrophoresis. Marker pA-381 was mapped 20 cM away from pUTG-132a on the USDA-ARS/ISU RFLP map. However, in the F₂ population created in Knoxville, pA-381 is only 4.5 cM from pUTG-132a (Landau-Ellis, personal communication; Figure 4.1). There may be significant errors in the determination of genetic distance, but if the lower value is correct, the distance between the markers pA-381 and

pUTG-132a may translate into a physical distance of only several megabases of DNA. This is within the range of resolution of PFGE. In preliminary physical mapping experiments no comigration of the two markers on distinct DNA fragments produced by the enzymes *SaII*, *NotI* and *SfiI* could be obtained (data not shown). It could be that the physical distance between the markers is in fact too large to be bridged by fragments that can be resolved by PFGE, or that more enzymes and/or partial digests should be tried.

VECTORS FOR CLONING LARGE DNA FRAGMENTS

By genetic mapping, the region of the genome containing a particular gene can probably not be delimited more accurately than a cM or a fraction thereof. This is true even in cases of complete cosegregation of a marker with the gene, because of the limited size of segregating populations and the associated standard error (Landau-Ellis *et al.*, 1991). In the case where a gene has been located between two markers, the region between them should encompass at least a part of a cM. A fraction of a cM represents tens or hundreds of kb of physical distance depending on the region of the genome and the species involved. Thus a region of the genome at least as large as this must be cloned and isolated before a search for candidate gene sequences can be launched starting from the neighboring markers. To do this using plasmid or phage cloning systems would require the isolation and alignment of numerous small overlapping clones. This would be laborious and fraught with error, and in large complex plant genomes might be practically impossible.

Several cloning systems exist that accept much larger insert sizes. The first of these is the yeast artificial chromosome (YAC) system (Burke *et al.*, 1987). Pieces of DNA ranging in size from 50 kb to more than 1 Mb can be propagated as artificial chromosomes together with the yeast host's native complement. The advantage of large insert size is offset to some degree by the effort required to build and array a complete representative library. A summary of the construction and use of plant YAC libraries is included in Chapter 7.

A second system is the P1 bacteriophage system (Sternberg, 1990). This phage vector contains a P1 packaging site which accepts inserts in the range of 80-100 kb. Recombination sites on the vector arms serve to circularize the packaged DNA after transformation into an *E. coli* strain expressing the appropriate recombinase. The P1 plasmid replicon is maintained at one copy per cell and can be amplified prior to isolation. A disadvantage relative to the YAC system is the limit on insert size which is determined by the "headfull" capacity of the phage. On the other hand transformation efficiency is much better. A partial P1 phage library from pine has been constructed (Gorman *et al.*, 1992).

The use of P1 clones in a hierarchical physical mapping and positional cloning strategy has been suggested. Thus large YAC clones might be used to select a subset of P1 clones, which could be tested for complementation of mutant phenotypes. The particular P1 clone which is determined to contain complementing activity can then be used to select a subset of smaller plasmid or phage clones to further localize the DNA segment that contains the target gene (Matallana *et al.*, 1992). Efforts are under way to extend the complementation strategy one more tier by developing a YAC-plant cell complementation system. If there are several overlapping YACs which could contain the gene of interest, the ability to transform plants with these large DNA fragments could accelerate the localization of genes in map based cloning strategies (e.g. Wordragen *et al.*, 1994).

More recently, the BAC (bacterial artificial chromosome) system has been developed (Shizuya *et al.*, 1992). This is based on the single-copy plasmid F factor of *E. coli*. Fragments as large as 300 kb are stably maintained in this vector, which also has the advantage of high transformation efficiency of the bacterial host.

Later in this thesis the construction of YACs containing soybean DNA is described. Reasonable insert sizes were obtained (average 250 kb) in clones that currently cover about 10% of the genome. The clones were stable through repeated subculture, and it was possible to isolate ends of YACs and hybridize them back to genomic DNA digested with different restriction enzymes.

TRANSFORMATION OF SOYBEAN TO CORRECT THE MUTANT PHENOTYPE

Candidate gene sequences would have to be tested via transformation of mutant plants to demonstrate that they can restore the wild-type nodulation habit. A sufficient number of independently derived transformed lines would have to be produced to compensate for position effects and varying patterns and levels of expression (De Block, 1993; Christou *et al.*, 1993).

Various techniques exist for the introduction of genes into plants, the most prominent being *Agrobacterium*-assisted transformation and biolistics. However, the delivery systems differ in their efficiency and their applicability to the problem at hand.

Agrobacterium tumefaciens has been used successfully to transform a wide spectrum of dicotyledonous plants (De Block, 1993). The system is relatively simple, and usually only one or a few precise insertion events are obtained. The success of this approach depends on

a number of factors, including the susceptibility of the explant material to *Agrobacterium*, which is determined by whether the plant exudes products that induce the bacterial *vir* genes. A further requirement is the stable integration of the transforming DNA into the plant genome. Finally, the transformed explant has to be capable of regeneration into a mature plant.

In soybean, varieties that are susceptible to *Agrobacterium* are rarely competent to be regenerated into whole plants. The only cultivar with which this has been achieved is Peking (Hinchey *et al.*, 1988). Even in this case, the transformation efficiency was low and only a few independent transformed lines were obtained. Because of the low susceptibility of soybean to *Agrobacterium*, research has been directed at artificially inducing the bacterial *vir* genes prior to cocultivation with the plant material, and towards the isolation of Ti plasmids and bacteria with chromosomal backgrounds that condition high virulence on soybean (Bond and Gresshoff, 1993; Kovács and Pueppke, 1993). At the same time, other DNA delivery systems that are less genotype dependent have been developed and have begun to supplant *A. tumefaciens* transformation in soybean.

The bacterium *A. rhizogenes* is the causal agent of hairy root disease and is virulent on a wide range of dicotyledonous plants. It transforms plants by inserting T-DNA from the root-inducing (Ri) plasmid into the host genome. To transfer specific desired genes to the plant, the sequences can be incorporated into the Ri plasmid, or a binary plasmid derived from *A. tumefaciens* may be used (for a review see Bond and Gresshoff, 1993).

A relatively high frequency of transgenic roots can be obtained, which has permitted the use of this gene transfer system to study the genetics of nodulation in soybean. For example, Miao and Verma (1993) studied the function of nodulin-26 in hairy roots induced by *A. rhizogenes* in both homologous and heterologous host plants. Hairy roots can be induced in a wide variety of genotypes, including *G. max* cv. Bragg and the mutant nts382. However, Bond and Gresshoff (1993), reported that transgenic soybean roots did not exhibit normal nodule morphology. The most important shortcoming of this DNA delivery system with respect to the problem at hand is that to date no whole plants have been regenerated from *A. rhizogenes*-induced roots of soybean. Since the *nts* gene appears to be expressed in the shoot portion of the plant, this means that *A. rhizogenes* is not a suitable vehicle for complementation experiments with candidate gene sequences.

The most promising transformation system for soybean appears to be particle bombardment of tissue with micron-sized DNA coated gold particles (Bond and Gresshoff, 1993;

Christou *et al.*, 1993). It has been demonstrated that this method can be used to deliver DNA to target tissues such as immature embryos and embryonic suspension cultures (Christou *et al.*, 1993; Finer and McMullen, 1991). Whole plants could be regenerated from transformed target tissues. Although some chimeric plants are obtained which show aberrant segregation in subsequent generations, the efficiency of the procedure is reportedly high enough that many plants can be generated in which the introduced genes are inherited in a stable Mendelian manner. Independently transformed lines can be screened for copy number and pattern and level of expression to compensate for variations caused by the chromosomal domain into which the gene is inserted. Christou *et al.* (1993) report that every soybean variety they have tested has yielded transgenic plants, suggesting that this transformation procedure is relatively immune to genotypic variation. Time considerations are also important in the overall strategy, and according to the timetable of Christou *et al.* seeds can be harvested from the primary transformed regenerant 5 months after particle bombardment. This compares to a generation time of 4 to 6 weeks in *Arabidopsis*.

Thus, in principle, complementation analysis does not pose a barrier to completing the positional cloning strategy in soybean. However, this depends on being able to reproduce the best reported transformation efficiencies, and sophisticated technology is often not easily transferred to a different laboratory. Thus while transformation and tissue culture technology for soybean do not appear to be as advanced or transferrable, and some disadvantages of soybean with respect to model plants such as tomato and *Arabidopsis* are a matter of basic biology, rapid progress in this area means that it is possible that similar success can be achieved in the future.

CHAPTER 5

PROCEDURES FOR THE ANALYSIS OF LARGE DNA MOLECULES FROM SOYBEAN USING PULSED FIELD GEL ELECTROPHORESIS

INTRODUCTION

Pulsed field gel electrophoresis (PFGE) is a technique for resolving large fragments of DNA ranging in size from 50 kb up to 12 Mb or more. In addition to physical mapping regions of particular chromosomes, the technique can be used to relate genetic distance to physical distance in these regions (see Appendix 5 for a review of the principles and applications of PFGE). This relationship varies throughout the genome, and PFGE can be used to identify areas of unusually high or low recombinational activity. It is a critical meterstick to possess en route to the cloning of genes that have no defined biochemical product. This information gives an indication of the physical distance that must be covered in search of the gene starting from a polymorphic marker that has been found to be linked to it.

The *nts* locus has been mapped close to the molecular marker pUTG-132a (1.1 cM), in linkage group H of the ARS/ISU RFLP map of soybean (Landau-Ellis *et al.*, 1991, 1992, 1994). As discussed in Chapter 4 of this thesis, there are no other markers in the vicinity of pUTG-132a that might be used to construct a physical map of the region. The genetic distance to the nearest flanking marker, pA-381, probably cannot be bridged by the size DNA molecules that can be produced by available methods of DNA fragmentation and analyzed by PFGE (see Chapter 4).

As an alternative, it was decided that another cluster of markers in the same linkage group would be analyzed. This was done mainly as a pilot study to develop an approach to physical mapping in soybean. Two markers in this cluster, pK-9 and pA-89, are inseparable by segregation analysis. pA-36 is 1 cM from this pair, while pA-69 is 1 cM distant on the other side (see Figure 6.2). Marker pA-36 is 12 cM from pUTG-132a and the *nts* locus. During the course of this study, the fact that all of the markers in the cluster have at least two copies in the genome was discovered to be an obstacle to the correct interpretation of Southern blots. This problem is discussed in more detail and resolved in the Chapter 6 of this thesis.

The object of the present chapter is to describe the methods that had to be developed for the analysis of large DNA from soybean. These include the isolation of high molecular weight

DNA in agarose plugs in a condition suitable for digestion by restriction enzymes. The range of fragment sizes produced by different enzymes had to be determined, and conditions for the transfer of large DNA to membranes had to be optimized to allow their detection by single copy probes.

PRINCIPLES OF PFGE

The separation of molecules by conventional electrophoresis is achieved by migrating DNA through a gel matrix under the influence of a single static electric field. The rate at which the molecules migrate depends on their length because they are sieved through the pores of the matrix. Short molecules move rapidly in a relatively straight path because they fit through most of the pores in the matrix. Longer molecules have a more convoluted path through the gel and therefore experience less overall movement in the direction of the electric field.

Molecules larger than about 30 kb have about the same cross-sectional area and are sieved through an agarose gel at the same rate (Birren and Lai, 1993). Therefore they cannot easily be separated from each other by conventional electrophoresis and migrate as a single high molecular weight band. The range of resolution can be extended by lowering the voltage and the agarose concentration, but less than satisfactory results are achieved at the expense of long run times and the inconvenience of working with low percentage gels.

PFGE exploits the principle of DNA relaxation to extend the range of resolution that can be achieved by electrophoresis more than 100-fold (Schwartz and Cantor, 1984). DNA molecules in solution exist as random coils, and these coils change shape in an electric field. The rate at which the molecules relax back to their random coil configuration when the field is removed is very sensitive to their molecular weight. In PFGE, DNA molecules are subjected to alternating electrical fields. Each time the direction of the field is switched, the molecules have to relax and assume another shape to begin migrating in the direction of the new field. Long molecules spend more time relaxing and realigning to the new field than shorter molecules, and spend a smaller proportion of each switch interval actually migrating through the gel. By changing the switch interval, different ranges of resolution can be achieved. The net movement of DNA molecules migrating in zig-zag fashion under the influence of regularly alternating fields of equal strength is straight down the gel, provided that the electrical fields are uniform across the gel.

In early PFGE units, the angle between the applied fields was not uniform across the gel. As a result, lanes were characteristically bent. Moreover, the field strength was not uniform across the gel and DNA molecules moved at different rates depending on their position,

making lane-to-lane comparisons and size determinations problematic (Lai *et al.*, 1989). These problems were overcome in the CHEF unit (contour clamped homogeneous electric field; Chu *et al.*, 1986; Figure 5.1). The CHEF unit from Biorad was used throughout the work described in this dissertation.

PREPARATION OF SOYBEAN CHROMOSOMAL DNA IN PLUGS

There are estimated to be 1.07×10^9 bp in the haploid genome of soybean (Blackhall *et al.*, 1991) and currently the the USDA-ARS/ISU RFLP map spans 2900 cM (see Chapter 4 for a description of genetic maps of soybean). Therefore on average there are about 350 kb of DNA per cM on the map. The study of large DNA fragments spanning one or more recombinational intervals requires that the starting material should have a very high average molecular weight.

DNA molecules in solution are extremely sensitive to shearing, and isolation procedures that involve pipetting and ethanol precipitation usually yield DNA with an average molecular weight no larger than 20-50 kb. To avoid handling DNA in liquid, suspensions of live cells are usually mixed with low melting point agarose and embedded in plugs or blocks. The embedded cells are then lysed with detergent in a high-EDTA buffer. Nucleases and cellular debris are removed by diffusion in the presence by proteinase K, leaving high molecular weight chromosomal DNA essentially intact and accessible to restriction enzymes.

This solution was first described by Schwartz and Cantor (1984), who used it to produce the first electrophoretic karyotype of yeast. It has been adapted to study DNA from many organisms. Plants pose a special problem because of the cell wall which must be removed. High nuclease activities result in a certain amount of degradation of the DNA, and deposits of starch or phenolic compounds in the agarose blocks may interfere with the diffusion or activity of restriction enzymes. A procedure for the preparation of high-molecular weight DNA from *Phaseolus vulgaris* used isolated nuclei embedded in agarose (Creusot *et al.*, 1992). The DNA appeared to be of good quality and in principle this method could help to reduce problems caused by carbohydrates and other compounds located in the cytoplasm.

Experiments were performed to determine the best source of soybean protoplasts. The different tissues used for starting material included young leaves, hypocotyls and root tips of 3 to 6 day old seedlings, and developing cotyledons removed from seed pods (Gamborg *et al.*, 1983; Newell and Luu, 1985; Lin *et al.*, 1987). It was found that protoplasts isolated from seedling hypocotyls and root tips were large and fragile, so that centrifugation during

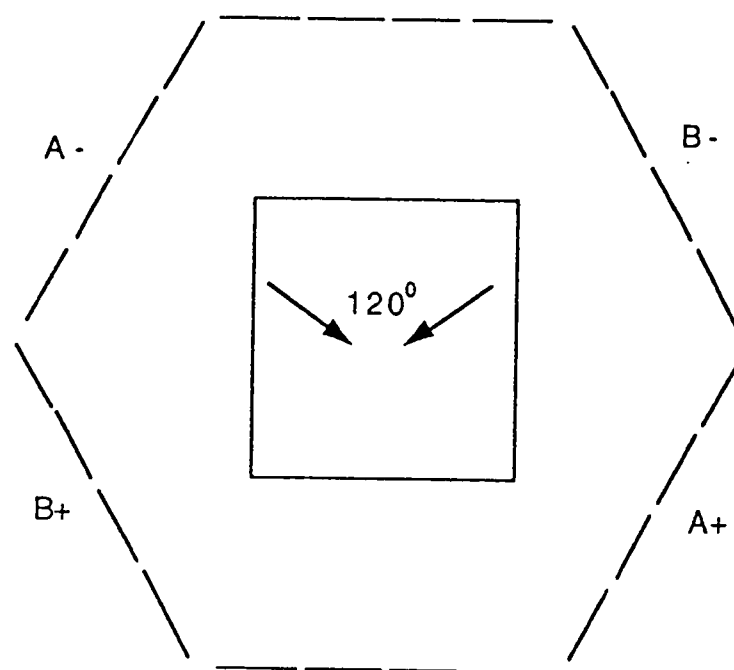


Figure 5.1. Scheme of the BioRad CHEF apparatus used for the work referred to in the text. 24 electrodes are arranged in a hexagonal array around the gel chamber to generate the electric fields A and B. With the position of the gel as shown, DNA molecules migrating within the contour undergo a 120° reorientation each time the field is switched. The electrodes are clamped to specific potentials calculated to make the electric fields homogeneous at every point within the array. This homogeneity ensures that the samples loaded in different lanes have the same trajectory down the gel and have the same mobility, thus allowing lane-to-lane comparisons.

the purification steps drastically reduced the yield. However, the DNA in plugs prepared from these protoplasts was of good quality and could be cut readily with restriction enzymes (Figure 5.2). Developing cotyledons produced numerous protoplasts free of debris from vascular tissue and hairs. The cells could be embedded at high density but proved to contain high concentrations of starch and phenolic compounds. This gave the plugs an opaque milky appearance and caused them to turn brown unless 100 mM DTT was included in all solutions used to rinse the plugs before incubation with a restriction enzyme. The deposits of starch could be removed with amylase, but clarifying the plugs in this way also degraded the DNA because of nuclease contamination of the amylase preparations. Leaf tissue was determined to be the best source of protoplasts in terms of ease of preparation, handling, and the eventual quality of the DNA in the plugs.

Glycine max (L.) Merrill plants were grown in the greenhouse. Young expanding trifoliate leaves (1 cm or less in length) or primary leaves were harvested for the preparation of protoplasts. Surface sterilized 1-2 mm leaf strips were digested with 2% driselase (Sigma Chemicals), 0.2% pectolyase Y-10 (Sigma Chemicals), and 1% cellulase Onozuka R-10 (Serva) for 5 hours with gentle shaking in a buffer (PB) containing 0.4M mannitol, 10 mM CaCl₂, 5mM MES, and 5mM ascorbic acid, pH 5.6. Following digestion the mixture was filtered through a 43 mm Nylon mesh and protoplasts were collected by centrifugation at 100xg for 10 minutes. The cells were rinsed once with PB and after a second centrifugation step were embedded in 0.5% LMP agarose (INCERT, FMC) at 2 to 3×10^7 per ml in 200 ml aliquots. Assuming that the diploid nucleus of soybean contains about 5 pg of DNA (Gurley *et al.*, 1979), this yielded plugs with a nuclear DNA concentration of about 100 to 150 μ g per ml. The total DNA concentration in the plugs was higher if the plastid and mitochondrial genomes are taken into consideration (120 to 180 μ g per ml, if organelles account for 15% of total cellular DNA; Grierson and Covey, 1988).

Plugs were transferred to at least three times their combined volume of ESP (0.5M EDTA, 1% sarkosyl, 1 mg per ml proteinase K) at 50°C while being gently agitated at 50 rpm. After 24 hours the solution was replaced with fresh ESP and left for a further 24 hours, at which point the plugs could be prepared for restriction digestion.

It was recently reported (Chen *et al.*, 1994) that in spite of the fact that embedded DNA is thought to be protected from mechanical shear, even gentle shaking (10 rpm) causes damage to DNA in lysed immobilized cells. All DNA plugs used during the experiments

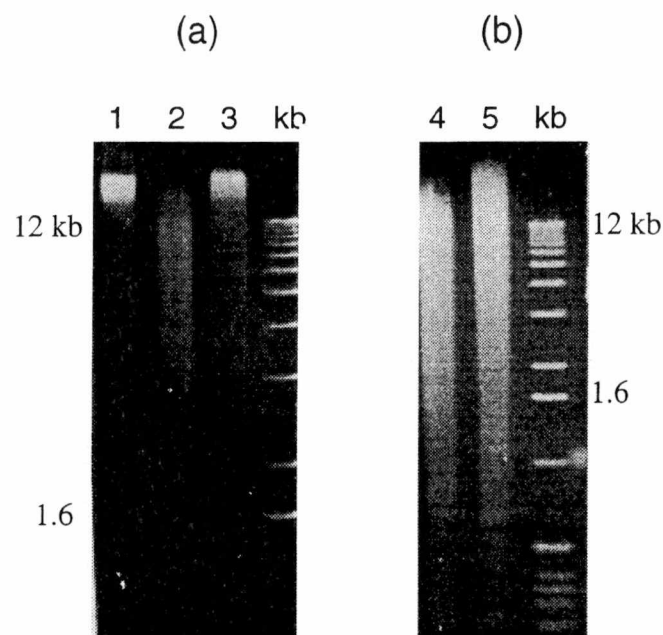


Figure 5.2. Preparation of high molecular weight DNA in plugs from root tip (a) and mesophyll (b) protoplasts, and illustration of different efficiencies of restriction enzymes acting on DNA embedded in agarose. 1= DNA plug not treated with a restriction enzyme. A portion of the DNA leaves the plug and is seen as a single brightly stained band that moves more slowly than the 12 kb marker of the kb ladder. 2= *Hind*III digest. 3= DNA plug incubated with 50 units *Eco*RI. 4 and 5= *Hind*III digests of different batches of DNA plugs prepared from leaf protoplasts.

described here were gently agitated during the incubations in ESP, and no improvement in the quality of the DNA was seen when this shaking was eliminated from the procedure.

RESTRICTION DIGESTION OF CHROMOSOMAL DNA IN PLUGS

Proteinase K was inactivated by washing the plugs twice for one-half hour in TE buffer containing 1 mM PMSF (100 μ l of 100 mM freshly prepared acetone stock per 10 ml of TE) at 50°C. The plugs were rinsed three more times with fresh changes of TE without PMSF.

One-third to one-quarter of a 200 μ l plug (about 5-10 μ g of DNA) was weighed in a tared microcentrifuge tube. Treating the plug as void volume, water and 10x restriction buffer were added to make 100 μ l of 1x buffer and the tube was placed on ice for one half hour. The liquid was removed and replaced with 1x buffer and 10-20 units restriction enzyme per μ g DNA. The mixture was left at 4°C for an additional half hour to allow the enzyme to diffuse into the plugs before becoming fully active. Digestions were carried out at the recommended temperature and terminated after at least 6 hours by adding 500 μ l of ESP and incubating at 50°C for one hour before commencing electrophoresis.

Figure 5.2 shows examples of DNA plugs prepared from embedded root tip and mesophyll protoplasts. The plugs were treated with enzymes that cut relatively frequently in the DNA and analyzed by conventional electrophoresis to give a rapid estimate of the quality and concentration of the DNA. A plug not treated with restriction enzyme is shown in lane 1. The high molecular weight DNA from the plug has entered the gel matrix and moves as a single unresolved band. A *Hind*III digest in lane 2 shows that the immobilized DNA is a suitable substrate for this enzyme. However, *Eco*RI (lane 3) is unable to digest the DNA in plugs to completion. A number of enzymes, including *Eco*RI, are known to cut embedded DNA with poor efficiency. Lanes 4 and 5 contain *Hind* III digests of DNA plugs prepared from mesophyll protoplasts.

CHOICE OF RESTRICTION ENZYMES FOR PULSED FIELD GEL ELECTROPHORESIS

To generate large DNA fragments of soybean DNA, restriction enzymes with long recognition sequences (6-8 base pairs) which are expected to occur relatively infrequently may be used. Enzymes containing CpG within their target sequences would also be expected to produce large fragments, because the genomes of leguminous plants, like those

of mammals, are relatively depleted of this dinucleotide (McClelland, 1983). In coding sequences, CpG is found at only 49% of the frequency predicted from a random distribution of bases.

Table 5-1 shows a list of the enzymes that were used to generate large DNA fragments from embedded soybean DNA. Table 5-2 shows the fragment sizes that are expected using these enzymes, calculated from a database of 157,227 base pairs of soybean nuclear genomic DNA (Honeycutt, 1991). The fragment sizes calculated from the mononucleotide frequency are derived by multiplying together the frequencies in the database of the individual nucleotides in the recognition sequence. The fragment sizes calculated from the di- and trinucleotide frequencies take into account the occurrence in the database of the overlapping di- and trinucleotides in the recognition sequence (McClelland *et al.*, 1987). Some bias in DNA sequence arrangement, and in particular the suppression of CpG, is evident from the increase in the predicted fragment size when the frequency of di- and trinucleotides is taken into account. In principle, restriction enzymes producing large fragments could be chosen by the criterion that their recognition sequence contain one or more of the rarest di- and trinucleotides in the genome. However, this method of predicting fragment sizes is less than perfect, because it uses sequences from a database containing protein coding sequences only. The noncoding part of the genome, which contains the majority of the DNA, could have different biases in base composition and arrangement.

Table 5-3 shows the sizes of fragments that were detected by hybridization with probes from the USDA-ARS/ISU RFLP soybean RFLP map. In general, the observed fragment sizes are larger than those predicted from the protein coding sequences in the database.

PULSED FIELD GEL ELECTROPHORESIS AND OPTIMIZATION OF CONDITIONS FOR SOUTHERN TRANSFER AND HYBRIDIZATION

A CHEF (Bio-Rad) electrophoresis unit was used in all experiments. Running buffer (0.5xTBE) was recirculated through the electrophoresis chamber and maintained at about 14°C in all experiments. Running conditions were varied according to the range of sizes of DNA fragments to be resolved.

For size standards in the 50-600 kb, range oligomers of λ DNA worked well. These were prepared by allowing λ DNA mixed with LMP agarose (30 μ g per ml) to concatemerize at 4°C in 100 mM NaCl, 10 mM Tris. HCl (pH 8), and 1 mM EDTA. Higher molecular weight size standards were produced by embedding spheroplasts of yeast (*Saccharomyces cerevisiae* YNN295) in LMP agarose to a concentration of 1×10^{10} cells per ml, and

Table 5-1: List of the rare-cutting restriction enzymes used in pulsed field gel analysis of soybean. The recognition sequences and methylation sensitivities of the enzymes are shown (*Source:* New England Biolabs Catalog, 1992).

Restriction enzyme	Recognition sequence	Sites cut	Sites not cut
<i>Mlu</i> I	ACGCGT	?	A ^{m5} CGCGT
<i>Not</i> I	GCGGCCGC	GCGGCCG ^{m5} C	GCGG ^{m5} CCGC GCGG ^{m5} CGC
<i>Sal</i> I	GTCGAC	GTCGA ^{m5} C	GT ^{m5} CGAC GTCG ^{m6} AC
<i>Sma</i> I	CCCGGG	^{m5} CCCGGG	^{m4} CCCGGG ^{m5} CCCGGG C ^{m4} CCGGG CC ^{m4} CGGG
<i>Sfi</i> I	GGCCN ₅ GGCC	GG ^{m5} CCN ₅ GG ^{m5} CC GGCCN ₅ GGC ^{m5} C	GGC ^{m5} CN ₅ GGCC
<i>Xho</i> I	CTCGAG	?	CT ^{m5} CGAG CTCG ^{m6} AG ^{m5} CTCGAG

Table 5-2: Expected sizes of fragments produced by the action of rare-cutting enzymes in the soybean genome. See text for discussion (*Source:* Honeycutt, R.J. (1991). Physical mapping of *Rhizobium meliloti* and *Glycine max* L. (Merr.) using pulsed field gel electrophoresis. Ph.D. dissertation, Iowa State University, Ames, IA).

Restriction enzyme	Recognition sequence	Predicted fragment size (kb)			
		sample ^a	mono ^b	di ^c	tri ^d
<i>Mlu</i> I	ACGCGT	31	9	35	37
<i>Not</i> I	GCGGCCGC	not found	750	1560	1987
<i>Sal</i> I	GTCGAC	22	9	21	26
<i>Sfi</i> I	GGCCN ₅ GGCC	158	750	371	359
<i>Sma</i> I	CCCGGG	17	26	26	33
<i>Xho</i> I	CTCGAG	10	9	16	10

^a Number of occurrences in a database of 157, 227 base pairs of soybean nuclear DNA

^b Calculated from the mononucleotide frequency of the database

^c Calculated from the dinucleotide frequency of the database

^d Calculated from the trinucleotide frequency of the database

Table 5-3: Sizes of restriction fragments detected by probes from the USDA RFLP map of soybean

Enzyme	Size of fragments hybridizing to probes (kb)	Size range (kb)	Approx. mean size (kb)
<i>Mlu</i> I	23, 50, 50, 60, 60, 80, 90, 100, 100, 110, 110, 125	23-125	80
<i>Not</i> I	280, 450, 1000,	280-1000	600
<i>Sal</i> I	45, 60, 70, 70, 70, 85, 90, 100, 180	45-180	80
<i>Sma</i> I	150, 200, 220, 245, 350, 370	150-370	250
<i>Sfi</i> I	350, 650, 750, 1100	350-1100	470
<i>Xho</i> I	50, 60, 60, 60, 75, 120, 130, 140, 150	50-150	90

preparing plugs for electrophoresis as described in Appendix 3. Whole chromosomes of yeast are resolved ranging in size from 225 to 2200 kb. As size standards for fragments less than 50 kb in size, *Hind*III digested λ DNA was used. In the size range 0.2-225 kb, and above 2200 kb, where whole chromosomes of more unusual lower eukaryotes are useful, a variety of high quality size standards have become commercially available (e.g. from Sigma (St. Louis, MO)). These were used in some of the experiments described later in this thesis.

It was necessary to optimize the conditions of transfer of large DNA fragments to membranes in order to achieve the strongest possible signals with Southern hybridization to soybean genomic probes. To do this, yeast chromosomes resolved on a pulsed field gel were nicked by UV irradiation (using StratalinkerTM 1800 from Stratagene), or depurinated for different lengths of time with hydrochloric acid before being transferred to a membrane and probed with ³²P labelled total yeast DNA (Figure 5.3). The use of yeast chromosomes, which range in size from 225 to 2200 kb, made it possible to assess whether there was a difference in the efficiency of transfer of small and large fragments. Figure 5.3 demonstrates that the strongest signal is obtained when ethidium bromide stained DNA in the gel is nicked by irradiation with 1500 μ J of short wavelength UV light. Nicking of DNA by this procedure depends on the transfer of UV light energy to DNA via the bound ethidium bromide. Therefore, it is important to make sure that the dye has diffused uniformly through the entire thickness of the gel and bound all the available DNA.

Following nicking, the DNA in the gel was transferred to Nylon membranes (Biotrans, ICN, or Zetaprobe, Biorad) by vacuum blotting (Vacugene, LKB) in 0.4 M NaOH, or, after 30 min denaturation in 0.4 M NaOH, by capillary transfer in 10xSSC (Sambrook *et al.*, 1989). After transfer, the membranes were rinsed with 2xSSC.

In a separate experiment, the effect of baking the membrane in a vacuum oven after transfer was compared to crosslinking the DNA to the membrane using 1200 μ J UV. Crosslinked membranes gave much stronger signals (data not shown).

Cloned inserts from the USDA RFLP map of soybean were radiolabelled with [³²P]-dATP using a random hexamer primer kit (Boehringer Mannheim). Membranes were prehybridized for 5 min at 60°C in 7% SDS, 0.25 M Na₂HPO₄, and hybridization was performed in a fresh volume of the same solution for at least 5 hours. Washes were carried out at 60°C in 5% SDS, 20 mM Na₂HPO₄, two times for 20 min, and for two more 20 min

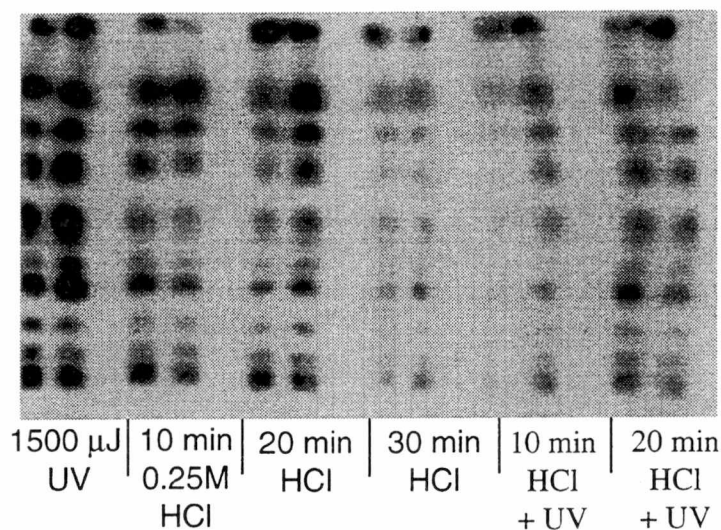


Figure 5.3. Optimization of conditions for depurination and UV nicking of large DNA fragments prior to Southern blotting. A gel containing 12 lanes of resolved yeast chromosomes was stained with 5 μ g/ml ethidium bromide and destained for 20 minutes in distilled water. The gel was cut into six slices containing two lanes of yeast chromosomes and each slice was treated as indicated below the photograph of the gel. UV nicking was performed in a Stratalinker 1800 (Stratagene). The slices were replaced in the gel casting tray and joined with molten agarose. The entire gel was vacuum transferred (Vacugene, LKB) in 0.4M NaOH onto a Biotrans membrane (ICN). The blot was probed labelled total yeast DNA.

periods in 1% SDS, Na₂HPO₄. Films were exposed to Kodak X-Omat AR film for 3 to 6 days at -70°C with an intensifying screen.

Figure 5.4 shows a pulsed field gel containing various digests of soybean chromosomal DNA run under conditions to resolve fragments of DNA up to 2 Mb in size. Ethidium bromide staining of the gel showed that most of the DNA in a plug not digested with a restriction enzyme did not enter the gel. A small proportion of it leaves the plug and is concentrated in the compression zone, which contains fragments of DNA larger than can be resolved under these conditions of electrophoresis (>2Mb). Many of the enzymes result in characteristic restriction fragment patterns when the ethidium bromide stained gels are viewed under UV light. This banding is a useful indicator of completeness of digestion. Most of the prominent bands in the region below 100 kb probably originate from organellar (mostly chloroplast) DNA, which has a high copy number in the cell and is not subject to methylation (Ganal *et al.*, 1989).

*Mlu*I digests show a striking ladder of bands with an interval of about 50 kb. This indicates that there are regions of the genome where unmethylated *Mlu*I sites are distributed at regular intervals. A similar phenomenon has been seen in cereal genomes (Moore *et al.*, 1993). It is not known what their significance to genome organization is.

The gel in Figure 5.4 was probed consecutively with the markers pA-703 and pA-69. In most digests, the probes detect fragments ranging from 50-300 kb. The *Sfi*I fragments homologous to these probes are beyond the resolution of this gel, and hybridization is to the compression zone. To improve resolution in the size range 50-300 kb, the switch interval during electrophoresis was shortened.

Figure 5.5 shows a gel run under conditions to resolve fragments up to 400 kb in size. After transfer, the membrane was probed consecutively with pK-9 and pA-69. The hybridizations show the effect of incomplete digestion. In lane 2, a *Sma*I digest, the majority of ethidium bromide fluorescence is seen in the compression zone rather than down the length of the lane, and the distinct bands below 100 kb are absent. All the hybridization is to the well and to the compression zone, where fragments larger than 400 kb are focused. The autoradiogram in Figure 5.4 showed that a complete *Sma*I digest should produce homologous fragments ranging in size from 100 to 300 kb. A prominent band is seen in the stained gel between 100 and 150 kb, and this may represent chloroplast DNA cut at a single site.

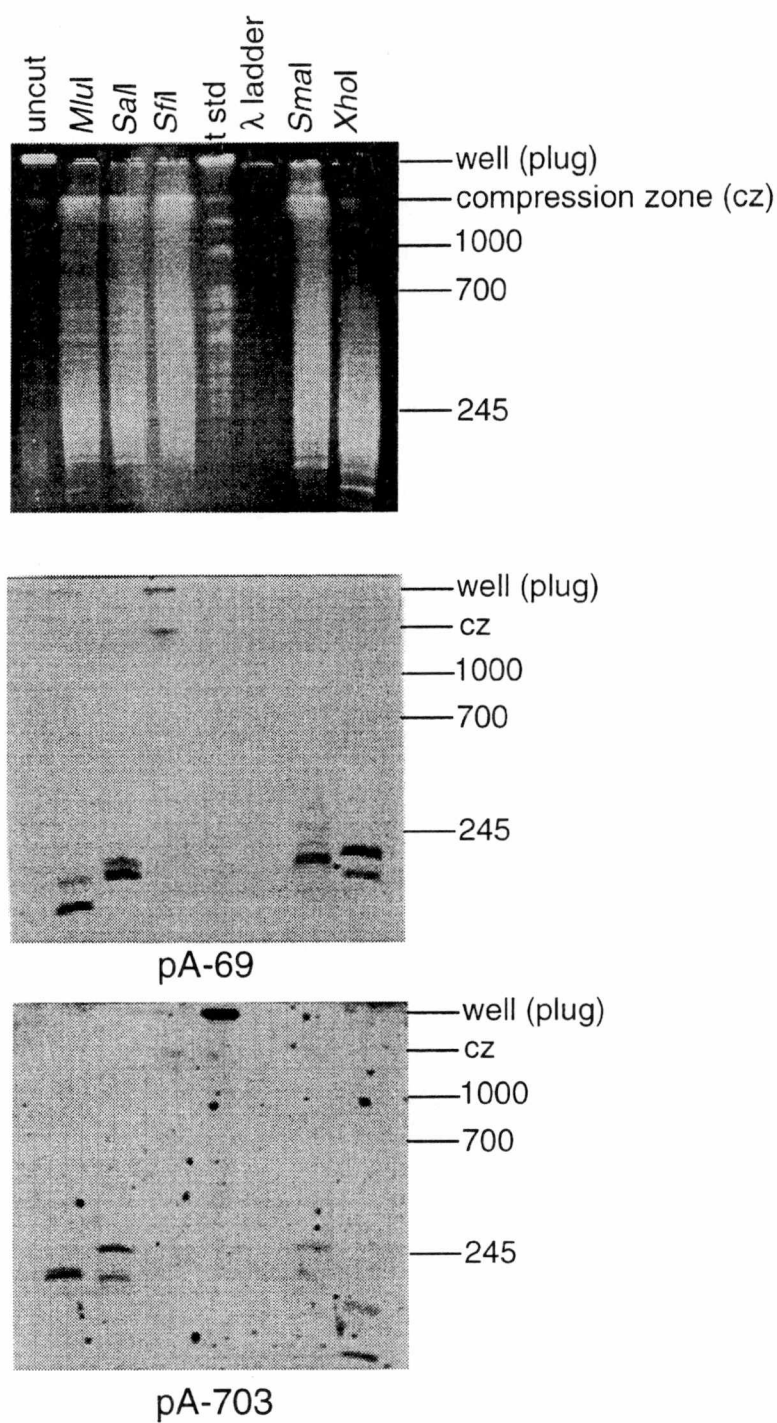
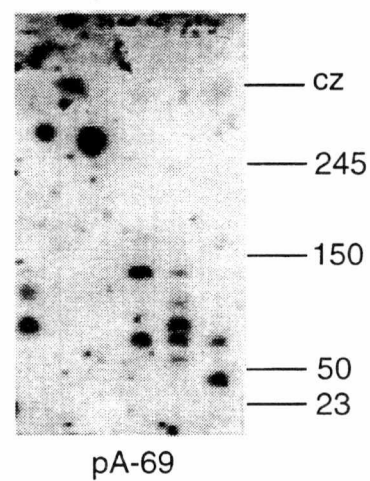
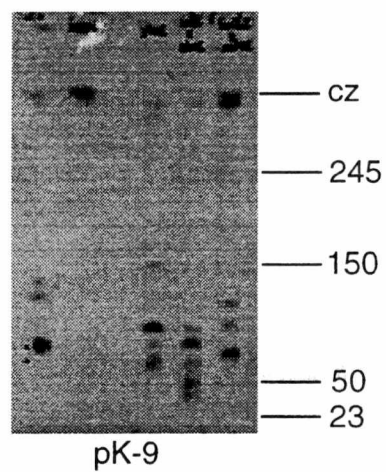
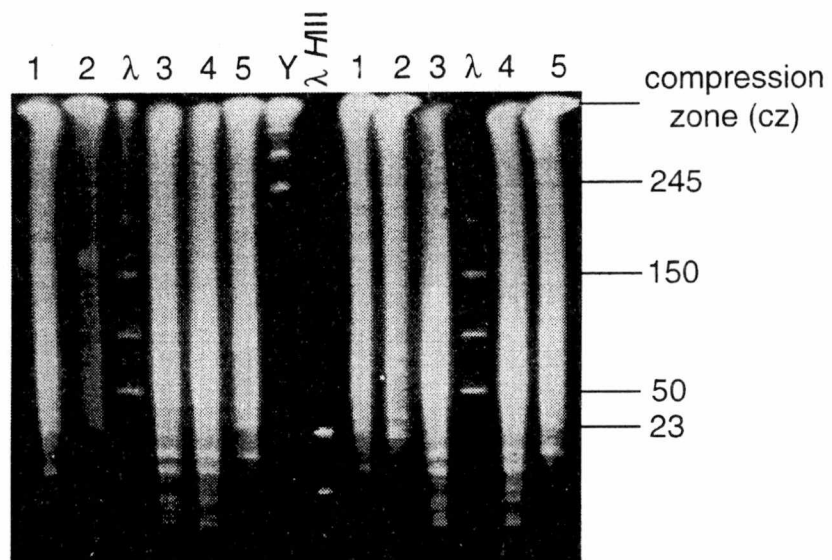


Figure 5.4. Pulsed field gel run under conditions to resolve fragments up to 2Mb in size: 1% agarose, 200V, 60 sec switch interval for 15 hours followed by a 90 sec switch interval for 9 hours with 0.5x TBE. A Southern blot of the gel was probed sequentially with markers pA-703 and pA-69.

Figure 5.5. Pulsed field gel run under conditions to resolve fragments up to 400 kb in size and illustrating the use of appropriate size standards to cover the whole range of resolution: 1% agarose, 200V, with a switch interval ramped from 5 to 20 sec over a 20 hour period in 0.5x TBE. 1= *SaII*, 2=*SmaI* (this digest is incomplete, the single sharp band that appears may be linearized chloroplast DNA), 3= *XhoI*, 4= *SaII* x *XhoI* double digest, 5= *SmaI* x *MluI* double digest. Size standards are λ concatemers (λ), yeast chromosomes (Y), and a λ *HindIII* digest (λ *HIII*). A Southern blot of the left side of the gel is shown probed sequentially with markers pK-9 and pA-69.



EFFECTS OF METHYLATION AT RESTRICTION ENZYME TARGET SITES

The DNA of higher plants has a high content of m^5C , and although this modification occurs at the dinucleotide C-G (which is the major methylated site in animal cell DNA), in fact all the methylated cytosines in plant DNA are part of the basic trinucleotide C-X-G (Gruenbaum *et al.*, 1981). The presence of 5-methylcytosine in the target sequence can prevent hydrolysis by restriction enzymes (McClelland and Nelson, 1988). Table 5-1 shows the methylation sensitivity of the restriction enzymes that were used in these experiments.

The effects of methylation can be of benefit to physical mapping experiments. Flanking markers sometimes comigrate on partial digestion products and yield data that could not be gleaned from complete digests (see Chapter 6). However, the benefits of partial digestion are best enjoyed when one is working with a single copy probe, where the explanation for more than one band in a hybridization pattern is relatively unambiguous.

This is illustrated by the blots in Figure 5.4. The marker pA-69 occurs twice in the soybean genome, as determined by hybridization to digests of DNA by methylation-insensitive enzymes (data not shown; the cloned fragment has no internal sites for the enzymes used here). In both the *MluI* and *SalI* lanes, two strongly hybridizing bands are seen, which probably but not necessarily represent the complete digestion products recognized by the probe. In addition, a larger fainter band is seen in each lane. This is a partial digestion product encompassing one or both of the smaller fragments, depending on whether the two copies of the marker are arrayed in tandem in the DNA or are located on different chromosomes.

The degree of methylation of sequences in the genome can vary between different cell types, and also between batches of DNA plugs prepared at different times. Finally, the extent of methylation may vary from one part of the genome to another. Figure 5.6 shows Southern blots of *SfiI* digests of three batches of DNA plugs prepared over a six month period. The blots were probed with the marker pA-36, which is present in two copies in the genome (see Chapter 6). In the first Southern, very little hybridization to the compression zone is seen, and a very intense single band in the *SfiI* digest is detected at 750 kb. In the second, the single band at 750 kb still hybridizes, but the majority of the signal is seen in the compression zone. This indicates that the DNA fragments homologous to the probe are all much larger than 750 kb, and that in this batch of plugs *SfiI* was unable to cut many of

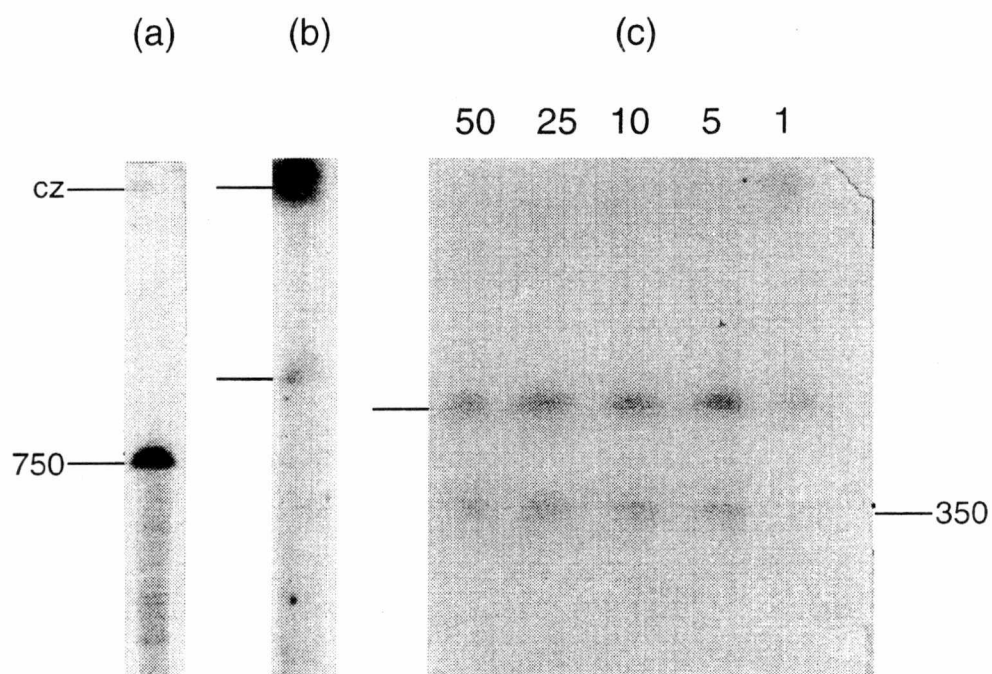


Figure 5.6. Southern blots illustrating variable degrees of methylation and completeness of digestion between different batches of DNA plugs. A different batch of plugs was used in each experiment. In a and b, 50 units of *Sfi*I were applied to a plug in a standard overnight digestion. In c, the indicated number of units of enzyme were applied to plugs for 5 hours. The gels used to produce the Southern blots were run under different conditions of resolution, and therefore the 750 kb fragment is detected at different distances from the compression zone in panels a, b, and c.

the sites surrounding the marker. Finally, in the third Southern, a second band hybridizing to pA-36 appears at 350 kb. When the DNA in this batch plugs was treated with decreasing amounts of *Sfi*I in partial digests, the smaller band continued to be produced, suggesting that an *Sfi*I site that was decisively methylated in the earlier batches of plugs was completely accessible to the enzyme in the third batch.

Batch to batch variation in the methylation status of DNA in plugs can be advantageous because of the extra data that can be gathered. However, the benefits should be exploited warily, and separate restriction maps of the region of the genome that is being studied should be constructed from each batch before they are compared.

These examples illustrate one of the challenges of mapping a complex genome like that of soybean. The presence of genes or markers with more than one copy makes it impossible to say whether multiple bands on pulsed field gel Southern are due to partial digestion or how many copies are present in each band. To correctly interpret a pulsed field gel Southern, it is usually necessary to perform second dimension electrophoresis (2D-DE; see Chapter 6).

DNA CONCENTRATION EFFECTS ON THE APPARENT SIZE OF HYBRIDIZING FRAGMENTS

The concentration of DNA in the original plug can affect the apparent size of fragments detected later by hybridization. In general, high concentrations of DNA in a lane will decrease the mobility of migrating fragments. If the original concentration of DNA in plugs is different, this can lead to different estimates of the size of a given fragment when it is measured from different gels.

For the purpose of accurately measuring the size of a fragment, therefore, it is important to limit the DNA concentration to the point where these artefacts are eliminated (about 80 µg/ml; Doggett *et al.*, 1992). In these experiments the calculated DNA concentration was slightly higher than this, partly because of the need to keep the nuclear DNA concentration high enough to permit detection of single copy sequences in the soybean genome on Southern blots. The presence of organellar and in particular chloroplast DNA increased the DNA load by about 10-15%.

A Southern blot containing *Not*I, *Sfi*I, and a *Not*I/*Sfi*I double digest was probed with the marker pA-89. The result illustrates the artefacts that can be created by high DNA

concentrations, and is shown in Figure 5.7. The smallest hybridizing fragment in the double digested lane appears to be larger than the smallest fragment detected in the *NotI* lane. This apparently paradoxical result is a DNA concentration artifact. The fragments are probably identical (i.e. there is a *NotI* fragment that is not shortened by *SfiI*). It also illustrates that it is not only the starting concentration in the plug that is important, but the proportion of DNA in the plug that migrates in a given size range. In the double digest, a greater mass of small fragments is produced, causing a general retardation of those fragments compared to their counterparts in the single digests.

Because mobility is so sensitive to DNA concentration, it is important for physical mapping projects to reprobe the same membrane to determine comigration of markers on a given fragment.

COMIGRATION OF MARKERS IN THE CLUSTER

Figure 5.8 shows Southern blots probed with the markers pA-36, pK-9, pA-89, and pA-69. These blots contain fragments less than 200 kb in size produced by the enzymes *MluI*, *SalI*, *XhoI* in single and double digests. A comparison of the pA-89 and pK-9 blots shows that copies of these markers comigrate on *MluI* and *SalI* fragments (indicated with asterisks, lanes 1 and 3), as well as on fragments produced by double digestion with *MluI*×*XhoI* (lane 5), *MluI*×*SalI* (lane 6), and *MluI*×*SfiI* (lane 8). The bands in lane 7 also coincide, but this is a repetition of the *SalI* pattern in lane 3 because *XhoI* enzyme failed to cut the DNA in this particular double digest. The coincidence of the hybridization patterns in lanes 1 and 5 suggests that copies of these two markers must also reside on the same *XhoI* fragments, because comigration on a fragment produced by digestion with two restriction enzymes is not possible unless copies of the markers also comigrate on the same fragments produced by the action of each enzyme separately. Unfortunately the signals in the *XhoI* lanes are poor. Because there are two copies of marker pA-89 in the genome, and at least four copies of marker pK-9 (see Chapter 6), it is not possible to say which of the restriction fragments correspond to linkage group H. However, none of the fragments is much larger than 100 kb, and there is no hybridization to the compression zone. Therefore the physical distance separating these markers is not more than about 100 kb. Comigration of markers that show complete linkage, like pA-89a and pK-9d in linkage group H, is not very informative, although it can give an estimate of the maximum distance by which markers can be separated without undergoing meiotic recombination (i.e. it may disclose areas of suppressed recombination).

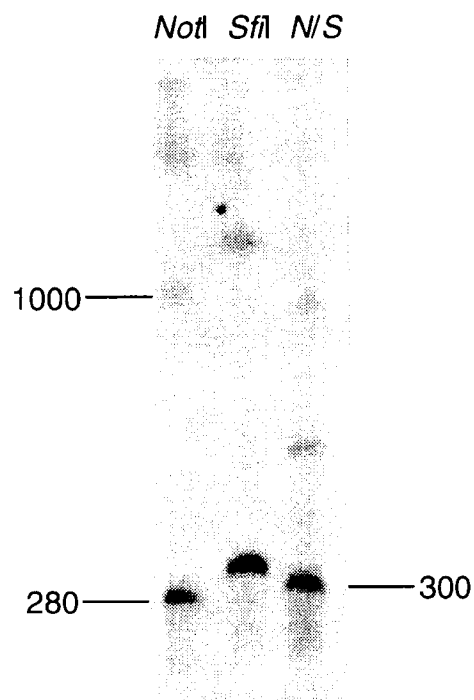
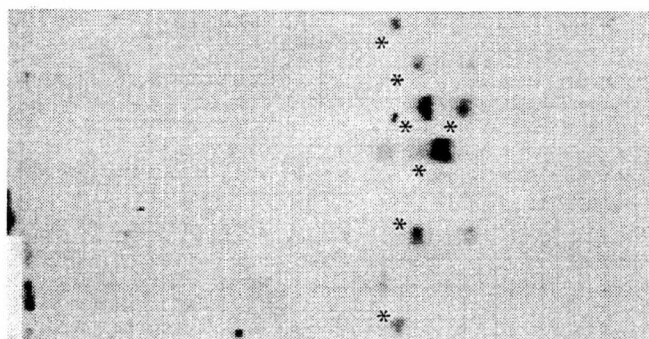


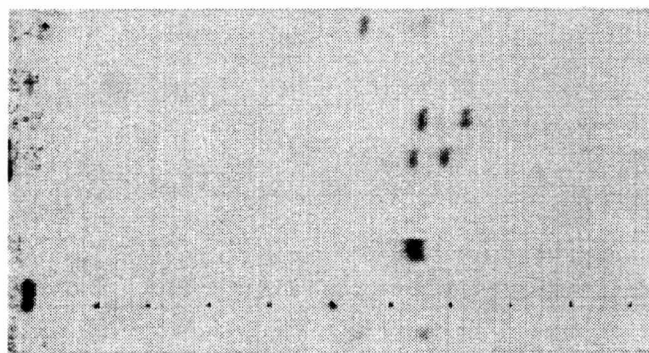
Figure 5.7. Southern blot demonstrating the effect of DNA concentration on fragment mobility in pulsed field gels. The gel was 0.8% agarose and was run at 200V for 15 hours with a 60 sec switch interval and for 9 hours with a 90 sec switch interval in 0.5x TBE. After transfer the blot was probed with marker pA-89. The fragment detected at 280 kb in the *NotI* digest experienced an apparent increase in size to 300 kb when the *NotI* digest was followed by an *SfiI* digest.

Figure 5.8. Southern hybridized with the markers in the pA-36 cluster. Markers in the cluster with measurable recombination distances between them do not comigrate on fragments in the size range of 50 to 100 kb and must therefore be separated by physical distances larger than this. Two gels were run under identical conditions: 1% agarose, 200V, with a switch interval ramped from 5 to 20 sec over a 20 hour period in 0.5x TBE. Both were transferred. One Southern was used to produce the pA-89, pA-36 and pA-69 blots, while the other was probed with pK-9. 1=*Mlu*I, 2=*Xho*I, 3=*Sal*II, 4= λ , 5=*Mlu*I x *Xho*I double digest, 6=*Mlu*I x *Sal*II, 7=*Xho*I x *Sal*II, 8=*Mlu*I x *Sfi*II. Lanes with *Xho*I show evidence of incomplete digestion, including strong hybridization in the well. In *Sal*IIx*Xho*I double digests, the hybridization pattern is identical to that produced by *Sal*II acting alone, showing that *Xho*I did not cut the DNA.

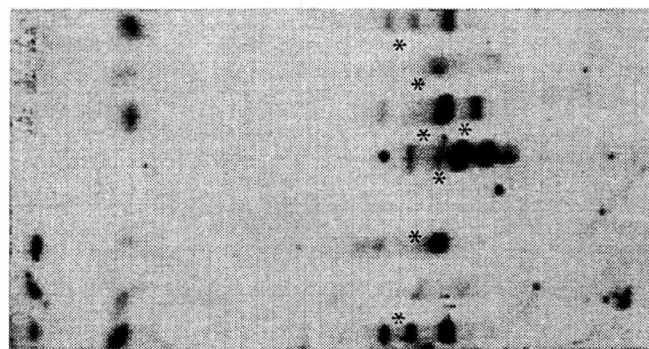
1 2 3 4 5 6 7 8



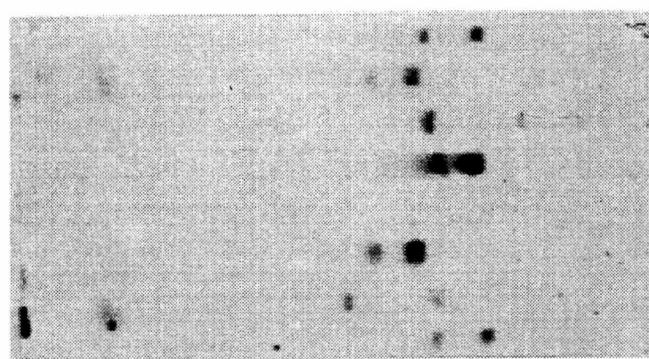
pA-89



pA-36



pK-9



pA-69

None of the markers in the cluster that are separated by measurable recombination distances cohybridize to the relatively small fragments produced by these enzymes. From considerations of genome size and the number of centiMorgans on the soybean RFLP map, it is expected that much larger fragments would be needed to span the genetic distances between these markers. These can be produced with restriction enzymes such as *NotI* and *SfiI* that recognize 8 bp target sequences (Table 5-3; see Chapter 6).

The blots exhibit the difficulties of interpretation that have already been alluded to. For example, the single band detected by the marker pA-36 in *SaII* digests raises the question of whether there are two overlapping fragments of the same size in that band, or whether the two copies of the marker are duplicated in tandem on an 80 kb *SaII* fragment. The use of second dimension electrophoresis to resolve questions like this is introduced in the next chapter.

CHAPTER 6*

PHYSICAL MAPPING OF A REGION IN THE SOYBEAN (*GLYCINE MAX*) GENOME CONTAINING DUPLICATED SEQUENCES

INTRODUCTION

Physical mapping in a complex genome such as that of soybean is complicated both by its large size and by the presence of duplicated chromosomal regions. Early in its evolution the soybean genome was duplicated by tetraploidization (Hadley and Hymowitz, 1973), and 36% of the genome has been estimated to consist of repetitive DNA (Gurley *et al.*, 1979). However, the pachytene karyotype of soybean shows that all chromosomes (20 in total) are visually distinct (Singh and Hymowitz, 1988). Genetic segregation of a number of morphological and developmental markers follows diploid genetics, suggesting that functional diploidization has occurred throughout much of the soybean genome (Lee and Verma, 1984). The ancestral tetraploidy of the genome is reflected by the fact that many markers on the RFLP map detect two or more independently segregating bands on Southern blots of electrophoretically fractionated DNA, suggesting that duplicate copies of the markers exist on what may be different homoeologous chromosomes (Keim *et al.*, 1989). Shoemaker *et al.* (1992) searched for conserved duplicate linkage of markers while constructing the soybean RFLP map and detected presumed homoeologous regions, albeit with internal linkage disparities. In analyzing Southern blots of pulsed field gels containing fractionated soybean DNA, it is usually not possible to assign a specific copy of a marker to given large fragment, or indeed to be sure of the number of copies of a marker that comigrate to the same position on the gel. Chromosome walking to a gene of interest may be complicated by the presence of multicopy markers and repeated elements because of the danger of being diverted to a region of the genome that contains similar sequences but that are not physically contiguous to the starting point of the walk.

In order to develop an approach to physical mapping in soybean we attempted to construct a partial map of a cluster of markers in linkage group H of the USDA-ARS/ISU soybean RFLP map (Keim *et al.*, 1989). To do this, we used pulsed field gel electrophoresis (PFGE; Schwartz and Cantor, 1984) to detect comigration of markers in the cluster on very large DNA fragments.

*Funke, R.P., Kolchinsky, A., and Gresshoff, P.M. (1993). *Plant Mol Biol* 22, 437-446.

All the markers in this cluster have at least two copies in the genome and hybridize to more than one large restriction fragment. The question arises which fragment detected by the probes is colinear with linkage group H on the RFLP map. In the present paper we demonstrate that by using a second restriction and a second dimension conventional electrophoresis following PFGE, individual duplicated markers may be assigned to specific large DNA fragments.

MATERIALS AND METHODS

Procedures for the preparation and restriction digestion of intact chromosomal DNA in agarose plugs were described in detail in Chapter 5.

For two dimensional electrophoresis, twice as much DNA (at the same concentration) was used in the first dimension PFGE. After staining with ethidium bromide, lanes of fractionated DNA were cut out of the gel and equilibrated on ice with the buffer appropriate for the enzyme (*Hind*III or *Dra*I) to be used in the second dimension for one hour. About 200 units of the enzyme were added and allowed to diffuse into the slice for an additional hour before placing the digests at 37°C overnight. The slices were briefly (15-20 minutes) equilibrated with running buffer. They were then placed in a casting tray perpendicular to the direction of electrophoresis and embedded in a large gel. A normal plug of soybean DNA not fractionated by PFGE and digested with the same enzymes (*Hind*III or *Dra*I) was loaded alongside the slice. Following electrophoresis, gels were blotted and probed as described above. A schematic representation of the procedure is shown in Figure 6.1.

RESULTS

Linkage group H (Figure 6.2) contains the pA-36 cluster of markers that we chose to analyze. The pA-36 polymorphism with *Hind*III between *G. max* (L.) Merrill cv. Bragg and *G. soja* PI468.397 (the parental lines used to map the *nts* locus) shows that this marker has two copies in the genome, one monomorphic (9 kb) and one polymorphic (3.5 kb in *G. max*) (Figure 6.3). The probe has no internal sites for *Hind*III. The polymorphic copy is 10 to 16 cM distant from both the single copy marker pUTG-132a and the *nts* locus in linkage group H (Landau-Ellis *et al.*, 1991; Figure 6.2). Because this band was mapped in our study populations, it served as an "anchor" in this cluster of markers. Segregation of the other duplicated markers in this cluster was not checked in our populations, and we

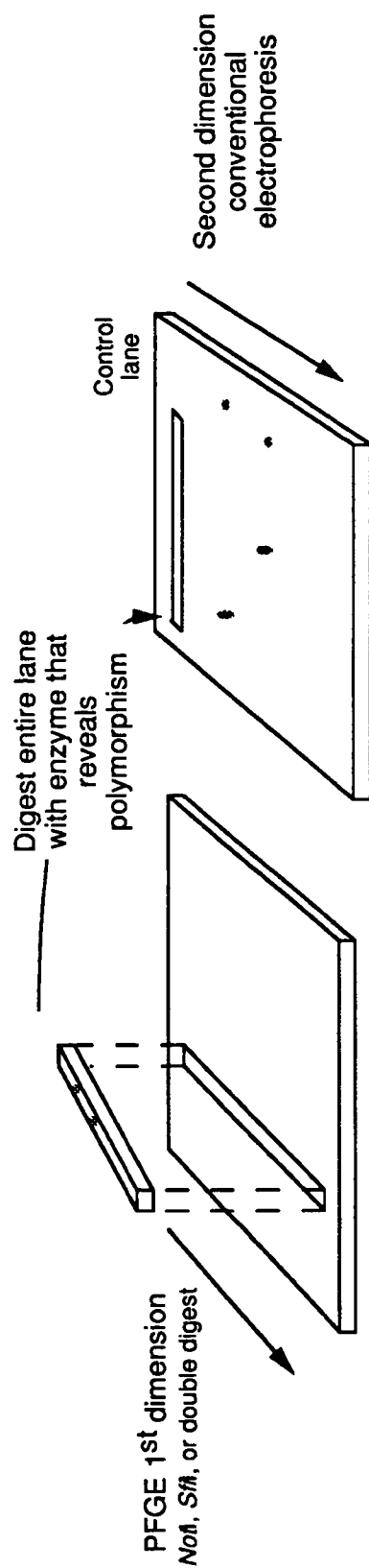


Figure 6.1. Schematic of the procedure for two-dimensional electrophoresis of soybean DNA. This procedure was used to assign individual copies of the duplicated markers in linkage group E to specific large DNA fragments.

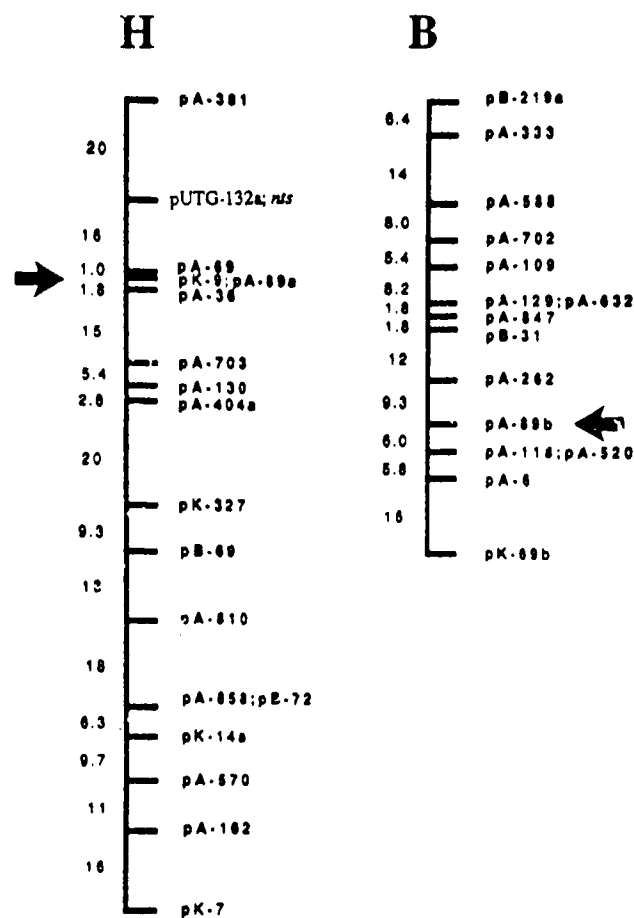


Figure 6.2. Linkage groups H and B of the RFLP map of soybean (Keim *et al.*, 1989). The pA-36 cluster of markers is located on linkage group H, about 18 cM (or 10 cM; Landau-Ellis *et al.*, 1991) away from marker pUTG-132a and the *nts* locus.

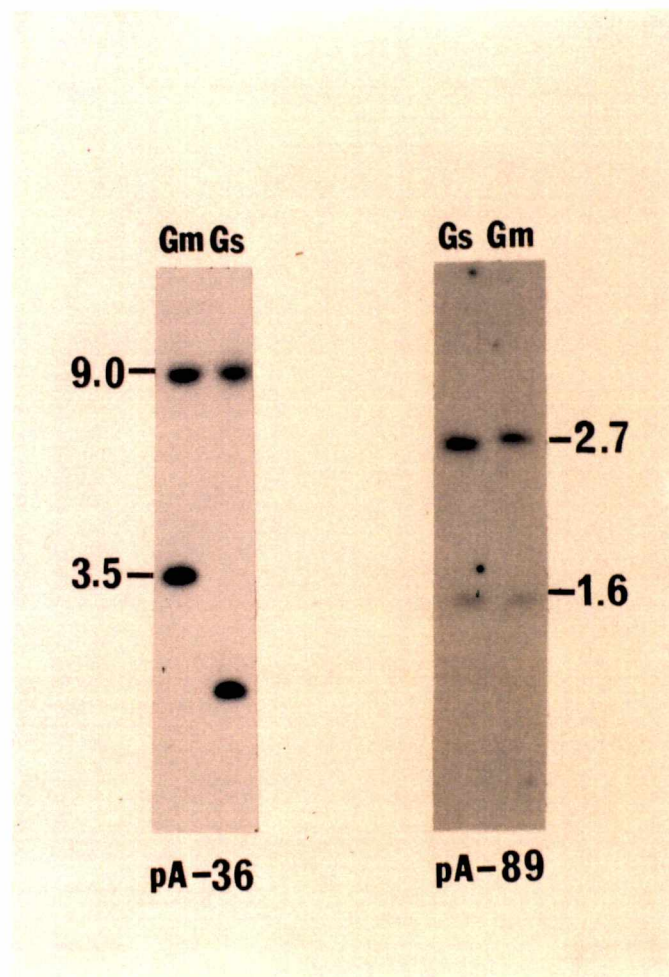


Figure 6.3. pA-36 polymorphism between *Glycine max* (L.) Merr. cv. Bragg and *G. soja* PI468.397, and absence of *Dra*I polymorphism for pA-89 in these parental populations. Sizes are indicated in kb. See text for discussion.

were able to state which copies mapped linkage group H only by virtue of their comigration on pulsed field gels with the polymorphic copy of pA-36. The marker pA-89 has two copies in the genome. In the lines used to construct the ARS/ISU map, polymorphisms for both copies (a and b) were revealed by *Dra*I and mapped to different linkage groups (Figure 6.2; T. Olson, pers. comm.). *Dra*I reveals no polymorphism with pA-89 for the parental populations used in this study. The probe hybridizes strongly to a 2.7 kb *Dra*I fragment and a less intense band is detected at 1.6 kb (Figure 6.3). We cannot be certain which of these corresponds to the copy in linkage group H (pA-89a) and which to the one in linkage group B (pA-89b), although we make a prediction on the basis of our PFGE analysis.

Throughout this work, we make the assumption that in spite of the occurrence of polymorphisms between the parental line used for the construction of the ARS/ISU maps and our lines of *G. max* and *G. soja*, gene order and the chromosomal location of molecular markers is conserved (Shoemaker *et al.*, 1992).

The probe pK-9 has an even more complicated hybridization pattern and is represented at least four times in the genome. For this reason we will restrict most of our discussion of cohybridization to the relatively simpler markers (pA-36 and pA-89 in particular), and will not present any two dimensional electrophoresis data for pK-9. Only one copy of this marker, pK-9d, was placed on the RFLP map, in the cluster that we chose to study. This copy of the marker was inseparable from pA-89a by segregation analysis (Figure 6.2). Another marker in this cluster, pA-69, is separated by only 1 cM from pK-9d and pA-89a. However, it never hybridized to large DNA fragments revealed by the three other markers in the cluster, and will not be discussed further.

Identification of Large DNA Fragments Hybridizing to Several RFLP Probes. In describing these data we make a clear distinction between the terms 'fragment' and 'band'. A band on a pulsed field Southern blot of soybean DNA may contain more than one fragment type. Considerable hybridization to the compression zone is often seen on pulse field Southern blots. This region of the lane contains DNA fragments that are too large to be resolved, either due to incomplete digestion or the absence of restriction sites in the DNA surrounding one or more copies of the marker. Preliminary results showed that with hexanucleotide recognizing restriction enzymes such as *Mlu*I, *Sal*I, and *Xho*I the majority of fragments identified by our probes range in size from 50 to 150 kb. Using single and double digests, no comigration of the markers in the cluster was obtained on fragments of this size. Since there are 1.07×10^9 bp in the soybean haploid genome (Blackhall *et al.*, 1991),

and currently the RFLP map contains 2466 cM (Keim *et al.*, 1989) there are an average of about 450 kb of DNA per cM. Therefore, in order to obtain cohybridization of two or more markers in the pA-36 cluster larger fragments would be needed. Using the octanucleotide recognizing enzymes *NotI* and *SfiI* fragments ranging in size from 280 kb to more than 1.0 Mb were detected by probes pK-9 and pA-89, and pA-69, respectively. Three of the markers, pK-9, pA-89, and both copies of pA-36 (there is no hybridization to the compression zone with this probe) comigrate in the same 1000 kb *NotI* band, and pK-9 and pA-89 in addition share 280 kb *NotI* and 350 and 650 kb *SfiI* bands (Figure 6.4a).

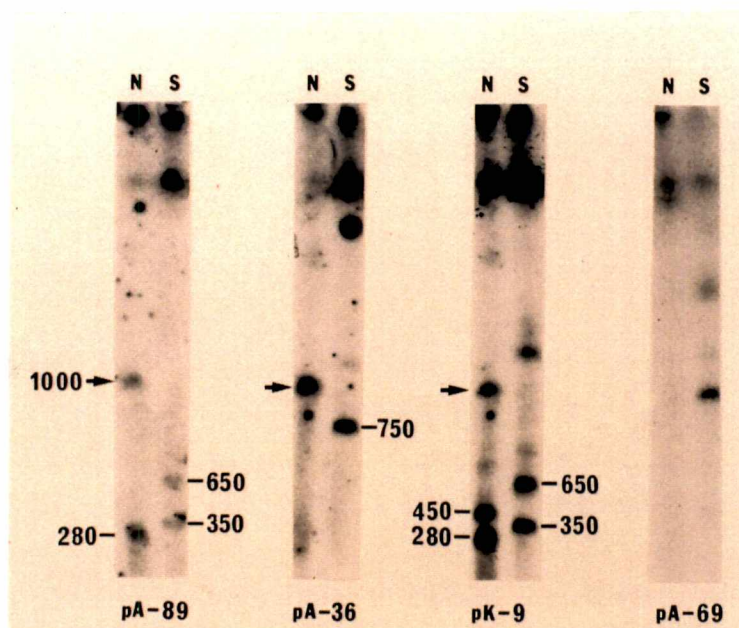
When we repeated the experiment with a different preparation of plugs, *SfiI* digests revealed an additional 350 kb band upon probing of *SfiI* blots with pA-36, as compared to the single 750 kb band seen before (Figure 6.4b). This band also contained one or more copies of pK-9. In *NotI/SfiI* double digests, a still smaller band (200 kb) appeared which contained copies of both pK-9 and pA-36 (this blot was not rehybridized with the other markers in the cluster; see below). We concluded that this depended on the variable digestion of the DNA, possibly caused by the methylation status of the tissue from which the plugs were prepared. In different experiments the bands were always the same size (750 kb and 350 kb in *SfiI* digests, with an additional 200 kb band appearing in double digests). Which band pattern we obtained did not depend on enzyme concentration. Plugs yielding a single band did so even when incubated overnight with 50 units of enzyme, whereas in material where we detected two bands as little as 1 unit of enzyme was sufficient to reveal them on Southern hybridization (data not shown).

The 1000 kb *NotI* band contained copies of three of the markers in the cluster (all except pA-69). In *SfiI* digests the copies of pA-36 migrated in separate 750 kb and 350 kb bands. The 350 kb band also contained copies pA-89 and pK-9. In double digests the physical linkage between copies of pA-36 and pK-9 appeared to be preserved in the 200 kb band, which does not contain pA-89. Also in the double digest a 970 kb band which seemed to be derived from the 1000 kb *NotI* band contained copies of both pA-36 and pA-89.

It now became important to resolve which large *SfiI* band contained the polymorphic copy of marker pA-36, and which band corresponded to the monomorphic copy. Using two-dimensional electrophoresis, we were able to assign each copy unambiguously to large pulsed field gel bands.

The 1000 kb *NotI* band contained both genomic copies of pA-36 (Figure 6.5a). A Southern of an *SfiI/HindIII* two-dimensional gel (Figure 6.5b) allowed two important conclusions to

a)



b)

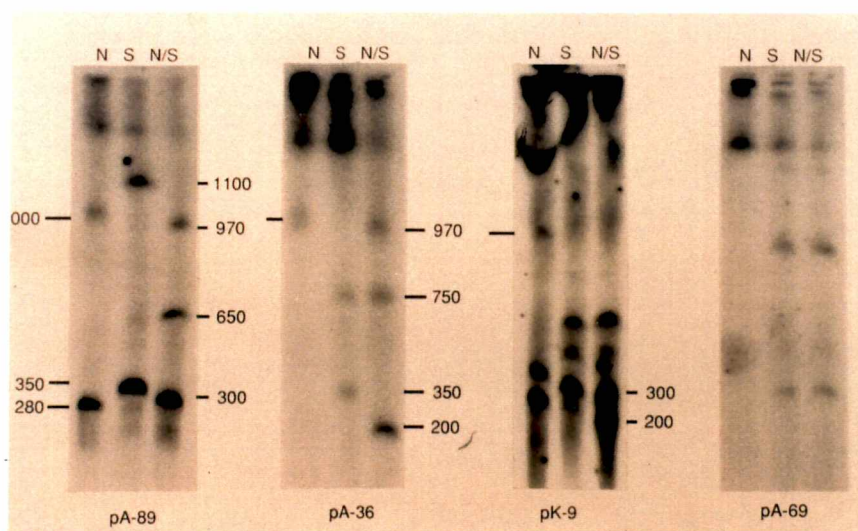


Figure 6.4. Southern blots of pulsed field gels carrying digests of different batches of *G. max* DNA. a) Southern blots produced from plugs showing a single band in *Sfi*I digests when probed with pA-36. The same blot was probed recurrently. b) Southern blots produced from plugs showing two bands in *Sfi*I digests probed with pA-36. pA-36 and pA-89 Southern blots represent two halves of a membrane blotted from the same pulsed field gel. The blot hybridized in the first round with pA-36 was reprobed to produce the pA-69 Southern. A second gel was run to produce the pK-9 Southern, and slight differences in conditions of electrophoresis are responsible for small differences in mobility between corresponding bands on this blot and on the other two. N=NotI, S=*Sfi*I, CZ=compression zone.

a)

b)

c)

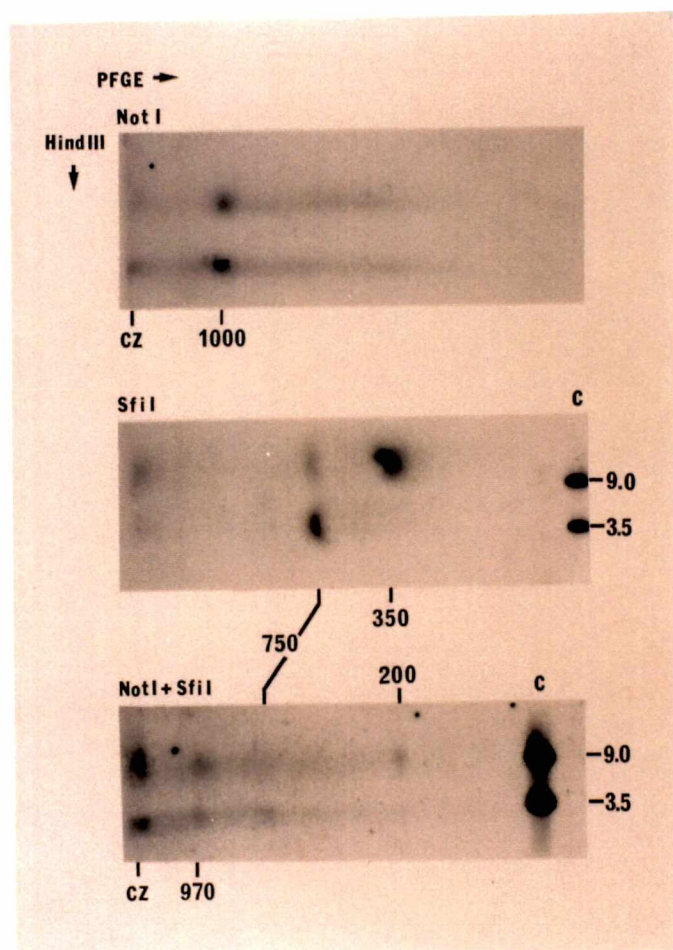


Figure 6.5. Two dimensional electrophoretic analysis of marker pA-36. Sizes for both the first and second dimensions are indicated in kb. a) *NotI* first dimension, *HindIII* second. b) *SfiI* first dimension, *HindIII* second. c) *NotI/SfiI* first dimension, *HindIII* second. In b) and c) a DNA plug digested with *HindIII* alone was electrophoresed alongside the sample in the second dimension to indicate the size of the two fragments detected by pA-36 on a Southern blot.

be drawn. First, only the monomorphic copy of pA-36 (at 9 kb) was located in the 350 kb band. Traces of it were also seen in the 750 kb band, which therefore contained a partially digested fragment with this copy of the marker on it. The partial *SfiI* digest proved that the copies of pA-36 were not tandemly duplicated on a single *SfiI* fragment. If they were, the appearance of the monomorphic copy on the 350 kb band should have been accompanied stoichiometrically by the migration of the other copy to 400 kb on pulsed field gels. Therefore, the 750 kb band contained two fragments, one of which is a partial digestion product. When we repeated this experiment, a more complete *SfiI* digest in the first dimension removed all traces of this copy of the marker from the larger band. The polymorphic copy (represented by the 3.5 kb *HindIII* fragment) was located exclusively in the 750 kb band. The two hybridization patterns obtained with pA-36 are explained by the failure of the second *SfiI* site near the monomorphic copy of the marker to be cleaved in some batches of plugs.

Extending this argument, it appeared likely that two separate fragments were also present in the 1000 kb *NotI* band (the alternative, that there is a single fragment with an internal *SfiI* site, is ruled out by the data in Figure 6.5b). This means that the two copies of pA-36 may be located in different parts of the genome, and that there is a high degree of conservation of 'rare cutter' restriction sites in the DNA surrounding them.

The Southern blot of an *SfiI-NotI/HindIII* two-dimensional gel further supported the interpretations obtained with single digests (Figure 6.5c). The *NotI* band (slightly shortened by *SfiI* and presumably also containing two separate fragments) contained both copies of pA-36, while the 750 kb *SfiI* fragment (which apparently has no internal *NotI* sites) contained only the polymorphic copy. The 350 kb *SfiI* fragment containing the monomorphic copy was shortened to 200 kb by treatment with *NotI*.

Analysis of pA-89. On PFGE Southern blots pA-89 comigrated with pA-36 in the 1000 kb *NotI* band, as well as in the 350 kb *SfiI* band. In *SfiI-NotI* double digests both markers detected a band at 970 kb, strengthening our other comigration data (Figures 6.4a and b).

In two-dimensional analysis of pA-89, partial products of *DraI* digestion are seen in the second dimension because the efficiency of restriction of agarose-embedded DNA is low with *DraI* (New England Biolabs Catalog, 1992). From a blot of a *NotI/DraI* two-dimensional gel (Figure 6.6a), it can be seen that the 280 kb *NotI* band contains only the smaller copy of the marker. Spots corresponding to the larger *NotI* band are not seen,

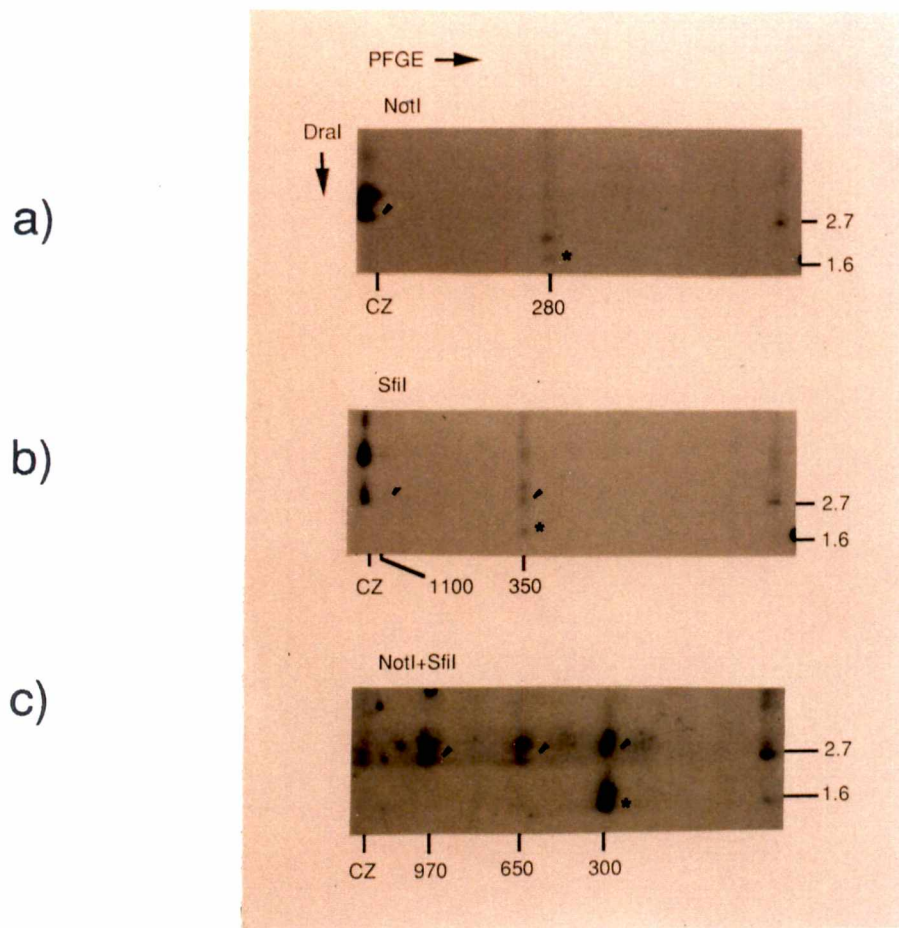


Figure 6.6. Two dimensional electrophoretic analysis of pA-89. Sizes for both the first and second dimensions are shown in kb. The *NotI-SfiI* digest (c) was run farther in the first dimension than the single digests in a) and b) in order to resolve the larger PFGE fragments better. a) *NotI* first dimension, *DraI* second. b) *SfiI* first dimension, *DraI* second. c) *NotI-SfiI* first dimension, *DraI* second. A DNA plug digested with *DraI* alone was electrophoresed alongside the sample in the second dimension to indicate the size of the two fragments detected. The faintly hybridizing 1.6 kb fragment is seen in the control lane in c, but not in a and b. This is because restriction of agarose-embedded DNA by *DraI* is quite inefficient (New England Biolabs Catalog, 1992), a problem that is more severe in the original DNA plugs than in lanes of fractionated DNA. An asterisk indicates the terminal digestion product of the 1.6 kb fragment, and an apostrophe indicates the terminal digestion product of the 2.7 kb fragment.

possibly because they are not completely resolved from the rather wide compression zone remaining after digestion with *NotI* in the first dimension. Only the 2.7 kb copy of the marker is found in the compression zone. *DraI* treatment of a lane containing fractionated *SfiI* products shows that both copies of the marker are present in the 350 kb band, while the 1100 kb band and the compression zone contain only the copy corresponding to the 2.7kb *DraI* fragment (Figure 6.6b). The analysis of a *NotI/SfiI* double digest shows that the 970 and 650 kb bands contain the copy of pA-89 represented by the 2.7 kb *DraI* fragment, while the band at 300 kb contains both copies of the marker (Figure 6.6c). The fact that both copies of pA-89 are present in the smaller bands on pulsed field Southern, as well as the greater retention in the compression zone of the copy represented by the 2.7 kb *DraI* fragment, account for the greater hybridization intensities of the shorter bands. According to the ARS/ISU RFLP map, the two copies of pA-89 are located in different linkage groups, and their coincidence in the same band on pulsed field gels again suggests that these bands contain separate fragments corresponding to each copy.

Thus the two copies of this marker, which have been unequivocally mapped to different linkage groups, also appear to have conserved long range restriction sites in their chromosomal surroundings.

These data allowed us to hypothesize which copy of the marker is pA-89a, on linkage group H, and which is pA-89b, mapped to linkage group B. Two-dimensional electrophoresis showed that only the copy of pA-89 represented by the 2.7 kb *DraI* fragment is found in the 1000 kb *NotI* and 970 kb *NotI/SfiI* band. We have explained our reasons for supposing that both the 1000 kb *NotI* and 970 kb *NotI/SfiI* bands contain two separate fragments, each with a copy of marker pA-36. Theoretically, the copy of pA-89 corresponding to the larger *DraI* fragment could lie on either of the fragments in this band. However, none of the PFGE data support a model that physically links the polymorphic copy of pA-36 (on linkage group H) to the copy of pA-89 that is located on the smaller *DraI* fragment. On the basis of this analysis we tentatively designated the copy of pA-89 represented by the 2.7kb *DraI* fragment as pA-89a in linkage group H.

It is possible that the other copy of pA-89 (pA-89b) comigrates with the monomorphic copy of pA-36 on a fragment in the 350 kb *SfiI* band, suggesting that the proximity of these markers is conserved in linkage groups H and B.

Comigration of pK-9 with other markers in the cluster. Because of its multiple copy number in the genome (four, based on *DraI* and *HindIII* genomic Southern) we did

not analyze pK-9 using two-dimensional electrophoresis. Copies of pK-9 always comigrated with both copies of pA-89 in *NotI*, *SfiI* and double digests (Figures 6.4a and b), suggesting that their genetic inseparability in linkage group H is manifested by tight physical linkage there as well as in another homoeologous region of the genome, linkage group B. The fact that the unmapped monomorphic copy of pA-36 also shows comigration with a copy of pK-9 (*SfiI* and double digests; Figure 6.4) strongly suggests that all three markers in the cluster are physically linked in both linkage groups H and B. Moreover, the physical distance between them appears to be about the same (1000 kb or less). The PFGE data presented here allow us to make a map of the cluster in linkage group H (Figure 6.7).

DISCUSSION

The kind of analysis presented in this paper had a relatively late start in soybean because of the late development of a usable RFLP map. A previous report on PFGE analysis of soybean DNA was published by Honeycutt *et al.* (1992), who embedded soybean protoplasts from stem-derived suspension cultures and used single-copy nodulin clones to detect large DNA fragments.

The experiments presented here show that the interpretation of pulsed field Southern blots of plant genomic DNA can be complicated and ambiguous. A major reason for this is the high degree of partial digestion that is seen on some of the blots. Partial methylation of a specific sequence in the genome can cause incompletely digested high molecular weight DNA to lag in the compression zone, and may prevent certain sites from being cut at all. Thus in some batches of plugs (represented by the Southern in Figure 6.4a), an *SfiI* site in vicinity of the monomorphic copy of pA-36 is completely inaccessible, whereas in others it is cut virtually to completion (Figure 6.4b). The data suggest that the site in one of the homoeologous chromosomal regions revealed here (in linkage group B) is preferentially methylated.

The comigration of different copies of a marker in the same pulse field band, even when these copies have been shown to segregate independently (the two copies of pA-89), presents another possible pitfall in the interpretation of Southern blots. We propose that there is marked conservation of 'rare cutter' restriction sites in homoeologous regions of the genome, and that many bands on pulsed field Southern blots in fact contain separate fragments.

The 300 kb band recognized by pA-89 in double digested lanes, which is larger than the 280 kb *NotI* band, is probably an artifact caused by local overloading of DNA in the lane.

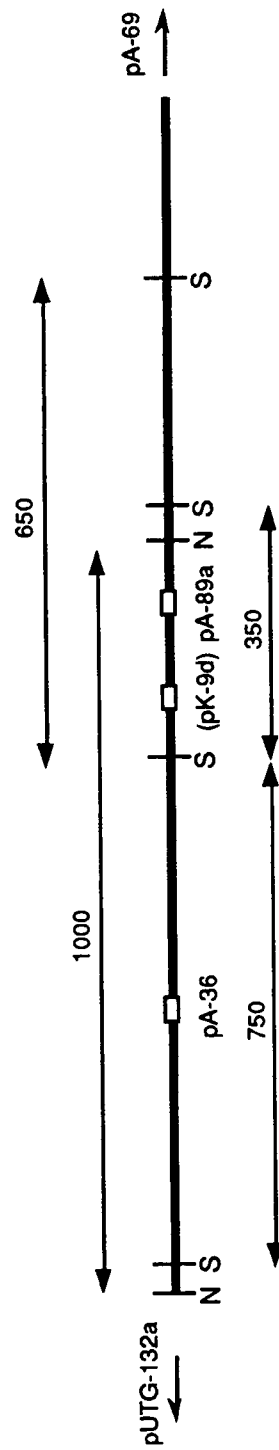


Figure 6.7. A physical map of the cluster of markers in linkage group E incorporating the results of *NorI* (N), *SfiI* (S), and double digests. Probe pK-9d is placed on the map based on the observation that a copy of this marker always comigrates with pA-89a. Sizes are in kb. The 1000 kb *NorI* fragment spans a recombination distance of 1.8 cM. The location of the markers within the fragments is random.

More DNA fragments of that size are produced as compared to the single digest, and it has been observed that the total concentration of DNA molecules of a certain size can affect their overall mobility in PFGE, leading to artifacts of this kind (Doggett *et al.*, 1992). We did not notice a consistent difference between plugs prepared from primary leaves and young trifoliates with respect to the pA-36 hybridization pattern, and did not investigate further or try to control the factors responsible for different methylation states of tissues used to prepare plugs. One way to overcome batch-to-batch variation is to prepare a large number of plugs from a single source tissue, and use the same batch for the duration of a single mapping experiment.

These data allow us to tentatively state which *Dra*I fragment corresponds to pA-89a and which to pA-89b, in linkage groups H and B, respectively. Two-dimensional electrophoresis showed that only the copy of pA-89 represented by the 2.7 kb *Dra*I fragment is found in the 1000 kb *Not*I and 970 kb *Not*I-*Sfi*II bands. We have explained our reasons for supposing that both the 1000 kb *Not*I and 970 kb *Not*I-*Sfi*II bands contain two separate fragments, each with a copy of marker pA-36. Theoretically, the copy of pA-89 corresponding to the larger *Dra*I fragment could lie on either of the fragments in this band. However, none of the PFGE data support a model that physically links the polymorphic copy of pA-36 (on linkage group H) to the copy of pA-89 that is located on the smaller *Dra*I fragment. On the basis of this analysis we tentatively designated the copy of pA-89 represented by the 2.7kb *Dra*I fragment as pA-89a in linkage group H.

It is possible that the other copy of pA-89 (pA-89b) comigrates with the monomorphic copy of pA-36 on a fragment in the 350 kb *Sfi*II band, suggesting that the proximity of these markers is conserved in linkage groups H and B. The conservation of motifs of molecular markers in different linkage groups revealed by both physical mapping and RFLP analysis may be a remnant of the ancient tetraploidization of the soybean genome. In a similar way, Helentjaris *et al.* (1988) postulated that the presence of duplicate sequences arrayed in order on different linkage groups in the maize RFLP map supported the idea that allotetraploidy may have occurred during the evolution of maize. Some homoeology between linkage groups is masked by the failure of some copies of the probes to reveal a polymorphism between the parental *G. soja* and *G. max* lines used to construct the USDA-ARS/ISU soybean RFLP map. An analysis using different parental strains could reveal the desired polymorphism and be used to complement the growing map. Alternatively, PFGE analysis can be used to detect linkage blocks that are invisible to RFLP analysis.

The 970 kb *NotI-SfiI* fragment that we propose contains the polymorphic copies of pA-36 and pA-89 spans a recombination distance of 1.8 cM, so that in this cluster of markers there are about 550 kb/cM. This relationship may vary significantly throughout the genome and depends on the accuracy of the genetic mapping data. For example, Ganai *et al.* (1992) reported that in the *Tm-2a* region of the tomato genome recombination was suppressed sevenfold compared with the expected average value, obtained by comparing the number of base pairs in the haploid genome to the size of the classical linkage map. In contrast, Segal *et al.* (1992) found that there were only 43 kb/cM (10-fold below the average for the tomato genome) in the region surrounding the *I₂* locus. It appears that the physical distance between copies of the markers studied here is about the same in linkage groups H and B. Because the copies of pA-36 and pK-9 that we propose are on linkage group B have not been placed on the RFLP map, we cannot determine if the similarity of the physical size of these clusters is reflected by a similar overall recombinational size.

We are interested in the cloning of the *nts* or supernodulation locus of soybean. This gene is involved in the autoregulation of nodule number in the nitrogen fixing symbiosis with the soil-borne *Bradyrhizobium japonicum* (Carroll *et al.*, 1985a,b). Since the product of the *nts* gene is unknown, we intend to proceed by a reverse genetical approach (positional cloning; Wicking and Williamson, 1991), beginning with the identification of molecular probes tightly linked to the Nts phenotype. Landau-Ellis *et al.* (1991, 1992) achieved this when they mapped both the *nts*382 and *nts*1007 alleles of the *nts* locus close to the molecular marker pUTG-132a (1.1 cM), located in linkage group H 18 cM away from the cluster of markers analyzed in this study (Figure 6.2). There are not yet other markers in the vicinity that could be used to construct a physical map of the area or to establish the directionality of a possible chromosome walk. However, the lack of recombination between pUTG-132a and the *nts* locus may allow us to 'land' on the locus if sufficiently large pUTG-132a-homologous fragments can be cloned from wild-type and mutant plants. Also, the marker pUTG-132a and the terminal marker of this linkage group, pA-381, are both single copy, which may mean that the *nts* gene lies on a diploidized chromosome arm. In this case, if a chromosome walk becomes necessary, we may be spared some of the complications of trying to locate a gene surrounded by duplicate genomic sequences.

The existence of polymorphisms between the parental lines used to construct the ARS/ISU RFLP map and the cultivars which were used to induce and map the *nts* locus is a potentially serious problem because of the presence of many multicopy markers in the genome. We were able to use pUTG-132a as an anchor in linkage group H because it is not represented anywhere else in the genome. The polymorphic copy of pA-36 was mapped relative to it and placed in the same linkage group. By cohybridization on pulsed field Southern blots we were able to do the same for the copy of pA-89 which is represented by the 2.7 kb *Dra*I fragment. Unless a single-copy anchor is available or positional cloning projects are started in the same populations from which the RFLP map is built, misapprehensions can arise about where in the genome one finds oneself during the course of a mapping experiment.

CHAPTER 7

PLANT YEAST ARTIFICIAL CHROMOSOME LIBRARIES AND THEIR USE: STATUS AND SOME STRATEGIC CONSIDERATIONS*

INTRODUCTION

Many genes that are now being pursued by plant scientists cannot be cloned by conventional procedures. These include disease resistance loci (e.g., Martin *et al.*, 1992, 1993) and genes that are involved in hormonal responses and developmental pathways (Chang *et al.*, 1993). The time and location of expression of these genes is usually not known, and often they are identified by mutational analysis. The genes may be characterized and placed on a genetic or RFLP map.

It has been demonstrated that these genes are accessible to map-based or positional cloning if the proper materials and experimental resources can be created. Once the gene of interest has been localized on a chromosome or linkage group by co-segregation with genetic or RFLP markers, it is envisaged that the genes may be approached by chromosome walking (see chapter 3; Wicking and Williamson, 1991; Gibson and Somerville, 1992).

Yeast artificial chromosomes (YACs; Burke *et al.*, 1987) have promised to be the vehicle of choice for such expeditions. This is because the nearest flanking marker may be several hundred kilobases (kb) from the locus of interest, and stretches of DNA as large as this may be stably propagated in yeast. It was hoped that YACs could bridge regions of repetitive DNA that would foil a chromosome walk in *E. coli* based vectors, and would also permit cloning of regions of DNA that are not stable in other vectors (Hieter *et al.*, 1990; Edwards *et al.*, 1992).

This chapter reviews the status of YAC projects in plants, and raises some strategic questions that grow out of the structure and composition of different plant genomes.

VECTORS

All existing plant YAC libraries were made in the original vectors described by Burke *et al.* (1987) or in modified derivatives thereof. The YAC vectors are maintained as plasmids in *E. coli*. Treating the plasmid with *Bam*HI and a second enzyme to expose the cloning site

*Funke, R.P., and Kolchinsky, A. (1994). In Current Topics in Plant Molecular Biology, P.M. Gresshoff, ed (Boca Raton, Florida: CRC Press), pp. 125-134.

releases centric and acentric artificial chromosome arms. The insert DNA is ligated between the arms, and *URA* and *TRP* selectable markers on the arms allow selection of transformed yeast cells which are able to grow on medium lacking these nutrients. To increase the usefulness of their *Arabidopsis* library for chromosome walking, Grill and Somerville (1991) modified the YAC4 vector to allow PCR-based amplification of the ends of inserts and simplify the production of end probes.

Several vector systems have been developed more recently (for a review see Hieter *et al.*, 1990). These include the pJS97/98 pair of vectors, in which the chromosome arms are maintained in separate plasmids and a T7 promoter has been placed next to the cloning polylinker. The development of a selectable high-copy-number YAC vector was reported recently which has proved useful in screening of colony blots of transformants and in the preparation of large amounts of gel purified cloned DNA (Moir *et al.*, 1993).

Even for plant species in which well-characterized YAC libraries are already in use, additional libraries may be constructed to take advantage of the features of new generations of YAC vectors. Improvements in the methods for preparing insert DNA should increase the average insert size of the new libraries (see below).

PREPARATION OF INSERT DNA

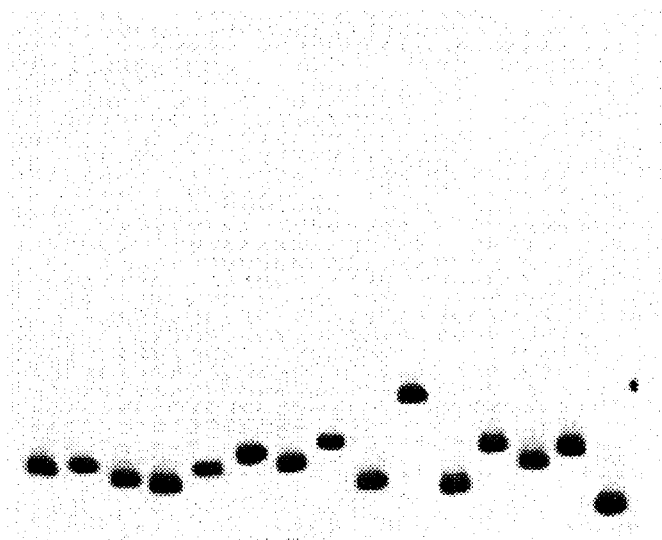
The isolation of very high molecular weight DNA pure enough to be suitable for YAC cloning presents special problems in plants because of the barrier posed by the cell wall and the presence of polysaccharides and other secondary products. In *Arabidopsis* there is the additional problem that the nuclear genome is relatively small (about 70 Mb; Meyerowitz, 1987). This makes it difficult to achieve sufficiently high DNA concentrations by lysing protoplasts embedded in agarose, which is probably the most reliable way to ensure that HMW DNA is not sheared (this is a relatively minor problem amidst the many advantages this model organism offers to map-based cloning strategies).

Guzman and Ecker (1988) prepared HMW DNA from *Arabidopsis* and carrot by centrifuging a protoplast lysate through a sucrose gradient. YACs were produced by partial *EcoRI* digestion of the pelleted DNA followed by ligation to the vector YAC4. When no additional size selection steps were applied prior to yeast transformation, the average insert size of the clones was only 75 kb. This is typical of libraries prepared in this way, because small molecules present in the mixture will transform at higher efficiency than larger ligation products (Figure 7.1).

(a)

200 kb

50kb



(b)

200 kb

50kb

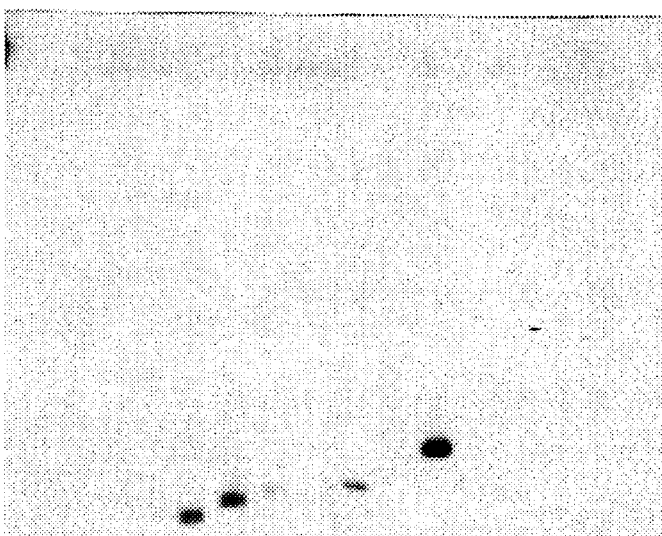


Figure 7.1. Soybean YACs produced without size selection of insert DNA. Soybean protoplasts at a high concentration were lysed by mixing with an equal volume of 0.5 M EDTA (pH 9.0), 1% sarkosyl, 0.5 mg/ml proteinase K at 50°C for one hour. After dialysis for 48 hours against two changes of large volumes of TE, the viscous green lysate was partially digested with *Eco*RI. The enzyme was heat inactivated and ligation to a molar excess of YAC4 was performed in the same reaction mixture. Aliquots of the mixture were used to transform yeast strain AB1380 (Burgers and Percival, 1987). Colonies containing recombinant YACs were prepared for pulsed field gel analysis according to Schwartz and Cantor (1984). YACs were detected on a Southern blot by hybridization with pBR322 plasmid (a) and nick-translated total soybean DNA (b).

By introducing size selection steps in the procedure, the average size of cloned inserts may be increased. Ward and Jen (1990) selected ligation products on a sucrose gradient. By pooling fractions containing DNA molecules larger than 50 kb, they produced a library with average clone size 150 kb. Size selection on CHEF gels involves compressing DNA molecules larger than a predetermined cut-off size in the limiting mobility band below the wells (Anand *et al.*, 1989; Albertsen *et al.*, 1990). Grill and Somerville (1991) achieved a 150 kb average insert size by size selecting ligation products on CHEF gels, while Guzman and Ecker (1990) report an insert size of 250 kb.

During the construction of the tomato YAC library an evolution of methods took place (Martin *et al.*, 1992). HMW tomato DNA was prepared by embedding protoplasts in agarose blocks and lysing them with detergent and proteinase K. The agarose blocks were liquefied by treating them with agarase prior to partial *EcoRI* digestion. A single size selection with CHEF was applied following ligation of the partial digestion products to the vector YAC4. The later additions to the tomato library were produced by size selecting partially digested DNA prior to ligation and performing a second size selection after ligation to remove excess cloning vector. YACs produced in this way averaged about 200 kb in size.

A rice YAC library produced by double size selection on CHEF gels had inserts averaging 350 kb, with a maximum size greater than 1Mb (Ashikawa *et al.*, 1992).

The double size selection appears to be critical for producing large clones, since a considerable number of molecules smaller than the selected size remains trapped in the compression zone of PFGE gels (Albertsen *et al.*, 1990). However, in many cases double size selection is reported to result in an intolerable decrease in transformation efficiency (Kleine *et al.*, 1993). This may be the result of difficulties in the extraction of HMW DNA from plant cells, and helps to explain why the average insert size of plant YAC libraries is considerably smaller than that reported for numerous mammalian libraries (Edwards *et al.*, 1992). During the construction of a partial barley YAC library, Kleine *et al.* (1993) reported that they achieved respectable insert sizes (albeit small by the standards of mammalian libraries) without any size selection steps (average size: 140 kb). Edwards *et al.* (1992) and Del-Favero *et al.* (1994) used a double size selection procedure that was unusual in that they retrieved DNA plugs from the wells of pulsed field gels and used large DNA remaining in the plugs for partial digestion and cloning. An initial size selection of using a 60 second switch interval for 4 hours was used to remove degraded DNA from the

plugs prior to partial digestion. A second less stringent size selection (10 second switch interval, 2-4 hours) was used after partial digestion to remove short products prior to ligation and transformation of yeast spheroplasts. The average size of the cloned inserts was 140 kb in spite of this type of double size selection.

CHARACTERIZATION AND USE OF YAC LIBRARIES

The test of a YAC library is its utility in physical mapping and chromosome walking experiments.

When they are produced, libraries are usually assayed for their representation of single-copy sequences. One *Arabidopsis* library containing 20,000 YACs of average size 150 kb was evaluated by probing with RFLP markers and judged to contain the entire genome once in every 800 clones (Grill and Somerville, 1991). The tomato library was characterized for the presence of repetitive and organellar sequences. Probing the library with unique markers linked to disease resistance loci suggested that the 22,000 clones represented three haploid genomes.

Edwards *et al.* (1992) calculated that the maize YAC library contained three haploid genome equivalents. Using probes or PCR primers for low copy number sequences in the genome, for example the alcohol dehydrogenase (*Adh1*) gene, they were able to identify between one and eight positive clones. Similarly, Del-Favero *et al.* (1994) assayed a sugar beet library containing five haploid genome equivalents for repetitive and single copy sequences, and isolated the expected number of positive clones in each case. A single *Hor1*-specific clone was isolated from the barley YAC library, and the clone proved to be colinear with barley genomic DNA (Kleine *et al.*, 1993).

Identification of YACs by their hybridization to RFLP markers does not guarantee that they are colinear with the plant genome. The fidelity of the YACs is essential to the success of a positional cloning project, as all the sequences from the chromosomal region containing the candidate gene must be cloned. This is best ensured by having some redundancy, so overlapping YACs verify each other's restriction maps. The restriction map of the resulting contig should coincide with that of the genome.

About 1700 kb of contiguous DNA were arrayed in overlapping YACs in the course of a search for the *CO* gene of *Arabidopsis* (Potterill *et al.*, 1993). The authors analyzed 67 YACs, and reported that 40% of them presented difficulties for chromosome walking because they contained repetitive sequences at their termini, were chimeric, or were missing

pieces of plant DNA. Some of the YACs suffered progressive deletions during propagation in yeast. Upon reflection, the authors would rather not have embarked on such an ambitious chromosome walk and stress the advantages of detailed genetic and RFLP mapping to land as close as possible to the gene of interest. Leung *et al.* (1994) discovered that 3 of 8 YACs they isolated during the assembly of a contig around the *ABII* locus of *Arabidopsis* were rearranged. Because of the redundancy of the *Arabidopsis* YAC libraries, they were able to isolate additional overlapping clones that aligned properly with the genomic area around the gene of interest.

In *Arabidopsis* work is in progress towards assembling an overlapping YAC library of the entire genome (Hwang *et al.*, 1991). If RFLP markers flanking a gene can be identified and clones are large enough, it is possible to identify a YAC containing the gene of interest. This would circumvent the technical difficulties of chromosome walking in YACs (Matallana *et al.*, 1992). This approach was used successfully by Leyser *et al.* (1993), who cloned an *Arabidopsis* auxin resistance gene after localizing it on a single YAC which hybridized to two flanking markers.

No other plant YAC library has been assessed for its usefulness in chromosome walking. A single YAC clone homologous to a linked molecular marker allowed Martin *et al.* (1993; see chapter 3) to isolate a gene for *Pseudomonas* resistance in tomato. Transfer of this gene (*Pto*) to susceptible plants conferred resistance to the transformed plants. Cloning a segment of DNA encompassing a gene of interest was more laborious in tomato than it was in the model plant *Arabidopsis*. Some of the reasons can probably be related to the fact the tomato genome, while relatively small and possessed of similar organizational advantages, is still an order of magnitude larger than that of *Arabidopsis*, and that the RFLP map is not as unified. Another problem was the presence of the *Pto* gene in a cultivar different from the parents that were used to create the tomato RFLP map. This had the consequence that the markers on the map had to be surveyed a second time to determine whether they showed useful polymorphisms in the material that was being used to isolate the gene. This also motivated a search for additional linked markers using NILs and single primer DNA amplification. In the event, a marker from the original RFLP map proved to be the most tightly linked to the *Pto* gene. The ends of the large YAC that was isolated using this marker were converted to polymorphisms that were essential in focusing the location of the gene. Only a single YAC was used, while in *Arabidopsis* the availability of several libraries means that a number of overlapping YACs can usually be isolated using any marker. This redundancy makes it possible to detect deletions or rearrangements in the YACs, and is useful in the construction of a contig. In the case of *Pto*, if several overlapping YACs had

been available to build a contig extending beyond the right arm of 400 kb clone, the effort needed to make sure that the gene was indeed located on the YAC and not just outside the right arm might have been avoided.

YAC STABILITY

The problems with YAC instability encountered by Potterill *et al.* (1993) are not unique to plant YAC libraries. It has been observed that some sequences of human DNA suffer deletions in yeast. For example, Neil *et al.* (1990) found that clones containing tandemly repeated *DYZ1* sequences from the human Y chromosome were underrepresented in their YAC library. Clones containing *DYZ3* sequences were not colinear with genomic sequences or were unstable. During the construction of the YAC contig of the euchromatic region of the Y chromosome, Foote *et al.* (1992) encountered YACs that sustained deletions, and chromosome segments containing known tandem repeats were especially susceptible to clonal instability.

There is some evidence that repetitive DNA from plants may cause similar problems. Martin *et al.* (1992) found that of YACs smaller than 100 kb a smaller number than expected contained the interspersed repetitive element *TGR11*. Larger clones contained the sequence in the expected proportion, and it was suggested that smaller clones containing *TGR11* are unstable.

Dunford *et al.* (1991, 1993) discovered that some barley YAC clones containing the dispersed retro-element *WIS11* were structurally unstable. By transforming a YAC containing these sequences into the recombination deficient yeast strain LIV2, the authors demonstrated that some stabilization could be achieved. However, because of low transformation efficiency, LIV2 was not deemed suitable as a primary host for library construction.

THE IMPORTANCE OF PLANT GENOME ORGANIZATION

An abundance of repetitive DNA in the genome might be expected to cause problems both with respect to the fidelity of YACs and the increased probability that end probes from the clones will hybridize to many noncontiguous clones in the library (Moore *et al.*, 1993).

Plant species with relatively small genomes, such as tomato and *Arabidopsis*, contain a relatively small amount of repetitive DNA (14% and 15-20%, respectively; Lapitan, 1992). In fact, the proportion of single-copy DNA sequences appears to be inversely related to

genome size (Flavell, 1980). About 60-80% of the genomes of wheat and maize consist of repeated sequences, while soybean contains an intermediate amount of repetitive DNA (30-40%; Lapitan, 1992; Gurley *et al.*, 1979).

Moreover, it appears that the interspersion pattern of repeated sequences and single copy DNA appears to be related to genome size (Lapitan, 1992). The average length of unique sequences in the *Arabidopsis* and tomato genomes is 120-130 kb. Species with larger genomes generally show a different pattern, in which short repetitive sequences are interspersed with short, single copy sequences. Soybean, which has a genome slightly larger than that of tomato (1000 Mb; Blackhall *et al.*, 1991), exhibits a short period interspersion pattern. About 60-70% of single copy sequences have an average length of 1.11-1.36 kb interspersed with 0.3-0.4 kb repetitive elements (Gurley *et al.*, 1979).

In *Arabidopsis* and tomato, more than 90% of RFLP markers map to single loci (Lapitan, 1992). In maize, however, many RFLP markers detect duplicate loci, and motifs of linked markers are conserved in different parts of the genome (Helentjaris *et al.*, 1988). Similarly, in soybean, which is regarded as an ancient tetraploid, multicopy markers detect what appear to be homoeologous linkage groups (Shoemaker *et al.*, 1992; Funke *et al.*, 1993). Physical mapping of a cluster of RFLP markers in soybean revealed that duplicate copies of the same RFLP marker that segregate to different linkage groups nevertheless detect identically sized *NotI* and *SfiI* fragments (200-1000 kb) on pulsed field gels (Funke *et al.*, 1993). This suggests that the sequence similarity in homoeologous chromosome segments may encompass a lot of DNA in addition to the duplicate loci detected by markers. Low copy number sequences that are part of conserved linkage blocks on different chromosomes create the danger of being diverted to a part of the genome that is not contiguous to the starting point of a walk.

The possible interference posed by these characteristics of plant genomes is causing would-be chromosome walkers to revise their strategies. Moore *et al.* (1993) proposed a small genome model approach to mapping and cloning projects in cereals. The genome of rice is the smallest among the *Triticaceae*, but gene order and linkage groups appear to be highly conserved among the genomes. The fraction of repetitive DNA is smaller than in barley and wheat, and YACs from rice may facilitate chromosome walking by bridging repetitive sequences that separate conserved single copy sequences in the genomes of other cereals. A similar approach may be useful among legumes, and the colinearity of small and large genomes, for example of mungbean and soybean, is being assessed with RFLP markers (Young *et al.*, 1992).

Flow sorting of chromosomes, as is now possible for hexaploid wheat, should enable the production of chromosome-specific YAC libraries (Moore *et al.*, 1993; McCormick *et al.*, 1993). Once a gene has been localized to an individual chromosome, this should simplify screening and walking in the library, as it would avoid the background created by duplicate low copy number sequences in a polyploid genome.

CONCLUSIONS

This chapter has addressed some of the strategic factors involved in map based cloning projects in plants. One important point was omitted: the conclusion of the cloning process, which would normally involve transformation of mutant plants to try to complement the phenotype with wild type gene sequences (Gibson and Somerville, 1992). If an efficient transformation system is not available for a particular species, other ways of screening for candidate gene sequences should be devised (see chapters 3, 4).

Perhaps the most important conclusion is the need to accumulate as many markers as possible in the vicinity of an unknown locus before touching down in a library (see, for example Martin *et al.*, 1991, and Michelmore *et al.*, 1991). Chromosome landing, rather than chromosome walking, will be desirable especially in the large, complex genomes of agronomically important crop species such as maize, barley, and soybean. In these species it will be a special challenge to develop the dense RFLP maps and redundant YAC libraries that are necessary for chromosome landing.

Finally, some questions, relating for example to the overall stability of YACs from particular species, will be answered as the experience of different researchers is shared. The publication of barley, maize, and sugarbeet libraries are important in this respect, since they possess large, complex genomes with many features which might intimidate researchers interested in positional cloning.

CHAPTER 8

CONSTRUCTION AND CHARACTERIZATION OF A PARTIAL SOYBEAN YEAST ARTIFICIAL CHROMOSOME LIBRARY

INTRODUCTION

In previous chapters, an argument has been built for a chromosome walking or landing strategy for the isolation of the *nts* gene, since there is no information available on the biochemical function of the gene product (Gresshoff, 1993). This strategy requires the availability of a representative genomic library containing large inserts (Kolchinsky *et al.*, 1994).

YAC libraries have become an indispensable tool for the analysis of complex eukaryotic genomes, and utility for positional cloning of genes known only by their phenotype has been demonstrated (see Chapter 3). Therefore, the construction of a soybean yeast artificial chromosome library was undertaken.

NUMBER OF CLONES NEEDED FOR A REPRESENTATIVE LIBRARY OF SOYBEAN

YAC libraries are normally stored in an array in 96-well plates (Foote, 1993). This means that each clone has to be picked, repicked, and stored individually so that it has a specific 'address' in the array. In the most favorable conditions, it may take between many months to two or more years to produce a complete library. In many cases clones that are added later have larger inserts and the library may go through generations of development. The investment that this represents in terms of space (storage), time, and effort makes it worthwhile to calculate what size library is needed.

The number of clones N needed to have a given probability P of finding a specific sequence in a genomic library is given by the relationship:

$$N = \frac{\ln(1-P)}{\ln(1-f)}$$

where f is the fractional proportion of the genome in a single recombinant (Sambrook *et al.*, 1989). The haploid genome size of soybean is estimated to be 10^9 bp (Blackhall *et al.*, 1991).

An important ingredient in this calculation is the average insert size that can be achieved. The best currently available plant YAC libraries have an average insert size of 250-350 kb, in spite of the capacity of the vectors to carry more than 1 Mbp of DNA. Most plant YAC libraries have an average insert size of 150 kb (see Chapter 7).

If an average insert size of 150 kb can be achieved, $\ln(1-f) = -1.5 \times 10^{-4}$. To have a 99% probability of finding a specific sequence in this library ($P=0.99$), 30,000 clones would be needed. For the same probability, an average insert size of 100 kb would require 46,000 clones, and if an average size of 200 kb can be achieved 23,000 clones would be needed.

For a 95% probability of finding a certain sequence, the library should contain 20,000 clones with an average size of 150 kb. Such a library can be stored in two hundred 96-well plates and would occupy about eight cubic feet in a large chest freezer.

PROCEDURES FOR THE CONSTRUCTION OF A SOYBEAN YAC LIBRARY: SURVEY OF THE CRITICAL FACTORS

YAC construction is a multi-faceted procedure. Many different components converge to produce a successful transformation, and each of them should be optimized. They include: vector preparation, preparation and size selection of insert DNA, ligation of vector to insert, the preparation of competent yeast cells, and transformation.

The vector

Most YAC libraries are constructed using the vector YAC4 (Figure 8.1; Burke *et al.*, 1987). This vector carries both the centric and acentric artificial chromosome arms on a single plasmid. pYAC4 is a relaxed plasmid that carries the ampicillin resistance gene and can be amplified in *E. coli*. Digestion with *Bam*HI exposes the telomeres, and digestion with *Eco*RI opens the cloning site (in other YAC vectors different cloning sites have been created. For example, in YAC2, *Sma*I exposes a blunt cloning site). The centric arm contains a gene to complement uracil auxotrophy, and the acentric arm contains a gene to complement tryptophan auxotrophy. Thus, a cloned piece of DNA flanked by the centric and acentric arms should be propagated as an artificial chromosome that corrects uracil and tryptophan auxotrophy on yeast hosts carrying *ura*⁻ and *trp*⁻ mutations.

Color selection distinguishes yeast cells transformed with religated vector containing no insert and those that are truly recombinant. The cloning site is located within the intron of the *supF* gene. *supF* encodes a tRNA that permits extension of a gene for adenine

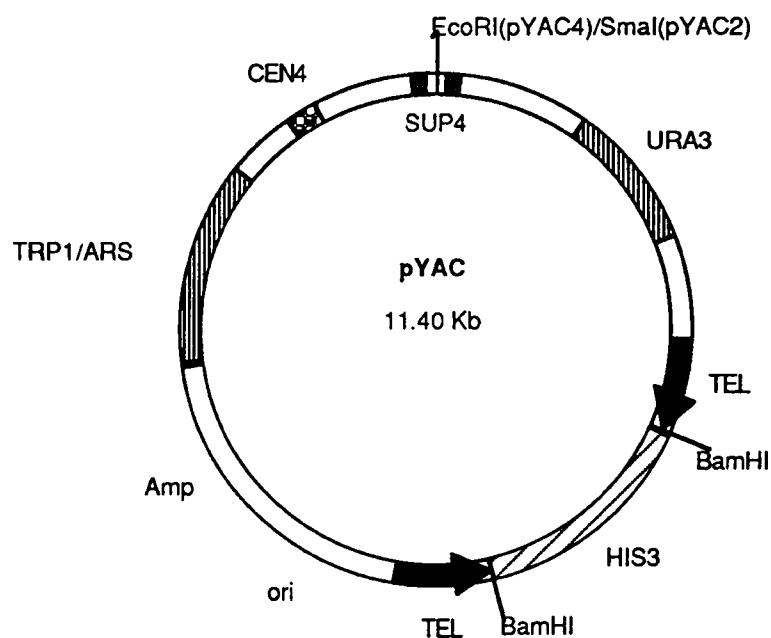


Figure 8.1. YAC cloning vector. Cloning sites for the in the pYAC4 and pYAC2 editions of the vector are shown. The cloning site is located in the intron of the SUP4 gene, which codes for an ochre-suppressor tRNA^{Tyr}. Digestion with the enzyme in the cloning site and *Bam*HI releases the centric (with the TRP1 gene; 6 kb) and acentric (with the URA3 gene; 3 kb) vector arms. The HIS3 fragment helps in stable propagation of the plasmid in *E. coli*, and has no role in the YAC cloning procedure.

biosynthesis that contains a premature stop codon in the yeast host AB1380. The *ade2* mutation of AB1380 causes the accumulation of a red intermediate product at the point where the biosynthesis is interrupted. When the *sup* gene is transformed intact into AB1380, as in the case where there is no insert DNA between the chromosome arms, the cell recovers from adenine auxotrophy and white colonies are produced even when adenine levels in the medium are limiting. If the gene is interrupted by the insertion of a foreign piece of DNA, yeast cells transformed with the recombinant artificial chromosome are red because of the precipitation of the adenine intermediate, as well as being *ura*⁺ and *trp*⁺.

The vector is usually dephosphorylated after double digestion with *Bam*HI and *Eco*RI to lower the proportion of nonrecombinant clones, as described in Appendix 3.

It is important that the vector is maintained in a *recA*⁻ *E. coli* host, because the telomeres otherwise undergo deletions and rearrangements which could compromise the stability of the artificial chromosome. A number of separate plasmid minipreps should be performed to check the integrity of the telomeres before choosing one for large scale isolation. Transforming a correct plasmid into a *recA*⁻ host, such as DH5 α , eliminates this problem.

Another potential problem with the vector preparations is the presence of contaminating bacterial chromosomal DNA. Although it is dephosphorylated along with the digested plasmid, the large number of ends of small chromosomal DNA molecules could compete with the insert DNA in a ligation reaction and would be favored during transformation of the yeast host. The resulting recombinant clones would be very small and indistinguishable phenotypically from clones carrying soybean DNA. This has not been a problem during the construction of the soybean library, since the presence of soybean DNA even in very small clones can be verified by hybridization to total soybean DNA. As a precaution, the vector was always purified over two cesium chloride gradients.

Preparation of insert DNA

For the preparation of YAC insert DNA, very high molecular weight DNA is a prerequisite. The starting DNA must have an average molecular weight at least three times that of the desired average insert size of the clones. The transformation procedure is biased towards smaller molecules, such that even in an equimolar mixture of small and large artificial chromosomes, the smaller molecules transform yeast at a frequency ten times that of the larger ones (Connelly *et al.*, 1991).

Procedures for the extraction of high molecular weight DNA from soybean have been described (Chapter 5 of this thesis). Initially, it proved difficult to achieve partial digestion of agarose-embedded soybean DNA with *EcoRI*. Even using pulsed field gel electrophoresis resolution, the addition of a large number of units of enzyme produced no visible downward shift in the average size of the DNA. This seemed to be in agreement with earlier observations that *EcoRI* could not digest soybean chromosomal DNA embedded in agarose, whereas *HindIII* could (see Chapter 5). Manufacturers of restriction enzymes include *EcoRI* among enzymes that have difficulty digesting DNA in agarose.

To overcome the apparent insensitivity of the embedded DNA to *EcoRI*, it was decided to start the cloning procedure using DNA isolated as gently as possible in liquid form. The objective of these pilot experiments was to see if any recombinant YACs could be produced, regardless of insert size, to test our fluency with the whole procedure, including vector preparation, ligation of insert DNA to vector, and preparation and transformation of competent yeast cells. Protoplasts were isolated from about 12 trifoliate leaves using the procedure described in Chapter 5. The protoplasts were resuspended in 400 µl of PB and pipetted into narrow dialysis tubing clamped closed at one end. ESP (400 µl) was very gently added to the protoplasts to lyse them and a second clamp was applied tightly to close the tubing.

The protoplast lysate was incubated overnight at 50°C. The dialysis tubing containing the viscous green mixture was dialysed against four changes of one liter of sterile TE buffer at 4°C. The mixture was transferred to an Eppendorf tube and proved to be sensitive to *EcoRI* digestion. However, there was a considerable amount of degraded DNA.

YACs were produced without applying any size selection steps to the insert DNA. Protoplast lysate was partially digested with *EcoRI* and the enzyme was heat inactivated by incubation at 65°C for 20 minutes. The mixture was supplemented with 1 mM ATP and DTT, and vector (equal in mass to the amount of DNA in the lysate) and T4 DNA ligase (NEB, 4 µl at 400 units/µl) were added. In a separate experiment, the randomly sheared ends of DNA in the lysate were filled in with Klenow enzyme in the presence of dNTPs, and the mixture was ligated to the blunt ended vector YAC2. After overnight incubation at 16°C, 10 µl of the viscous ligation mixture was used to transform 100 µl of spheroplasts using the procedures described in Appendix 3.

These experiments consistently produced a number of YACs. The clones that were picked after 5 days' growth on Ura⁻ medium were red and were able to grow on double selective

(Ura⁻Trp⁻) medium. The artificial chromosomes in these clones were shown to be recombinant by pulsed field gel electrophoresis and hybridization (Figure 7.1). All of the extra chromosomes hybridized with pBR322 (almost all the pBR322 sequences are retained in the YAC vector), and probing with total nick-translated soybean DNA showed that they contained inserts from the soybean genome. The average size of the inserts was about 60 kb, with a range from 30 to 200 kb.

It proved difficult to improve the average insert size of the clones starting from protoplast lysate. The viscous mixture was difficult to work with, and size selection by sucrose density gradient centrifugation, which proved successful in some cases (see Chapter 7), was not practical. One reason for this was that significant losses were incurred during centrifugation, and the volume of starting material was limiting. Fractions from the gradients could be analysed by pulsed field gel electrophoresis, but they were very dilute and attempts to concentrate them were not successful. Each manipulation of liquid high molecular weight DNA resulted in material losses and degradation. Accordingly, it was decided that this was not a promising approach to improving the average size of the YAC inserts.

Later, suitable conditions for partial digestion of embedded soybean DNA were worked out. Work with agarose blocks was more attractive because it was possible to start with relatively small amounts of DNA, and all manipulations and buffer changes could be accomplished simply by washing the plugs without pipetting the DNA or changing its concentration. Extensive washing and dialysis of the agarose blocks left DNA that was exquisitely sensitive to *Eco*RI (Figure 8.2). This level of sensitivity is not reported by all laboratories producing YAC libraries from plant DNA. For example, during the construction of a *Beta vulgaris* library, 50 units of enzyme were applied twice for one-and-a-half hours to similar amounts of DNA to achieve levels of partial digestion that resulted in a library with an average insert size of 140 kb (Eyers *et al.*, 1992).

If the starting quality of the DNA in the plugs was good, it was possible to titrate the extent of partial digestion with different amounts of *Eco*RI directly and choose a point with which to scale up the reaction and set ligations. However, the quality of the DNA plugs can be variable, and degraded DNA can represent a sizable fraction of all the DNA in the plug. This may not interfere with the production of large numbers of clones, but the presence of degraded DNA makes it more difficult to titrate visually *Eco*RI partial digestion conditions and lowers the average insert size of the clones. To improve the average molecular weight of the starting DNA, a pre-electrophoresis step was eventually added to the procedure,

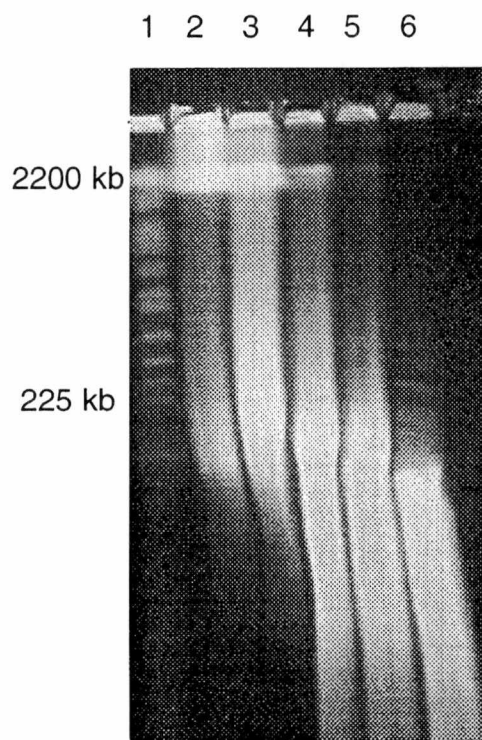


Figure 8.2. Partial digestion of soybean DNA in agarose plugs with *EcoRI*, according to procedures described in Appendix 2. The plugs were not preselected in this case, because the starting material was substantially free of degraded DNA. Lane 1, yeast size standard; lane 2, uncut DNA; lane 3, 0.2 unit *EcoRI*; lane 4, 4 u; lane 5, 10 u; lane 6, 20 u.

based on the recommendations of Edwards *et al.* (1992). By subjecting soybean DNA plugs to electrophoresis for four hours at 200 V using a 60 second switch interval, a significant amount of degraded DNA is effectively removed. The plugs that are retrieved from the wells contain primarily high molecular weight DNA with an average size greater than 2 Mbp (see Figure A3.1).

After partial digestion of agarose embedded DNA, the DNA blocks were processed in different ways before ligation to vector and transformation. The size of the clones produced by each procedure was determined by pulsed field gel electrophoresis and hybridization.

When the partially digested DNA was ligated to dephosphorylated vector without further manipulation or size selection, numerous clones were produced consistently with an average size in the range of 90 to 130 kb, depending on the transformation. The insert sizes ranged from 40 to 300 kb.

The average size of inserts was improved when the partially digested DNA was subjected to size exclusion by pulsed field gel electrophoresis before ligation and transformation. A pulse time of 3s was used initially, which focuses DNA 150 kb and larger in the limiting mobility zone of the gel. By ligating the DNA fragments that were compressed in the limiting mobility zone to dephosphorylated vector, 400 clones with average sizes of 150 kb were produced in three independent transformations. The size of the inserts ranged from 40 to 900 kb, with most clustered in the interval from 100-200 kb. These clones were added to the library for permanent storage.

Although this is a straightforward approach, the results are not reproducible. In many apparently identical transformations using a single size selection, all of the inserts were less than 100 kb in size, with a substantial number of the transformants carrying less than 50 kb of soybean DNA. The presence of soybean inserts even in the smallest clones was confirmed by Southern hybridization. Increasing the stringency of size selection by using a pulse time of 30 seconds, which should exclude all molecules smaller than about 600 kb from the compression zone of the gel, did not alter the size distribution of the clones that were produced (Figure 8.3).

The ability of the DNA to withstand size selection and other manipulations was routinely tested by checking aliquots from intermediate steps in the procedure to check for degradation. In addition, DNA plugs containing yeast chromosomes were subjected to all the washing and melting steps in the procedure and checked by pulsed field electrophoresis

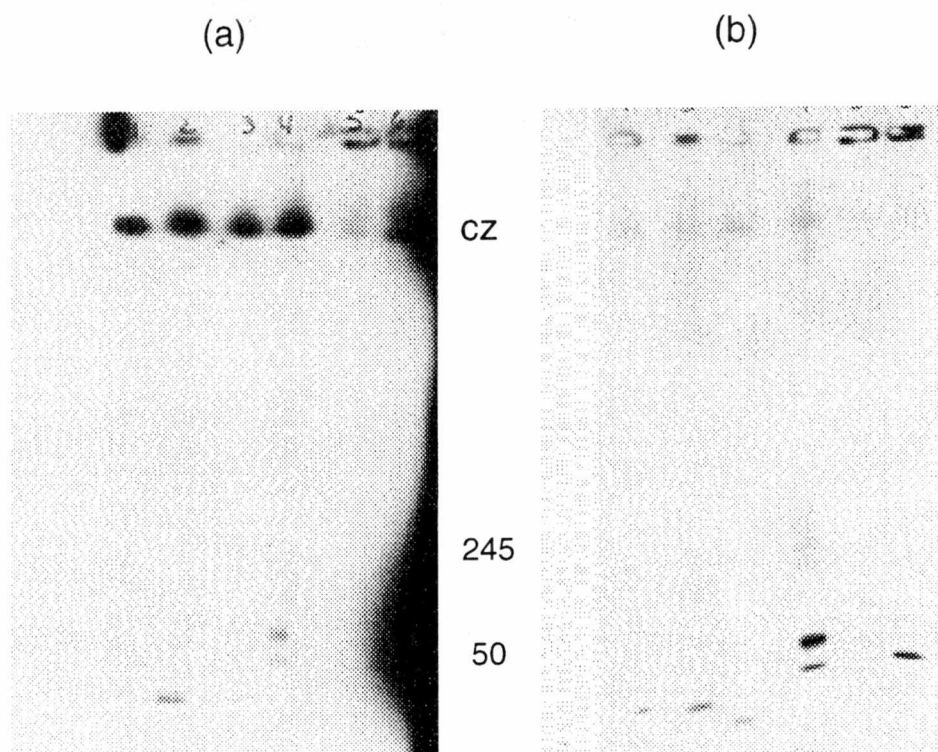


Figure 8.3. Small YACs produced in spite of stringent PFGE size selection of insert DNA. High molecular weight soybean DNA was partially digested with *EcoRI* and the products were compressed as described in Appendix 2 using a 30 second switch interval. This should produce a cut-off size of about 400 kb. In spite of this, clones with an average size of less 50 kb were obtained in this transformation. a) Hybridization of PFGE-resolved YACs to total soybean DNA, proving that the inserts contain soybean DNA. b) Hybridization of a replicate panel of the same clones to pBR322, to show the presence of an artificial chromosome in each lane.

to verify that even the largest chromosomes of more than 2 Mbp remained intact at the end of all manipulations (Figure 8.4). For high molecular weight DNA to withstand melting of the LMP agarose matrix in which it is embedded (65°C for 20 minutes), it is necessary to maintain NaCl concentrations of at least 25 mM and to include polyamines (0.75 mM spermidine and 0.3 mM spermine) at the time of melting (McCormick *et al.*, 1990; Larin *et al.*, 1991).

The preponderance of small clones resulting from these transformations could be due to entrapment of small *EcoRI* digestion products in the compression zone of the size selection gel. As mentioned previously, these small molecules are ligated and transformed at a much higher frequency than large molecules. It was demonstrated that one source of entrapment is the presence of high levels of EDTA salt in the blocks of partially digested DNA that are loaded on the size selection gel (Funke *et al.*, 1994). The EDTA has a retarding effect especially on small molecules migrating in PFGE (see Appendix 4).

However, this is not the only factor that contributes to the entrapment of small molecules and the abundance of small clones. High DNA concentrations in the agarose blocks also cause retardation of DNA in pulsed field gels. This effect becomes especially severe at very high DNA concentrations in excess of 200 µg/ml (Doggett *et al.*, 1992; Chapter 6). By subjecting the DNA to a second size selection on a pulsed field gel after ligation, the effects of entrapment are reduced (see Chapter 7). However, in our hands this produced such a small number of clones that it would not be practical to build an entire library in this way. Moreover, visual estimation of the amount of soybean DNA in the agarose blocks as well as calculations based on the density of the embedded protoplasts suggested that the concentration of DNA in the blocks was well below the limit of 200 µg/ml.

It is possible that there are other reasons for entrapment and the inconsistency of the size selection procedure. One of these could be the presence of residual carbohydrates and cell wall material in the plugs to which high molecular weight might adhere during electrophoresis. The production of YACs smaller than the cut-off size determined by the PFGE size selection regime is not unusual for plant YAC libraries. Later additions to the tomato YAC library used a 30 second switch interval to focus ligation products prior to transformation. Although this represents a cut-off size of about 600 kb, the average size of the clones produced was less than 200 kb (Martin *et al.*, 1992). However, rather than attributing the results to some unusual property of embedded soybean DNA, it would be more useful to emphasize other steps in the procedure that are known to influence the

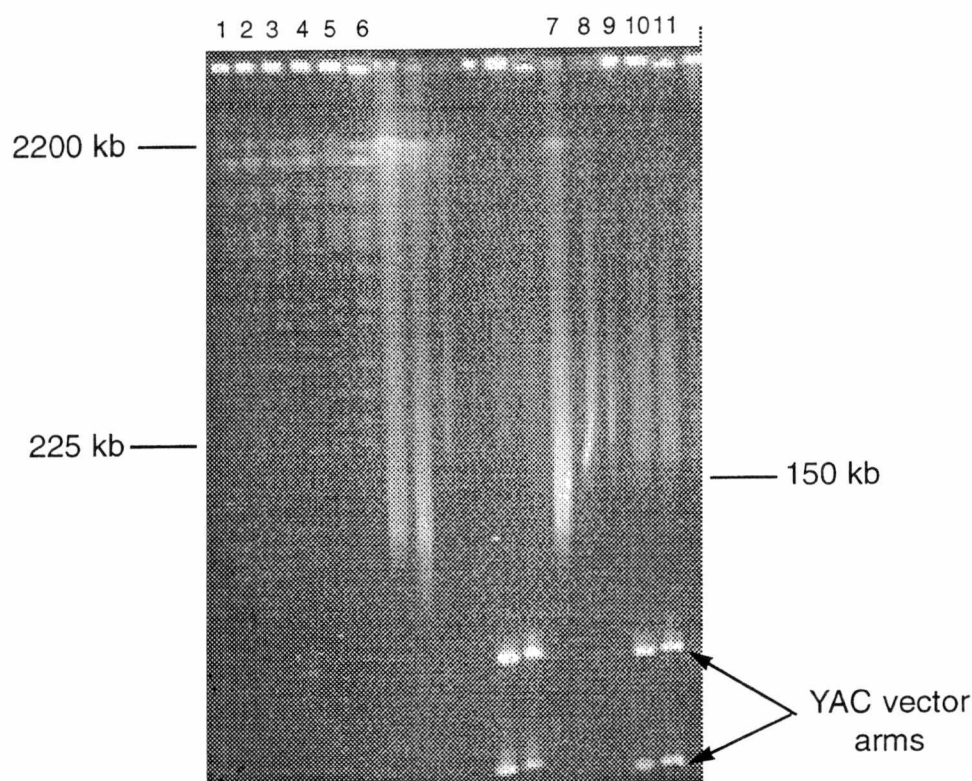


Figure 8.4. Integrity of large DNA fragments during size selection and ligation procedures before transformation. These procedures are described in detail in Appendix 2. Yeast DNA plugs were rinsed with all solutions used in the procedure, and small blocks from each step were assayed by PFGE (lanes 1-6). Aliquots from each stage of a ligation of soybean DNA to YAC vector are shown in lanes 8-12. Lane 1, yeast plug dialyzed with TE buffer; 2, *EcoRI* restriction buffer, no enzyme, incubated 37°C for 1 hour; 3, equilibrated with ligation buffer; 4, melted; 5, remelted after simulated overnight ligation; 6, liquefied with gelase enzyme. Lane 7, Soybean DNA partially digested with 0.2 u *EcoRI*; 8, partial digestion products size selected by PFGE, 3 second switch interval; 9, melted after equilibration with ligase buffer; 10, melted after overnight ligation to YAC vector; 11, ligation products liquefied with gelase enzyme just before transformation. The results show that the solutions used are free of nuclease activity, and that the DNA is adequately protected during melting. Size selection by PFGE eliminates molecules smaller than the expected cut-off size (150 kb with a 3 second interval), and leaves molecules with an average molecular weight much bigger than the cut-off size.

outcome of transformations. Apart from the quality of the insert DNA, the treatment of yeast cells prior to transformation is a critical step in the procedure.

Transformation and preparation of yeast competent cells

Although various procedures exist for the transformation of yeast, the preparation of competent cells by removing the cell wall is the gentlest and most efficient and is used universally for the construction of YAC libraries (Burgers and Percival, 1987).

The yeast host is a complex eukaryotic organism, and its competence and ability to be transformed are determined by more or less obvious experimental factors leading up to the moment of transformation. For reasons that are often not clear, large fluctuations in transformation efficiency are encountered in most laboratories working with YAC libraries (Foote, 1993). The transformation procedure used throughout the experiments described here is described in detail in Appendix 3.

The efficiency of the transformation procedure may be monitored by using a CEN plasmid containing a *ura* gene, such as pRS316. The highest efficiency reported in the literature using such a plasmid is about 10^6 cfu/ μ g (McCormick *et al.*, 1990; Foote, 1993). We routinely obtained efficiencies in the range of $1-2 \times 10^5$ cfu/ μ g using the plasmid pRS316. Critical factors that influence the transformation efficiency are the density at which the cells are harvested (late log phase, 3×10^7 /ml), the degree of spheroplasting, which is monitored by observing the decrease in OD₈₀₀ of the cells during treatment with lyticase, and the quality of reagents, including in particular sorbitol and PEG. Large laboratories have been known to reserve certain lots of these reagents with manufacturers for continued use throughout the construction of a library (personal communication, CSH Course on the Cloning and Analysis of Large DNA Molecules, 1993). For the experiments described here, batches of all reagents from different manufacturers were tested over a period of about two years, and recommendations for which of these proved to be the best are included in the procedure in Appendix 3. It has been found that the degree of spheroplasting that results in the highest plasmid transformation efficiency does not coincide with the point that is optimal for transformation with ligation mixtures (Foote, 1993). Thus, although the plasmid is a useful control and can be used a general guideline, the most favorable conditions of lyticase treatment should be determined by titration using proper ligation mixtures.

The batch of lyticase that is used may be critical. Different grades of yeast lytic enzyme are

available from different manufacturers, ranging from crude powdered extracts that are not tested for DNase and RNase activity, to lyophilized extracts that contain no detectable levels of DNase and RNase. Finally, recombinant grade lyticase is available. All of these enzymes are able to degrade the yeast cell wall, but may produce spheroplasts with varying levels of competence. Moreover, even low levels of DNase in the enzyme will linger until the moment of transformation and degrade large recombinant molecules in the ligation mixture, contributing to the enhanced transformation of the more resistant (and more numerous) YACs with small inserts. All of the batches of lyticase used throughout these experiments have been checked and compared to recombinant grade lyticase, and no nuclease activity was detected (the largest yeast chromosomes remained intact after being liquefied and incubated with different amounts of enzyme, data not shown).

The need for polyamines during melting steps in the preparation of insert DNA has been mentioned and is widely accepted. On the other hand, the inclusion of polyamines during the transformation procedure itself is recommended by some researchers but is thought by others to have a negative effect on the overall transformation efficiency (Connelly *et al.*, 1991 vs. e.g. Kleine *et al.*, 1992). It is proposed that polyamines cause compaction of the DNA and reduce the size discrimination against YACs with large inserts during transformation. In our experience, including polyamines during transformation reduces the proportion of very small YACs that are obtained without having a noticeable effect on the transformation efficiency, and they are routinely added during transformation (Figure 8.5).

Different YAC libraries exhibit different degrees of chimaerism, in which a single YACs contain DNA from widely separated regions of the genome. Although it was thought that the primary source of chimaerism was coligation of partial digestion products from different parts of the genome, it has been suggested that recombination activity by the yeast host at the time of transformation influences the level of chimaerism in YAC libraries. Experiments in which partial filling in of ends prevented ligation between disparate partial digestion products resulted in unreduced levels of chimaerism in the resultant YACs (Wada *et al.*, 1994). This implies that the reduced levels of chimaerism achieved by some researchers may be due to superior handling of yeast (e.g. Albertsen *et al.*, 1992).

Although all the factors contributing to transformation efficiency, overall average insert size, and level of chimaerism may not be known, there is enough evidence to suggest that as much attention should be paid to the yeast host and the conditions of transformation as to the preparation of insert DNA. The use of fresh, well-aerated cultures and good grades of

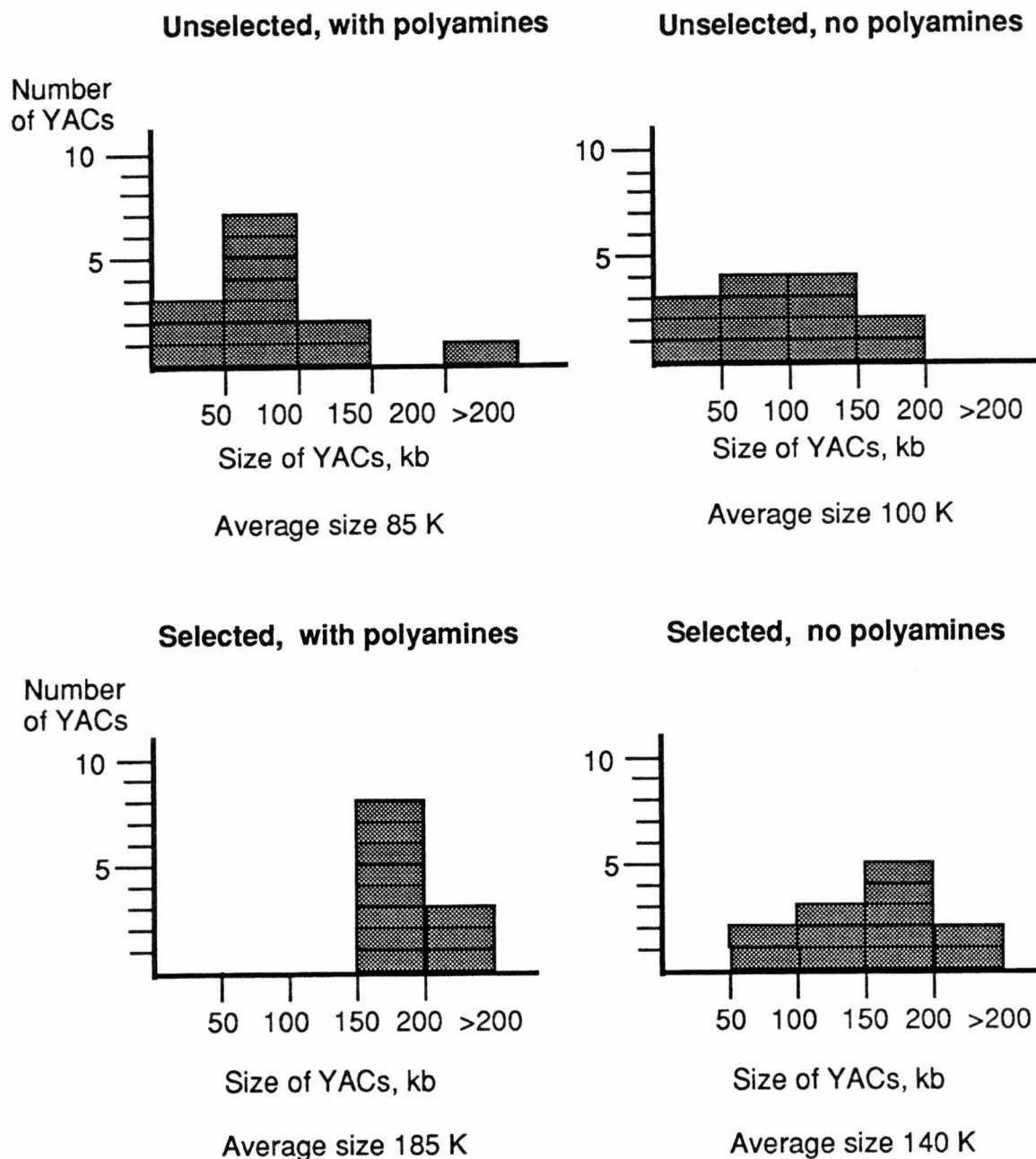


Figure 8.5. Summary of the results of two YAC transformation experiments testing the effect of polyamines and PFGE size selection on insert size. PFGE size selection (using a 3 second switch interval as described in the text) improved the average insert size of the clones in these experiments, while the inclusion of polyamines appeared to reduce the contamination with small clones.

lyticase, as well gentle treatment of spheroplasts should be regarded as bare essentials in the success of a transformation experiment.

REPRESENTATIVENESS OF THE LIBRARY

Ten to fifteen clones from each transformation experiment were checked for insert size as described in Appendix 3, and only if the average size of the clones was greater than 100 kb were all the clones from that experiment repicked and included in the library. The average insert size of all the clones that were checked in this way was 250 kb (Figure 8.6). The majority of the clones are in the range 100-200 kb, with some smaller and a few very large clones carrying inserts as large as 900 kb. Recently, 96 clones were picked from the library at random and stored in a separate plate in order to have a subset of characterized YACs with specific addresses for the purpose of collaboration and distribution to other researchers. All the clones in that plate were characterized for insert size and for the contents of repetitive sequences as assayed by hybridization to total soybean DNA. The average size of these clones was 175 kb (Debbie Landau-Ellis, personal communication).

At present, the partial soybean YAC library contains 600 clones that have been arrayed in 96-well plates. With an average size of 200 kb, this represents about 12% of the genome. The probability P of finding a specific sequence in the soybean genome among these clones, assuming all regions of the nuclear genome are equally represented, may be calculated as follows:

$$600 = \frac{\ln(1-P)}{-2 \times 10^{-4}}$$

Solving for P :

$$P = 0.11 \text{ or } 11\%$$

Since about 10% of the clones in the library originate from the chloroplast genome (Figure 8.7), the actual probability may be 9 or 10%. Evidence is accumulating that mitochondrial sequences are essentially not represented in plant YAC libraries. The mitochondrial DNA may have some intrinsic structural features that make it difficult to propagate in yeast (Lonsdale, 1994).

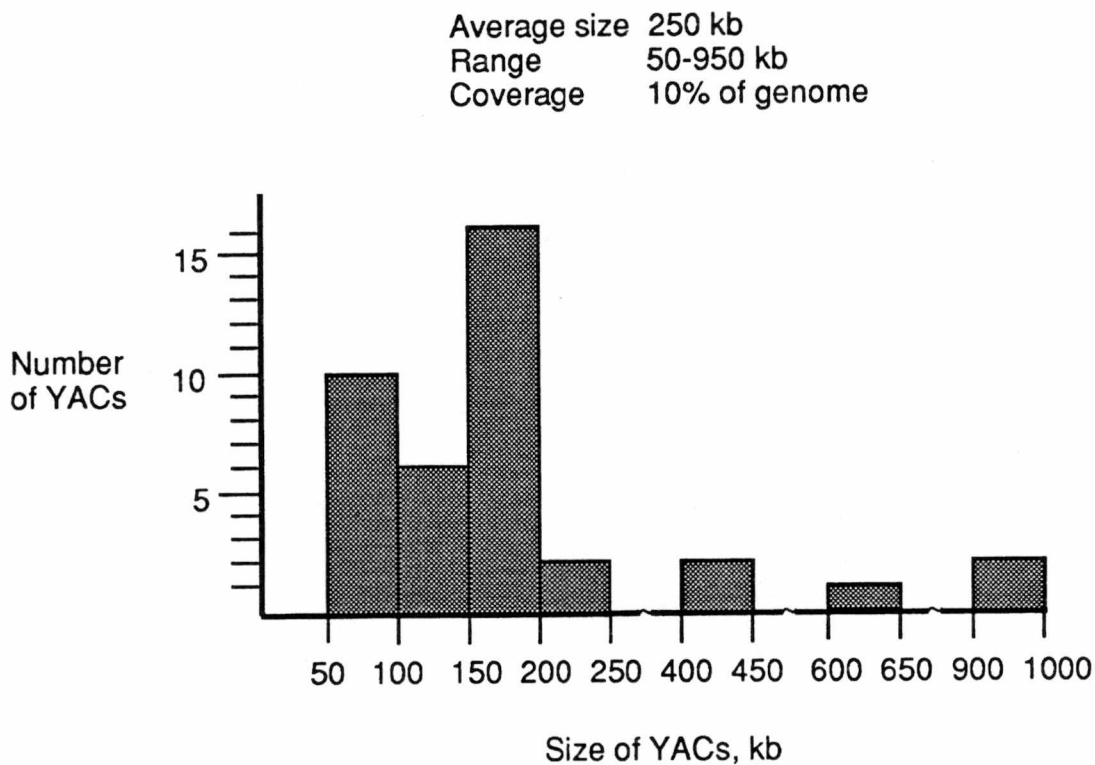


Figure 8.6. Current status of the soybean YAC library. Included are all the clones that were checked from transformation experiments whose average insert sizes were judged to be large enough to be included for permanent storage in the library. 96 clones chosen at random from the library in September 1994 had an average insert size of 175 kb (see text).

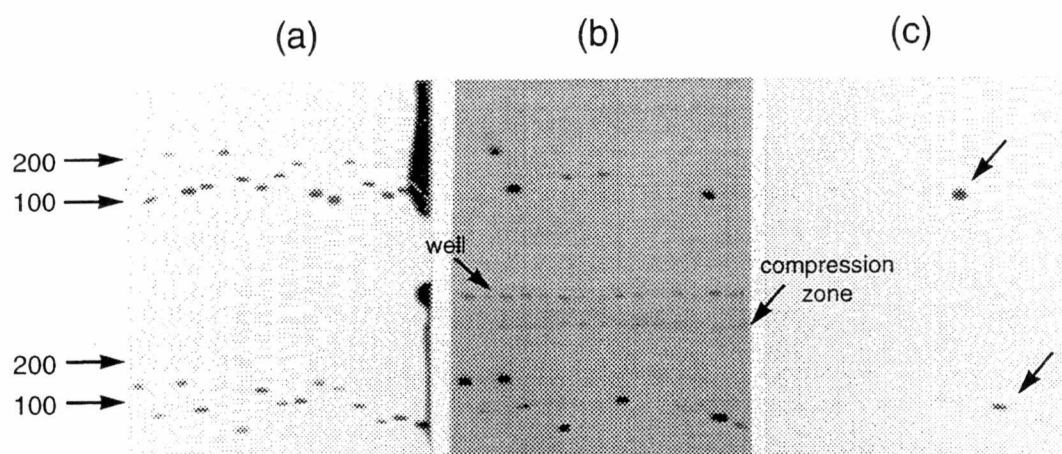


Figure 8.7. Effect of PFGE size selection on the average size of recombinant YACs. YAC clones in the top row of the panel were produced from insert DNA which was size selected after partial digestion and before ligation (3 second switch interval). YAC clones in the bottom row of the panel were produced without size selection. Polyamines were used throughout the procedure in both transformations. In b), the wells and compression zones of each lane are visible because of homology of rDNA sequences on the largest yeast chromosomes to the total soybean DNA probe. The same membrane was probed consecutively with labelled pBR322 (a), total soybean DNA (b), and a cloned fragment from the mungbean chloroplast genome (c; clones a gift from R. Shoemaker, Iowa State University). The average size of clones produced with size selection of insert DNA is larger than without any selection (see Figure 5). Only clones of the type shown in the top row of the panel were added to the permanent library. 2 of the 30 clones analyzed, or about 7%, hybridize with the chloroplast probe (c). Although more YACs might be detected with other cloned fragments from the chloroplast genome, the average size of the soybean YACs is large enough that those that do contain chloroplast DNA should encompass most or all of the genome.

Screening by PCR

The most efficient way to screen the partial library for specific sequences is by using PCR (Green and Olson, 1990). PCR-based screening has been used to detect unique sequences in pools of YACs containing as many as 1920 individual clones (Green and Olson, 1990; Del-Favero *et al.*, 1994). Each PCR reaction in these experiments contained about 100 ng of pooled DNA, containing 2000-3000 copies of each individual YAC.

Each plate from the library was replicated into a second 96-well plate and individual clones were grown in separate wells in 200 μ l of 1.5xAHC medium. After 2 or 3 days when a pinkish layer of cells was visible in the bottom of each well, a 12-channel pipettor (Costar Corp., Cambridge, MA) was used to pool the YACs into a single tube. Two DNA agarose blocks were prepared from each plate according to the procedures described in Appendix 3. The plugs were dialysed against TE buffer and 1 mM PMSF to inactivate proteinase K and remove traces of detergent. They were then melted at 68°C for 20 minutes and mixed with four volumes of warm TE. The mixture was immediately vortexed to disperse the DNA evenly.

One microlitre of the pooled DNA from each plate was assayed by PCR. This was estimated to contain about 10,000 copies of each YAC on the plate. The pooled DNA was assayed for the ability to amplify yeast actin sequences (primers kindly donated by Dr. Henry York-Steiner, University of Tennessee; see Appendix 3 for primer sequences and PCR conditions). Two strong products between 500 and 1000 bp are obtained with these primers. The primers were able to amplify the same segments of the yeast actin gene from a 1 in 100 dilution in water of the pooled plate DNA, conditions that mimic the amplification of a unique sequence from the soybean genome cloned in a single YAC on that plate (Figure 8.8).

It was important to make sure that the yeast DNA in the pools did not interfere with the amplification of soybean sequences. Using primers for an alfalfa histone gene that gave strong products of the expected size when used to amplify soybean DNA (see below), the same products were obtained when 20 ng of soybean DNA (or about 5000 diploid nuclei, corresponding to 10,000 copies of a unique sequence in the genome) was amplified in the presence of an aliquot of yeast DNA equivalent to 1-2 μ l of pooled YAC DNA. This verified that yeast DNA did not inhibit the amplification of soybean-specific sequences.

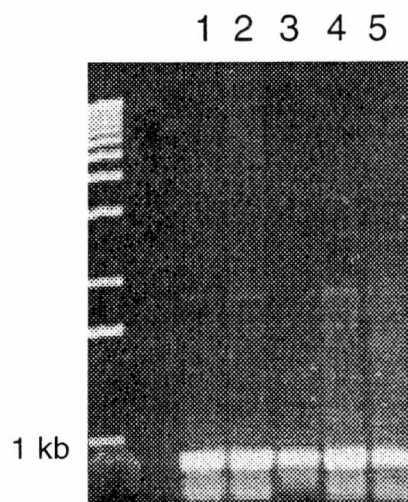


Figure 8.8. Amplification of pooled DNA from 96-well plates of the soybean YAC library with yeast actin primers. See Appendix 2 for a description of the procedure. Lane 1, plate 3; lane 2, plate 6; lane 3, plate 4; lane 4, 1 μ l of a 1:10 dilution of plate 4 DNA; lane 5, 1 μ l of a 1:100 dilution of plate 4. See text for discussion.

Only a fraction of the soybean genome is present in the partial YAC library, but this fraction should contain the expected number of copies of dispersed repetitive sequences from the soybean genome. Few such sequences from the soybean genome are known (see Chapter 4). Fluorescent In Situ Hybridization to soybean chromosomes has been performed using one dispersed repetitive sequence, SB120 (Zhu *et al.*, 5th Biennial Conference on the Molecular and Cellular Biology of the Soybean, July 25-27, 1994, Athens, GA). However, the probe for SB120 is not publicly available pending the submission of the sequence to the EMBL data base (Dr. M. Morgante, Udine, Italy, personal communication). In addition, the sequences of the different repeat units seem to be diverged, so it would be difficult to find conserved regions from which primers could be designed.

A single primer for a dispersed repetitive sequence from the soybean genome was obtained from Dr. K. Gordon Lark (University of Utah). This sequence was discovered as an RFLP marker with multiple alleles located on different linkage groups of the genetic map of soybean (Lark *et al.*, 1993). No information on the copy number of the repetitive sequence has been published, and the sequence of the primer was not provided. The primer amplified a prominent product from soybean DNA, but no reproducible positive signals could be obtained from the pooled YAC plates.

A concern with primers for repetitive elements is that it is not clear how homogeneous the sequence of the element is in the genome. It is possible that it is divergent, so that primers would amplify only one or a few elements, and not others that are present in the pools of clones but have slightly different sequences. Thus it is necessary to know how many separate elements were isolated and sequenced, and how the sequence of the primer was derived.

Southern hybridization of a filter replica of the library with cloned repetitive sequences might be a more reliable way to assay for the presence of those sequences in the genome (see below). This would overcome the limitation of PCR that it is too stringent to pick up slightly diverged copies of those elements. On the other hand, the considerable amount of background hybridization that can be encountered with colony hybridization might be particularly aggravating in a search for clones containing a repetitive element, and a large number of false positives would probably be encountered. In addition, it was observed that some sequences cloned from individual yeast chromosomes are homologous to sequences in the soybean genome (data not shown), and this could be problem because each of the

YAC colony patches contains many copies of all the yeast chromosomes as well as the extra soybean artificial chromosome.

The highly conserved H3 histone gene is present in multiple copies in the genomes of higher eukaryotes, including plants (Wu *et al.*, 1988; Chaubet *et al.*, 1992). No data on the representation of H3 sequences in the soybean genome could be found, but there are about 160 copies of the histone H3 gene in the alfalfa tetraploid genome (Wu *et al.*, 1992). Nothing is known about the organization of the genes, and the isolation of a YAC containing one or more copies could establish whether they are dispersed or clustered. If they are dispersed in the genome, nearly every 96-well plate representing 0.5-1% of the soybean genome would be expected to contain a copy of the gene. Primers were designed from the sequence of an alfalfa H3 gene (forward 5'-ATG GCA CGT ACC AAG CAA ACC G-3' and reverse 5'-C GAG CTG GAT ATC CTT AGG CAT-3'; Wu *et al.*, 1988). These primers amplify a prominent fragment from the soybean genome at about 400 bp, corresponding to the expected size of the product from the alfalfa genome (Figure 8.9). Two weaker products of about 700 and 900 bp are also obtained. The forward primer includes the first 22 bp of the alfalfa histone gene nucleotide sequence. In this region, there is 100% homology to the yeast H3 gene at the protein level (Müller and Schmitt, 1988). The reverse primer was chosen to discriminate against the yeast H3 gene. It begins 27 bp upstream of the termination codon, and is 22 bp long. In this region there are two amino acid differences between yeast and alfalfa, including one determined by the last codon at the 3' end of the selected primer.

The alfalfa H3 primers yielded strong amplification products with some of the pooled plate DNAs. However, these results were not always reproducible, and, on occasion, pooled DNA from each of the plates produced weak amplification products of the expected size. When the rows and columns of an individual plate (plate 3) were pooled in order to determine the exact 'well address' of a YAC containing a soybean histone gene, weak products of the expected size were obtained in all lanes. This faint product was also obtained when the H3 primers were applied to yeast DNA alone or used for direct colony PCR of yeast host AB1380 (Figure 8.9). Thus the alfalfa primers did not discriminate against the DNA of the yeast host. Although the product from soybean is much stronger, the amplification of even a faint product of the same size from the yeast host makes these primers less useful for the isolation of soybean histone gene containing YACs. The product amplified from yeast may represent a yeast histone gene with imperfect homology to the primers.

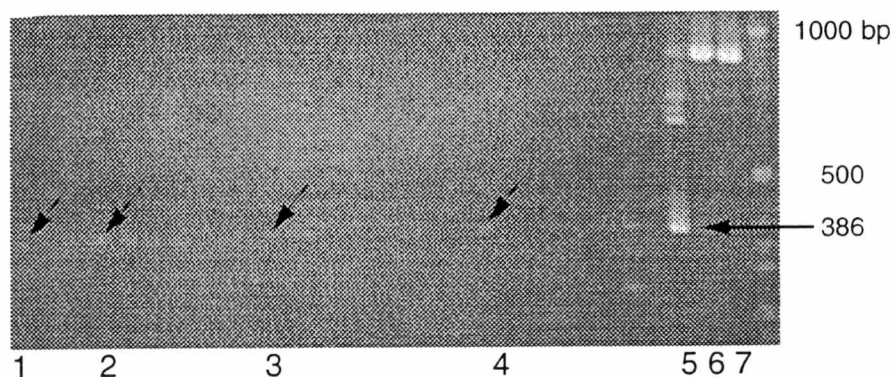


Figure 8.9. Amplification of rows, columns, and individual wells of a 96-well plate of soybean YACs using alfalfa histone gene primers. DNA from wells in the rows and columns of plate 3 was pooled and amplified using histone gene primers to determine the exact address of a well with a histone gene-containing YAC. Arrows indicate faint products in the numbered lanes. Lane 1, 10 ng of yeast host AB1380 DNA; lane 2, 5 μ l of one colony of AB1380 yeast lysed by boiling in 10 μ l water; lane 3, DNA from wells of column 10 from the plate; lane 4, DNA from well 10F; lane 5, positive control with 20 ng soybean DNA; lane 6, 20 ng yeast host DNA amplified with yeast actin primers (see Figure 8); lane 7, 5 μ l of one colony of AB1380 yeast lysed by boiling in 10 μ l water amplified with yeast actin primers

A more trivial explanation for the faint amplification product is that there was prevalent low-level contamination with soybean DNA in the solutions or equipment that were used to set up the amplification reactions. However, when the experiment was repeated and the amplifications were set up in replicate at different work stations, using different pipettes, pipette tips, and batches of autoclaved water, the faint product was still obtained. This leaves the possibility that the water in which the DNA from the 96-well plates was dissolved was contaminated. The fact that direct colony PCR of the yeast host using freshly sterilized water on the day of the experiment gave the same product argues against this explanation. If contamination with soybean DNA is invoked to explain these results, the problem would have to be very widespread in the laboratory.

The probability of isolating a YAC containing a unique sequence from the soybean genome from a partial library is certainly smaller (theoretically about 10%, see above). Only a small number of primer pairs for unique sequences are available in the laboratory. The following primer pairs were tried: pUTG-132a (Gresshoff, 1993), ENOD40 (Kruchinina *et al.*, 5th Biennial Conference on the Molecular and Cellular Biology of the Soybean, July 25-27, 1994, Athens, GA), and primers for two soybean specific sequences that are duplicated in the genome, A199 and pNAU163, that were sent to us by Dr. Paul Keim, Flagstaff, AZ. None of the primers reproducibly yielded the expected amplification products from any of the pooled plates of YACs. The occurrence of occasional false positives is to be expected in a laboratory where soybean DNA is extracted and handled on a daily basis, and emphasizes the need to reproduce positive results before going on to the next tier of the screening procedure, the pooling of YAC DNA from the wells in the rows and columns of individual plates. Aliquots of the pooled DNAs have been sent to the laboratory of Dr. Paul Keim, where they will be surveyed using primers for probes on the USDA/ISU RFLP map that are available in that laboratory. Any positives that are obtained will be reproduced in Knoxville. Pooled columns and rows of YAC-containing wells from the plate that produced the positive signal will then be tested with the primers. In this way it should be possible to identify an individual YAC containing the amplified sequence.

One reason for the failure to isolate YACs containing known sequences from the soybean genome could be that they are simply not represented among the clones that have been produced up to this point. Either there are too few clones, or they do not represent a random sample of all parts of the genome. Other reasons for the inability to find YACs containing repetitive or unique sequences could be technical. The primer concentrations used in these experiments, 0.1 μ M, are ten to 20 fold lower than is normally used for PCR

screening (*e.g.*, Green and Olson, 1990; Del-Favero *et al.*, 1994). These primer concentrations were sufficient for the amplification of specific products from 20 ng of soybean DNA, and from pooled YAC DNA 'spiked' with 20 ng of soybean DNA (representing the same number of copies of a unique sequence as would be the case if one of the clones in the plate contained such a sequence). However, these controls do not mimic precisely the situation where one YAC in a pool contains an amplifiable sequence. Thus, it could be useful to raise primer concentrations and perhaps modify the conditions of PCR, although the latter conform closely to published practices (Green and Olson, 1990).

Finally, it could be that melted, diluted DNA from the 1 μ l aliquot for PCR is taken is not uniform. In spite of melting and vortexing after the addition of warm TE buffer, the mixture is not completely homogeneous after freezing. The presence of small clumps of reaggregated agarose could cause variation between 1 μ l aliquots taken from the tube. This is a trivial point that could easily be overcome by boiling the DNA before removing the sample, or increasing the volume of the DNA in the reaction to 5 or 10 μ l.

Screening by Southern hybridization

All the YACs in the library have been replicated onto Hybond-N membranes for screening by Southern hybridization. Using a 12-channel pipettor (Costar Corp., Cambridge, MA), the clones from each plate from the library (6 total) were replicated by dipping the pipette tips into the wells of the plate and touching the tips to a membrane placed on YPD agar (see Appendix 3). After 24 hours the membranes were lifted from the surface of the agar and prepared for hybridization (Matallana *et al.*, 1992). In short, membranes were placed consecutively on 3MM paper saturated with solutions containing lyticase and β -mercaptoethanol (spheroplasting), and 0.5 M NaOH (lysis and denaturation). Before hybridization, the membranes were neutralized by rinsing with 0.5 M Tris, 10xSSC, and cell debris was removed by rubbing with gloved hands in a tray containing 2xSSC.

Using a 'multiprobing' approach, all the membranes were hybridized with a mixture of six probes from the USDA/ISU map (pA-36, pA-69, pA-89, pA-130, pA-381, and pK-9). Since all these probes represent duplicated sequences in the soybean genome, the probability of isolating a YAC with sequences homologous to one of the probes should be double 10%, or 20%. A number of strongly hybridizing clones can be seen, and these will be retrieved from their wells in the library to determine which of the probes in the mixture is giving the positive hybridization signal (Figure 8.10).

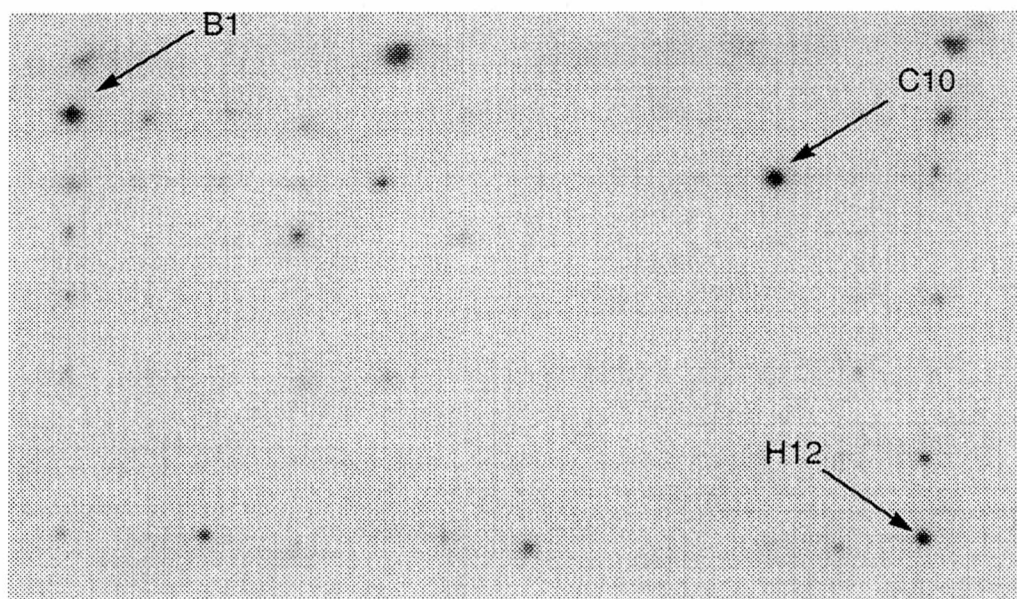


Figure 8.10. Screening of filter replicas of the soybean YAC library by Southern hybridization with RFLP probes. The entire library was replicated onto Nylon membranes as described in the text, and each membrane was probed with a mixture of soybean RFLP probes (see text). Membranes were prehybridized with unlabelled pBR322 DNA to suppress signals from the YAC vector arms. A number of colonies on this filter (plate 3) produce strong signals, as indicated with arrows. The probability of isolating a unique sequence from the soybean genome on a single filter is about 2%. Since five out of the six probes used have 2 or more copies in the genome, the probability that at least one colony per filter will hybridize to one of the probes is as high as 20%.

Potential bias in the library

An important consideration is whether or not all areas of the genome are represented equally in the library. An unequal distribution of *EcoRI* restriction sites in the genome could cause non-representation of some areas of the genome. This was found to be the case with the EG YAC library of *Arabidopsis*. Although this library contains three genome equivalents, for 8 out of 45 probes from the tops of chromosomes 4 and 5 no positively hybridizing YACs could be identified (Schmidt and Dean, 1993). The clones could be isolated from other *Arabidopsis* libraries which were made using a different restriction enzyme. Thus the use of a single restriction enzyme could result in non-representation of some parts of the genome. That this could be the case with *EcoRI* in soybean is suggested by the observation that even when a large excess of the enzyme is added and distinct bands signalling a complete digest appear, some very high molecular weight fragments remain to be seen (Figures 8.2, A3.1). These very large fragments may represent stretches of DNA that are unclonable due to a scarcity of *EcoRI* sites in them. This phenomenon is an argument in favor of experimentation with other vector systems such as BACs, which are constructed using *HindIII* or *BamHI* partial digests, or with YAC vectors that have a different cloning site.

In summary, although it is possible to show that the YACs contain soybean DNA by hybridization with total nick-translated DNA, by hybridization of the ends of YACs to restriction digests of soybean DNA (see below), and by Fluorescent In Situ Hybridization of whole YACs to soybean metaphase chromosome spreads (see below), the isolation of specific YACs containing nuclear DNA sequences is still necessary to check the representativeness of the library.

ISOLATION OF GENOME-WIDE REPETITIVE SEQUENCES FROM YACs

Because of the shortage of available dispersed repetitive sequences from the soybean genome with which to characterize the YAC library, it was decided that such sequences might be isolated from the YACs themselves. If sequences from one YAC are represented in a proportion of the other clones, their occurrence in the library could be measured against their overall copy number in the genome. This approach, while apparently circular, could provide one measure of the representativeness of the library. In addition, such a sequence would be interesting because its interspersions with low copy number sequences and coding regions could be studied in discrete, relatively large cloned segments of the genome (YACs). Knowledge of genome organization and the interspersions pattern of repeated

sequences is important for positional cloning experiments (see Chapter 7), and until now for soybean has been based mostly on DNA reassociation studies carried out on the late 1970s (Gurley *et al.*, 1979). In addition, repetitive sequences could have application for cytogenetic analysis by chromosome painting (Carter, 1994; Zhu *et al.*, 5th Biennial Conference on the Molecular and Cellular Biology of the Soybean, July 25-27, 1994, Athens, GA).

Three clones were chosen from among the 96 clones characterized for insert size and intensity of hybridization to total soybean DNA (see above). These clones, E10 (245 kb), B1 (220), and A7 (300), all produced moderate to strong signals on hybridization to total soybean DNA. When total DNA from these YACs was digested with *Hind*III and the digests were probed with labelled total soybean DNA, distinct strongly hybridizing bands were seen (Figure 8.11b).

To avoid cloning unwanted yeast DNA, the YACs were resolved from the yeast host chromosomes on a preparative LMP agarose gel (Figure 8.11a). They were cut from the gel, melted, treated with gelase, and restricted with *Hind*III. Following precipitation assisted by glycogen carrier, the digested DNA (about 200 ng) was ligated to 20 ng *Hind*III cut pBS+ vector (Stratagene, La Jolla, CA) in a final volume of 10 μ l.

The ligation mixtures were diluted with distilled water to a volume of 100 μ l. 2 μ l of the diluted mixture were used to transform 40 μ l of competent *E. coli* DH10B cells using the electroporation procedure (Sambrook *et al.*, 1989). The transformation mixtures were spread on two LB agar plates containing X-Gal, IPTG, and 100 μ g/ml ampicillin.

For each of the YACs, 180 white bacterial colonies carrying recombinant pBS+ molecules were picked in replicate onto LB agar containing ampicillin and onto a membrane layed on top of LB agar. After 12 hours growth at 37°C the membranes were prepared for colony hybridization (Sambrook *et al.*, 1989).

The three membranes were hybridized to total labelled soybean DNA (MegaPrime Kit, Amersham) and exposed to Kodak X-Omat AR film. After overnight exposure 3 to 4 colonies on each membrane could be seen to hybridize intensively (Figure 8.11c). It would appear that the cloned fragments in these colonies, representing 5-10% of the soybean DNA in the original YAC, are responsible for the majority of the hybridization of those YACs to total soybean DNA.

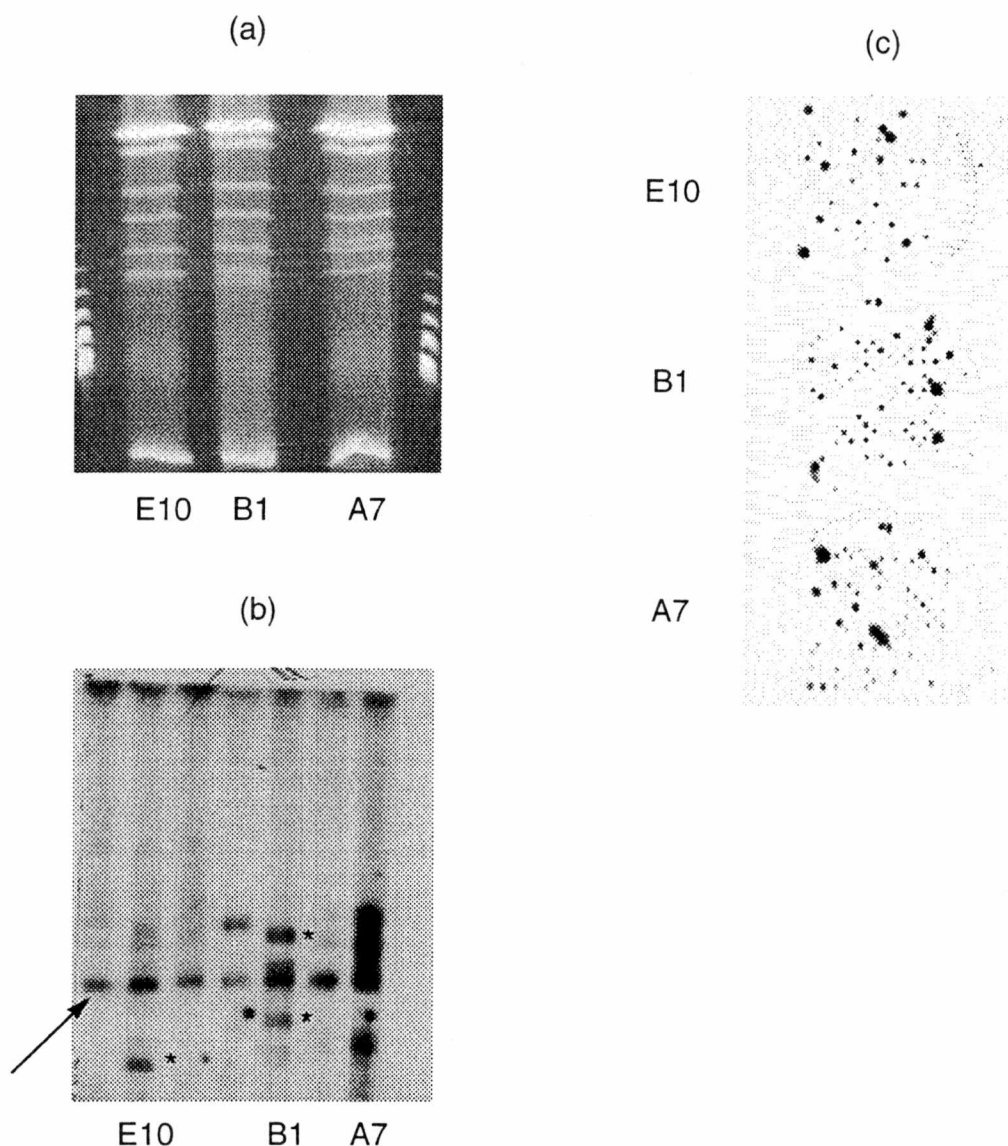


Figure 8.11. (a) Three YACs that gave intense signals when hybridized to total soybean DNA were resolved from host chromosomes on a preparative LMP agarose gel. E10 is the same size as the smallest chromosome of the yeast host. (b) When total YAC DNA (host chromosomes + artificial chromosome) was digested with *Hind*III and probed with total soybean DNA, several distinct hybridizing bands were detected (*). The arrow indicates a hybridizing band that is common to all YACs. These are yeast host rDNA sequences that are homologous to sequences in the soybean genome. (c) *E. coli* DH10B transformed with YAC *Hind*III fragments cloned into pBS+ were arrayed on a membrane and hybridized with total soybean DNA. Some of the recombinant molecules contain sequences that are reiterated in the soybean genome, as judged by the intensity of hybridization, and these will be analyzed further.

The plasmid inserts from these colonies are candidates for hybridization to panels of YACs and to complete and partial digests of soybean genomic DNA, as well as for the determination of their copy number in the genome (according to e.g. Gullis *et al.*, 1984). Moderately hybridizing recombinant bacterial colonies from the same YACs could be analyzed in the same way.

FLUORESCENCE IN SITU HYBRIDIZATION USING YACS

Four clones were chosen from among the 96 clones characterized for insert size and intensity of hybridization to total soybean DNA that showed virtually no signal on hybridization with total soybean DNA, and thus were judged to contain little or no repetitive DNA. These YACs were isolated by preparative gel electrophoresis as described above (Figure A5.5). Agarose slices containing the isolated YACs, as well as the YAC colonies themselves, were sent to Dr. Paul Keim (Northern Arizona University, Flagstaff, AZ) for use in fluorescence in situ hybridization (FISH) on metaphase spreads of soybean chromosomes.

The purpose of these experiments was to study duplicate regions in the soybean genome. Until now, duplicated regions in the soybean genome have been detected by two-factor inheritance of traits, the identification of duplicated genes and RFLP markers, and by physical mapping (see Chapter 6). FISH with whole YACs containing duplicated sequences would extend the size of those regions that could be physically studied at the molecular level by an order of magnitude. FISH could show whether a segment as large as a few hundred kb hybridizes primarily to two loci (one homologous to the YAC, the other homoelogenous), as is the case for shorter cloned duplicated regions, or whether duplications are more interspersed and involve more than two regions of the genome. Within those YACs, the interspersion of duplicated sequences with unique sequences and repetitive elements could be studied.

So far, results have been obtained with one of the YACs, A12 (Figure 8.12). Four bright hybridization foci can be seen, which is what would be expected for a YAC comprising mostly duplicated sequences (in a diploid nucleus, two signals would be expected on the chromosomes that are homologous to the YAC, and two for the chromosomes that contain homoelogenous sequences).

The left and right ends of YAC A12 have been amplified by vectorette PCR. These ends will be labelled and hybridized to restriction digests of DNA from *G. max* (L.) Merrill cv.

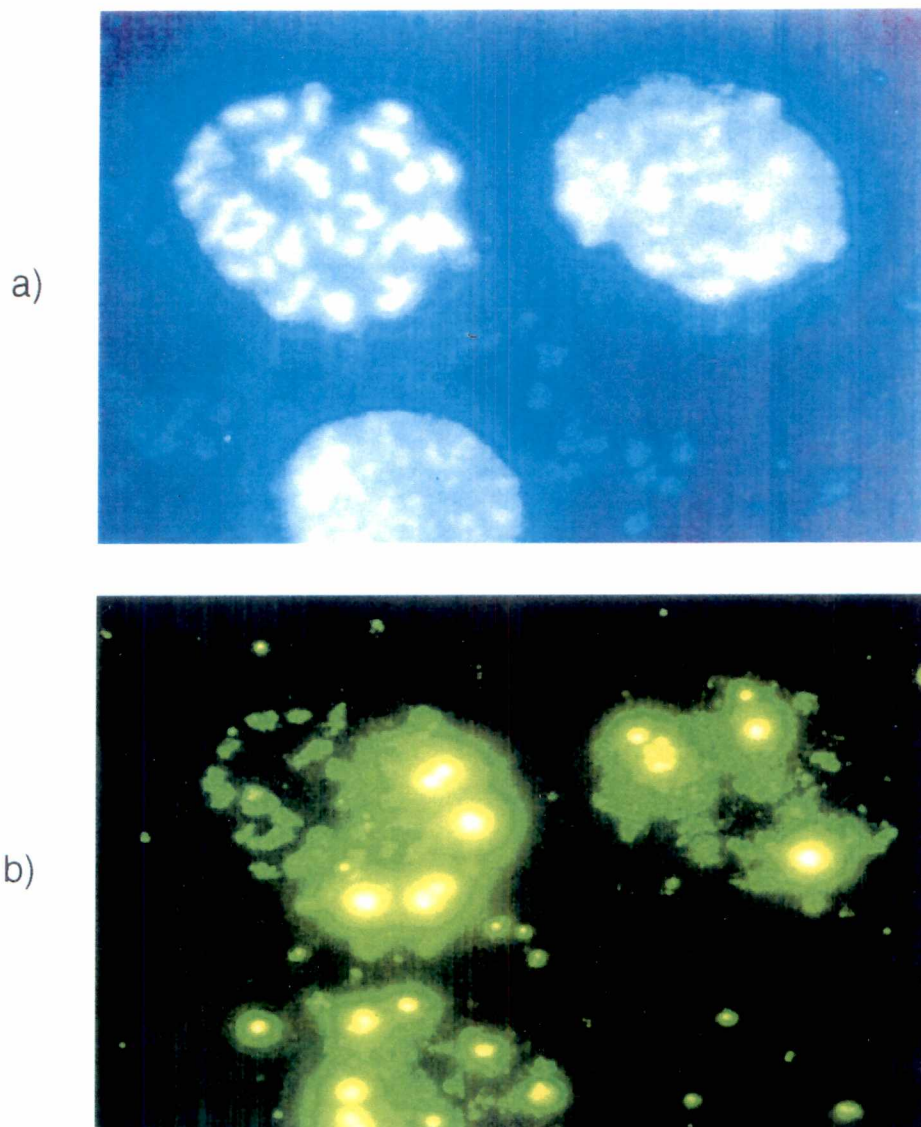


Figure 8.12. Fluorescent In Situ Hybridization (FISH) of YAC A12 to soybean metaphase chromosomes. a) Nuclei stained with DAPI, a DNA-binding dye which fluoresces in the blue range. Individual chromosomes in the nuclei can be distinguished. b) FITC FISH image of the same nuclei. The FISH image appears on four chromosomes in the nucleus at the top lefthand corner. YAC A12, which was selected because it appeared to have a low repetitive DNA content, provides intense staining at relatively few chromosomal locations. Four hybridization foci suggest that the YAC contains duplicated sequences and detects homologous and homoelogenous chromosomal regions. FISH images were produced by Liang Shi in the laboratory of Dr. Paul Keim at the University of Northern Arizona.

Bragg and *G. soja* (PI468.397), and *G. max* (L.) Merrill cv. Minsoy and Noir1, the parents used in the genetic map produced at Utah (Lark *et al.*, 1993). If polymorphisms can be found, they will be mapped relative to each other. If polymorphisms for the opposite ends of the YAC are genetically linked, as expected, any recombination between them will allow a direct and very precise measure of the relationship between physical and genetic distances in that cloned segment of the genome. There are two explanations for the potential observation that polymorphisms for the ends map to different linkage groups of the genetic map. One is that the YAC is a chimera, and contains fragments from unrelated, non contiguous parts of the genome. Alternatively, the polymorphism for one end might correspond to the chromosomal region that is homologous to the YAC (that is, the region that is cloned in the YAC), and the other could map to the related, duplicated region located on a second (homologous) chromosome.

STABILITY OF YACS CONTAINING SOYBEAN DNA

It is known that some repetitive sequences are not stable in YACs and that YACs containing such sequences can undergo deletions or rearrangements. This represents a potential problem for soybean which has a high proportion of interspersed repetitive elements in the genome (Chapter 7).

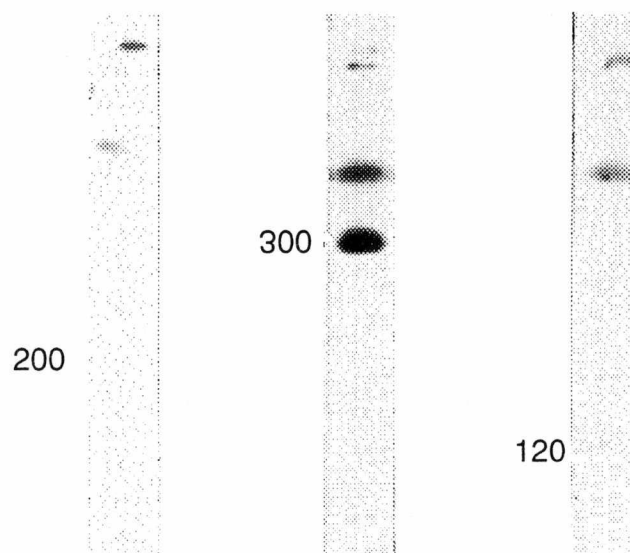
To assess the stability of soybean YACs, we checked whether their size and integrity were preserved after a known number of generations of growth in selective and rich media. An experiment was designed to mimic what happens during storage and handling of a YAC library as individual clones are streaked out repeatedly and propagated in liquid culture. Clones were selected that contained different proportions of repetitive sequences from the soybean genome, as judged by their intensity of hybridization to labelled total genomic DNA (Figure 8.13a).

The YACs were resolved on a pulsed field gel by electrophoresis with a pulse time of 5-20 seconds. After Southern transfer, the membrane was probed with labelled pBR322 DNA (Figure 8.13b).

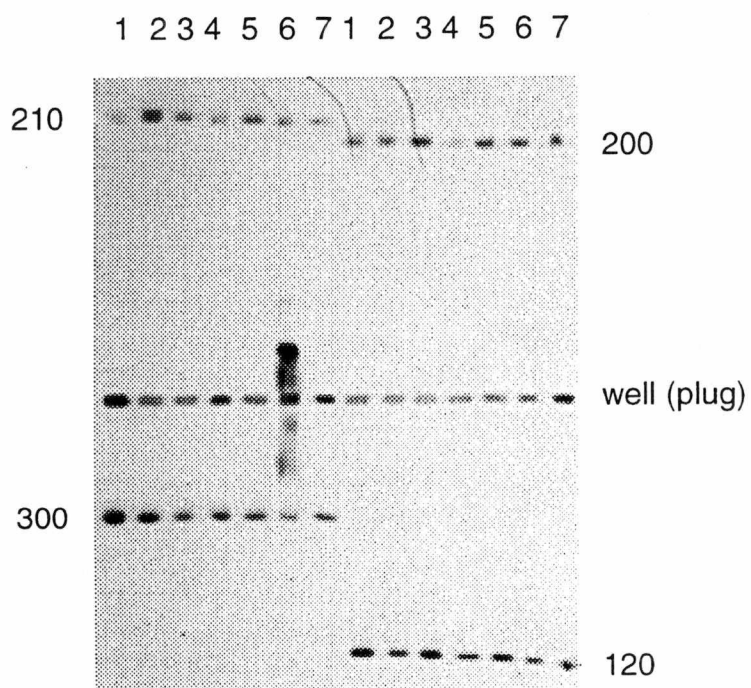
The soybean YACs on the whole did not show visible rearrangements or changes in length. The clone with high repetitive DNA content was as stable as the YACs containing primarily low copy number sequences. In lane 6 a small extra band is seen which could be a derivative of the 210 kb YAC. The reason for its intense hybridization is probably that shorter DNA molecules are transferred more efficiently than large molecules.

Figure 8.13. a) A 200 kb clone which was clearly visible on ethidium bromide stained gels showed no signal after 70 hr exposure. This YAC was judged to contain primarily unique sequences. A 300 kb YAC showed an intense signal, and a 120 kb clone produced a band of intermediate intensity. Hybridization to the compression zone is caused by homology of soybean rDNA to sequences on chromosome XII (at 1000 kb) of the yeast host strain. A YAC of 210 kb was also checked for stability, but had not been characterized for repetitive DNA contents. b) A DNA plug was prepared from each YAC patched onto double selection medium from the primary transformation plates (lane 7). The patch was streaked to single colonies. One colony was used to inoculate 10 ml of YPD medium (lane 3) and a second was inoculated into Ura⁻ liquid medium (lane 6). When the cultures reached saturation, DNA plugs were prepared. A third colony was frozen in YPD/20% glycerol and streaked out to single colonies the next day. Again three colonies were picked, of which two were used to prepare DNA plugs after growth in YPD and Ura⁻ liquid medium (lanes 5 and 2), and a third was frozen. This process was repeated one more time to produce the plugs in lanes 4 and 1. Each plug represents 30 generations of division starting from an individual YAC containing cell.

(a)



(b)



This experiment does not provide a real estimate of the stability of the YACs (e.g. the rate of loss of artificial chromosomes under nonselective growth conditions). However, it does show that, for practical purposes, YAC-containing yeast cells retained the clones faithfully whether they were grown under selective conditions or in rich medium.

ANALYSIS OF THE ENDS OF YACS TO ESTIMATE THE FEASIBILITY OF CHROMOSOME WALKING IN SOYBEAN

The isolation of the ends of YAC clones is important for a number of reasons (see also above, Fluorescence In Situ Hybridization using YACs). First, for chromosome walking experiments, the ends of clones are valuable for isolating overlapping clones in the library by hybridization. Second, the ends of clones could be a valuable source of additional markers in a particular chromosomal region.

Since the soybean genome consists of 30% repetitive DNA, it was a concern that that chromosome walking using YACs would be too arduous because end clones would hybridize to many unrelated YACs with cognate repetitive sequences.

To investigate this, the “vectorette” procedure was used to amplify the ends of three YACs (Figure 8.14a; see Appendix 3). Two of these, a 120 kb clone and the 300 kb clone used in the stability experiment, were judged to have high repetitive DNA contents by their intensity of hybridization to soybean DNA. A 200 kb YAC which contained largely single copy sequences was also used.

To verify that the amplification products represent the ends of the artificial chromosomes, the bands obtained from the 120 kb YAC (amplification products not shown) were sequenced using the Silver Sequence™ kit (Promega; Figure 8.15). Sequencing primers were the same as those that were used for vectorette amplification of the ends. Within the sequence the *Eco*RI cloning site was discovered (underlined). To one side of the cloning site were sequences from inside the vector arms (Riley *et al.*, 1990), and to other was soybean DNA (as shown by hybridization, Figure 8.14b).

The amplified YAC ends were cut from LMP agarose and hybridized to *Eco*RI and *Hind*III digests of *Glycine max* and *G. soja* DNA (Figure 8.14b). One of the four ends (1) contains sequences that are reiterated many times in the soybean genome. Another (3) hybridizes strongly to a single genomic *Eco*RI fragment and detects a number of others at a

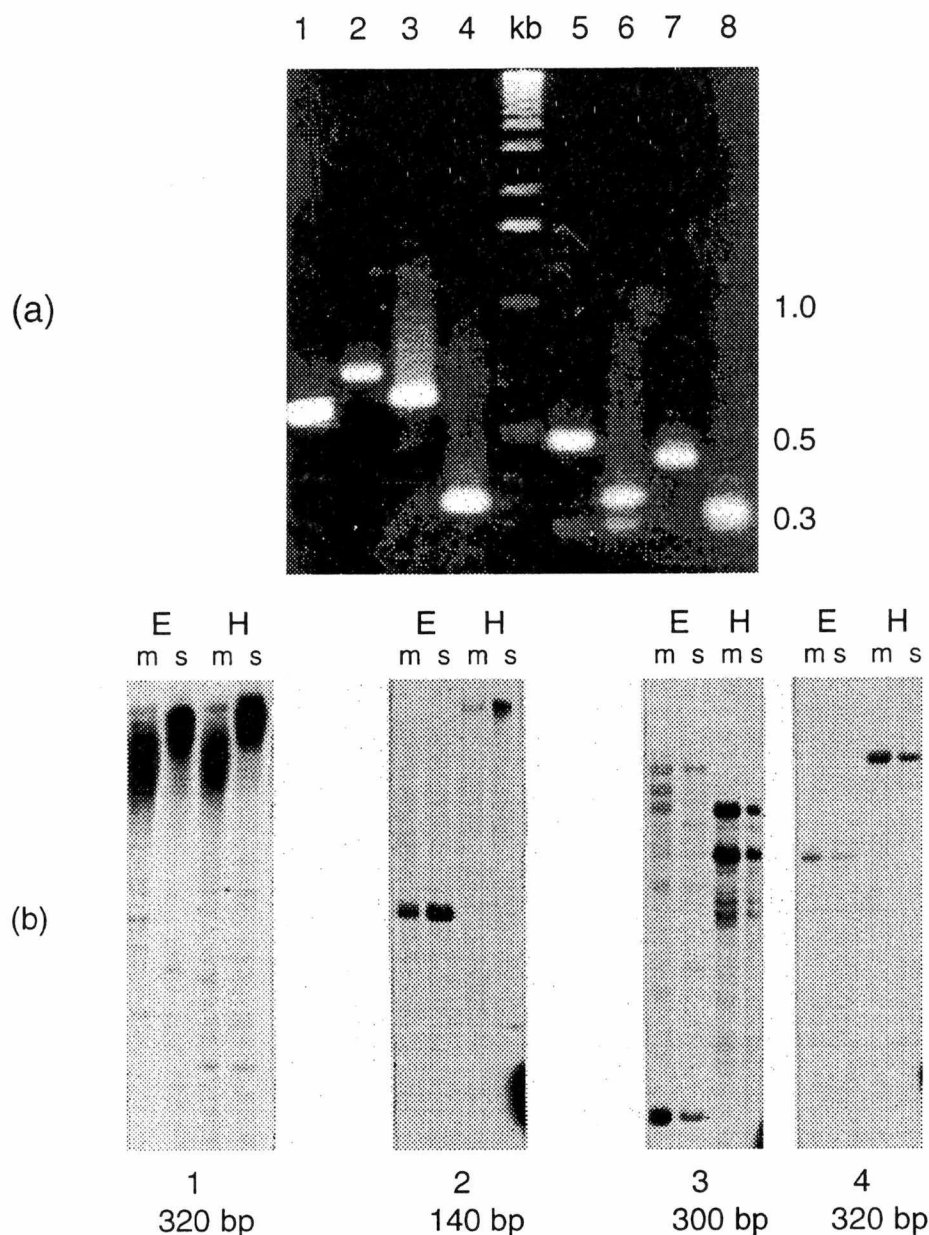


Figure 8.14. a) "Vectorette" amplification of the ends of YACs. Lanes 1 and 5 are the left and right ends of the 200 kb YAC with primarily low copy number sequences. Lanes 4 and 8 are the left and right ends of the strongly hybridizing 300 kb YAC. b) The amplified YAC ends were cut from LMP agarose and hybridized to *Eco*RI (E) and *Hind*III (H) digests of *Glycine max* (m) and *G. soja* (s) DNA. The purpose of this experiment was to check the copy number of the YAC end sequences in the genome. 1, right end of the strongly hybridizing 120 kb YAC; 2, left end of the strongly hybridizing 300 kb YAC; 3, left and right ends of the 200 kb YAC containing primarily low copy number sequences. The number of base pairs of soybean DNA in the labelled amplification product is indicated below the corresponding blot.

YAC/vectorette left PCR-product, YBE29-LFTVECT

5-TCCTCTATTA GTACGAGGTG GAAGAGTGTA AGGGCTCACT ATAGGGTCAG AATTTGGGGT
AGAAGAGACA TTGCATTGTA TATAAGTGGA AACATAGTTT TGTGTATTAG CAACATAATG
AAAAAGGCGT AAGAGTTTCA GGTGATACCT CTACAAAATT ATTTTCAGAT TGGATCAATA
GTGGATTTAA GGTTGAAGGA GAGATTATAC TATCATCTTC GAOCATATTC TCATTTACAT
TATTCTAAAT TTGTCCTCAT GTTCTTATCT CCATTATCTT GAATTCGTAG TGAAATTAAA
GTCTTGCGCC TTAACCAACT TGGCTACCGA GAGTATTTAA TTGTT-3'

YAC/vectorette right PCR-product, YBE29-rtvect

5-GTGTGGTTTT ACACATTACT TGTTTGTGCA TGAATTGTTG AAGGCATGCT AGAATAAGTT
TTTTTAGTTT TGAATAAGGG TAGGTTTCTC TTAAGCTCTT ATTCACAGAG GACCOCTAGGG
CTGGGTACCT TACTCTCTTT TTCTGGGGTA GGAAGTGAAG TTGCTTCCTG ATTGCTTATA
AGAATTCCGT AATCTTGAGA TCGGGCGTTC GACTCGCCCC CGGGAGATTT TTTTGTTTTT TT-3'

YAC/vectorette right PCR-product, YBE29-rtyac (complementary to YBE29-rtvect)

5-AGATTACCGA GAATTCTTATAA GCAATCACAA GCAATCTCAG TTCTACCCC AGAAAAAGAG
ACTAAGGTAC OCAGCCCTAG GGTCTCTGT GAATAAGAGC TTAAGAGAAA CCTACCCCTA
GCCAAAATA AAAAACTTA TTCTAGCATG CCTTCAACAT TCATGCACAA ACAAGTAATG
TGTAACCA CACGAAAAAG TAGAGTCAAA GATAGACACA ACCTGATCAA TCCTTTTCGC
AAAAACCTTG CTTGACTGAA TGAGGTGTCT TCTTGAAGTC AAGAAAGATC ATCTCACTAT
CACAAAGT-3'

Figure 8.15. Sequence of the ends of YAC YBE29, a 120 kb YAC cloned in the vector pYAC4. The procedure for obtaining the ends by vectorette PCR is described in Appendix 1. Both strands of the right end product were sequenced. The *Eco*RI cloning site is seen, with soybean DNA 5' to it in YBE29-rtvect, and 3' to it in the complementary strand, YBE29-rtyac. Sequencing by Alexander Kolchinsky.

much lower intensity. Two others (2 and 4) are uniquely represented in the genome. This is comparable to the overall fraction of repetitive DNA in the genome.

While the right end of the strongly hybridizing 120 YAC is highly reiterated, the sequence from the end of the strongly hybridizing 300 kb YAC detects unique fragments in restriction digests of genomic DNA. On the other hand, one end of the low copy number 200 kb YAC hybridizes to several cognate sequences in the genome, although only one very prominent band is seen on hybridization to *EcoRI* digests. Thus there is no correlation between the overall repetitive DNA content of a YAC, and the probability that the ends of the clone will contain repetitive sequences.

The ends of soybean YACs do not appear to be enriched for repetitive sequences (which could be the case if a common repetitive element contained an *EcoRI* site) and chromosome walking could be a feasible complement to the preferred objective of landing on a specific large YAC using tightly linked markers.

For mapping of YAC ends, it is necessary that the amplified or cloned ends reveal polymorphisms between the parents of a cross. In the case where the end reveals more than one copy in the genome, mapping would be complicated by the difficulty of not knowing which band on a genomic Southern corresponds to the end of the YAC, and which to a cognate sequence in another part of the genome.

CHAPTER 9

CONCLUSIONS AND PERSPECTIVES

HMW-DNA methods are necessary for positional cloning because of the gap in resolution that exists between genetic and physical mapping (see Chapters 3 and 4; Appendix 5). While saturating the relevant area of the genome with molecular markers is useful and necessary, segregation analysis will only refine the location of the gene to within a fraction of a centiMorgan. Even a total lack of recombination between a gene and a linked marker represents a chromosomal segment of tens or hundreds of kb in which the gene might be located.

In this thesis some of the HMW-DNA methods that will be necessary for map-based or positional cloning of genes from soybean have been established. These include procedures for cloning, separation, and analysis. This work was started with the goal of cloning the *nts* gene, but the technology will be relevant to the pursuit of other genes, too. In particular, however, soybean will be an important source of genes controlling developmental processes, such as nodulation and nitrogen fixation, that are unique to legumes. (see Chapter 1 and Appendix 2).

The isolation and characterization of *nts* mutants led to advances in the understanding of nitrate tolerance and autoregulation that would not have been possible otherwise. For example, it was shown that the early effects of nitrate in soybean are exerted by arrest of the nodule initiation process, and that the mechanism of inhibition intersects with the pathway by which the plant normally regulates total nodule numbers in the root system. (Day *et al.*, 1989). Second, interspecific grafts using the mutants showed that the autoregulation signal originated in the shoot (Delves *et al.*, 1986). Shoot apex removal experiments later demonstrated that it is probably the leaves and not the apex region itself that is the source of the signal (Delves *et al.*, 1992). The mutants also provided a genetic basis for the observation that nitrate inhibits early and late nodulation events by different mechanisms (Gresshoff *et al.*, 1988). The full potential of the mutants for experimentally unravelling the physiological basis of nitrate inhibition and autoregulation has probably not been realized. For example, in spite of the lack of reproducibility in the past, comparisons of the vascular sap between wild-type and mutant plants could reveal informative differences in composition, and might even lead to the discovery of components that could be candidates for the shoot-derived autoregulation signal. In addition, some recently developed, very sensitive techniques for comparing and measuring gene expression,

discussed in Chapter 2, have not been applied to the mutants yet.

Having analyzed the phenotype of *nts*- plants, and having mapped the gene onto a linkage map, the next step in dissecting the relationship between the gene and the phenotype is to identify the relevant DNA, RNA, and protein sequences. This would open a whole new range of questions to be studied. A putative function of the gene could be suggested by a search of gene data bases (Bork *et al.*, 1994), and the role of homologous genes in other legumes and non-legumes could be investigated. Tailor-made chimeric genes using the promoter region of *nts* and a reporter gene could be used to probe the importance of sequences in the upstream region in regulating the time and site-specific expression of *nts* (de Bruijn and Schell, 1993). The contribution of different parts of the gene to the plant's phenotype could be probed by making specific changes in the nucleotide sequence and reintroducing the gene into the plant (reverse genetics; Berg, 1993). In short, not having the cloned gene in hand is presently the main obstacle to understanding the molecular basis of the action of *nts*.

Since the approximate location of the *nts* gene relative to a linked marker (pUTG-132a) is known, the most immediate obstacle en route to the gene is the isolation of large cloned DNA fragments (YACs) containing pUTG-132a or another tightly linked marker.

The work described in this thesis shows that it is possible to make large soybean YACs, and that they can be used to address a variety of questions concerning the structure and organization of the soybean genome. However, the construction of a comprehensive YAC library will be arduous, and represents a large investment in a technological bottleneck (it should be mentioned that the investment has been made with apparent success in other important crop plants with large, complex genomes; see Chapter 7).

Many permutations of the YAC cloning procedure were tried in addition to what was described in Chapter 8 of this thesis. For example, having shown that the linked marker pUTG-132a detects a 180 kb *SaII* fragment on Southern blots of pulsed field gels, YAC vectors with a *SaII* cloning site were obtained (pJS97/98, Shero *et al.*, 1990; pYACRC, Marchuk and Collins, 1988). This fragment does not necessarily contain the *nts* gene, but the ends of the clone would provide additional molecular landmarks for physical and genetic mapping in the relevant chromosomal region. Theoretically, the cloning of fragments generated by a complete digest with a rare-cutting enzyme such as *SaII* is more efficient than the use of partial digests, because a higher proportion of the DNA molecules will have two suitable cohesive ends, and the digestion does not have to be meticulously

titrated. However, even with PFGE size selection of the *SaII*-digested DNA, very few clones of the expected size were obtained. Perhaps entrapment of small molecules in the compression zone of size selection gels is especially severe with complete *SaII* digests, because the vast majority of fragments produced by this enzyme are much smaller than 180 kb. Our experience confirmed that the conventional and most widespread cloning method using partial *EcoRI* digests and the YAC4 vector is also the most reliable.

Given the low percentage of successful transformations, either the number and scale of transformations set will have to be increased dramatically, or the underlying reason for the persistent appearance of YACs with small inserts will have to be determined. Maybe then compensating changes in the cloning and transformation procedures can be made.

If the underlying reason is peculiar to the sequence organization of soybean DNA, then YACs are not a suitable cloning vehicle for large segments of DNA from soybean. It is conceivable that the interspersed pattern of repetitive sequences in the genome presents a very suitable template for recombination by the yeast host. The fact that the yeast host does recombine sequences of introduced artificial chromosomes is suggested by work in which the rate of chimerism of YACs was not reduced when steps were taken to prevent ligation of unrelated insert molecules to one another during the ligation reaction (Wada *et al.*, 1994). However, it would have to be supposed that the recombinational activity of the host fluctuates from one batch of spheroplasts to the next, because some transformations do yield numerous clones of the expected insert size, and even very large clones are stably maintained once they are established. Recombination might take place intramolecularly, causing 'looping out' of sequences between recombined sites, or between different molecules that enter the same spheroplast. A simple experiment to check for the importance of recombination would be to ligate partially digested size-selected DNA separately to the two artificial chromosome arms. By mixing the ligated DNA only at the point of transformation, any resultant Ura+Trp+ clones with both centric and acentric artificial chromosome arms would have to be the result of intermolecular recombination events between two molecules that entered the same spheroplast. In addition, large stable YACs from previous experiments could be retransformed into new batches of spheroplasts. If the clones undergo deletions and rearrangements upon retransformation, this would be evidence of batch to batch variation in the spheroplasts.

Fluctuations in transformation efficiency and insert size are widely documented in YAC library construction, and are difficult to explain (Foote, 1993). Perhaps the introduction of a 'recovery phase' in the transformation procedure, in which the spheroplasts are incubated

in YPD medium with 1M sorbitol for 30 minutes (Albertsen *et al.*, 1992) could alleviate some of the variation from one transformation to the next.

At the same time, it would be useful to invest in other cloning vectors. Although no plant BAC libraries have been reported in the literature, a number of them have been constructed, reportedly with relative ease (Pamela Ronald, UC Davis, personal communication). The average insert size of a rice BAC library was 120 kb, with most clones being 60 kb in size. Three months were sufficient to produce a library with 95% coverage of the rice genome. The rate of chimerism in BAC libraries is negligible, and clones are stable. Although the average insert size is smaller than what is reported for plant YAC libraries, the greater probability of finding relevant clones in a much shorter time makes this an attractive alternative. Most of the procedures for handling large DNA that have been developed in this laboratory can be applied easily to the production of BACs, and the vector and suitable host have already been obtained. If the construction of BAC libraries does proceed rapidly, it might be feasible to construct one for both the *nts382* and wild-type plants. Clones covering the same genomic library could then be scanned directly for differences and mutations using DAF technology (Caetano-Anollés *et al.*, 1991; Pena *et al.*, 1994).

Alternatives to the cloning of large DNA fragments might soon be feasible. The discovery of heat-stable polymerases with very high processivity has enabled the amplification of PCR products up to 35 kb in length (Barnes, 1994). In the future, this may be improved to permit the amplification of even greater lengths of DNA comparable to the average insert size of plant YAC and BAC libraries. It has already been shown that vectorette PCR (Riley *et al.*, 1990) can be adapted to amplify unknown or uncloned stretches of DNA that are adjacent to markers for which sequence information is available (Arnold and Hodgson, 1991). In theory at least, large consecutive steps can be taken from a linked marker in the direction of a gene of interest. The specificity with which PCR primers can be designed would obviate the difficulties that might be encountered in trying to build a contig of overlapping clones from a BAC or YAC library using end clones, especially in a genome with numerous repetitive or duplicated sequences. The large amplification products would be a source of new molecular markers for genetic and physical mapping, and could be scanned for expressed sequences in the same way that YACs and BACs are currently being screened for candidate genes.

Once a large segment of DNA is obtained encompassing the *nts* gene, candidate genes will have to be identified by screening cDNA libraries prepared from the relevant tissues at the appropriate time of development (see Chapter 3). The *nts* gene is known to be expressed in

the shoot portion of the plant, most likely in the leaves (Delves *et al.*, 1992). Thus cDNA libraries from the leaves of wild-type and mutant plants will be needed, and these are currently under construction.

Transformation technology is a critical ingredient in the success of the positional cloning project. This will be necessary both for complementation of the mutant plants with candidate gene sequences (a requirement for confirming the identity of the gene), and for eventual reverse genetics experiments to probe the function of different parts of the gene. Probably on the order of tens or hundreds of whole regenerated plants will be needed for these kinds of experiments, and this rate of production is not possible now. It may be possible to do complementation in another tropical legume that makes determinate nodules, such as *Lotus japonicus*, that is more readily transformable (Appendix 1; Handberg and Stougaard, 1992). As discussed in Appendix 1, saturation mutagenesis of *L. japonicus* is under way, and this may yield *nts*-like mutants that could be tested for complementation with candidate *nts* genes isolated from soybean. Alternatively, it might be feasible to transform the plants with an antisense version of each candidate gene, in the hope of inducing a mutant phenotype in the transformed plants. The use of antisense technology to probe the importance of nodulin genes has already been pioneered (Miao and Verma, 1993).

REFERENCES

- Akao, S., and Kouchi, H. (1992).** A supernodulating mutant isolated from soybean cultivar Enrei. *Soil Sci. Plant Nutr.* **38**, 183-187.
- Albertsen, H.M., Abderrahim, H., Cann, H.M., Dausset, J., Le Paslier, D., and Cohen, D. (1990).** Construction and characterization of a yeast artificial chromosome library containing seven haploid human genome equivalents. *Proc. Natl. Acad. Sci. USA* **87**, 4256-4260.
- Albertsen, H.M., Abderrahim, H., Cohen, D., and LePaslier, D. (1992).** Construction and Application of a YAC library. In *Plant Genomes: Methods for Genetic and Physical Mapping*, J.S. Beckmann, and T.C. Osborn, eds (Netherlands: Kluwer Academic Publishers), pp. 141-165.
- Ambrose, H.J., Byrd, P.J., McConville, C.M., Cooper, P.R., Ctankovic, T., Riley, J.H., Shiloh, Y., McNamara, J.O., Fukao, T., and Taylor, A.M.R. (1994).** A physical map across chromosome 11q22-q23 containing the major locus for Ataxia Telangiectasia. *Genomics* **21**, 612-619.
- Anand, R., Villasante, A., and Tyler-Smith, C. (1989).** Construction of yeast artificial chromosome libraries with large inserts using fractionation by pulse-field gel electrophoresis. *Nucleic Acids Res.* **17**, 3425-3432.
- Arnold, C., and Hodgson, I.J. (1991).** Vectors PCR: A Novel Approach to Genomic Walking. *PCR Methods and Applications* **1**, 39-42.
- Arondel, V., Lemieux, B., Hwang, I., Gibson, S., Goodman, H., and Somerville, C.R. (1992).** Map-based cloning of a gene controlling omega-3 fatty acid desaturation in *Arabidopsis*. *Science* **258**, 1353.
- Ashikawa, I., Katayose, Y., Momma, T., Miyazaki, A., Saji, S., Shimizu, T., Tanoue, H., Umehara, Y., Yasukochi, Y., Yoshino, K., and Kurata, N. (1992).** Physical Mapping of Rice Genome. *Rice Genome* **1**, 8-9.
- Barnes, W.M. (1994).** PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Nat. Acad. Sci. USA* **91**, 2216.
- Berg, P. (1993).** Co-chairman's remarks: reverse genetics: directed modification of DNA for functional analysis. *Gene* **13**, 261-264.
- Birren, B., and Lai, E. (1993).** In *Pulsed Field Gel Electrophoresis: A Practical Guide*, (San Diego, CA: Academic Press, Inc.).
- Blackhall, N.W., Hammatt, N., and Davey, M.R. (1991).** Analysis of variation in the DNA content of *Glycine* species: A flow cytometric study. *Soyb. Gen. News.* **18**, 194-200.
- Bond, J.E., and Gresshoff, P.M. (1993).** Soybean Transformation to study Molecular Physiology. In *Current Topics in Plant Molecular Biology*, P.M. Gresshoff, ed (Boca Raton, Florida: CRC Press), pp. 25-44.
- Bork, P., Ouzounis, C., and Sander, C. (1994).** From genome sequences to protein function. *Current Opinion in Structural Biology* **4**, 393-403.
- Boyes, D.C., and Nasrallah, J.B. (1993).** Physical linkage of the SLG and SRK genes at the self-incompatibility locus of *Brassica oleracea*. *Mol Gen Genet* **236**, 369-373.
- Brar, G.S., and Carter, T.E. (1993).** Soybean *Glycine max* (L.) Merrill. In *Genetic improvement of vegetable crops*, G. Kalloo, and B.O. Bergh, eds (NY: Permagon Press, Ltd), pp. 427-463.

- Brewin, N.J. (1993). The *Rhizobium*-legume symbiosis: plant morphogenesis in a nodule. seminars in CELL BIOLOGY 4, 149-156.
- Burke, D.T., Carle, G.F., and Olson, M.V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236, 806-812.
- Burmeister, M., and Ulanovsky, L. edB (1992). In Methods in Molecular Biology, Vol. 12: Pulsed-Field Gel Electrophoresis, (Totowa, NJ: The Humana Press Inc.).
- Caetano-Anollés, G., Joshi, P.A., and Gresshoff, P.M. (1990). Spontaneous nodules induce feedback suppression of nodulation in alfalfa. Planta 183, 77-82.
- Caetano-Anollés, G., and Gresshoff, P.M. (1991). Plant genetic control of nodulation. Annu. Rev. Microbiol. 45, 345-382.
- Caetano-Anollés, G., Bassam, B.J., and Gresshoff, P.M. (1991). DNA amplification fingerprinting using very short oligonucleotide primers. Bio/Technology 9, 553-557.
- Caetano-Anollés, G.C., Paparozzi, E.T., and Gresshoff, P.M. (1991). Mature nodules and root tips control nodulation in soybean. J. Plant Physiol. 137, 389-396.
- Calvert, H.E., Pence, M.K., Pierce, M., Malik, N.S.A., and Bauer, W.D. (1984). Anatomical analysis of the development and distribution of *Rhizobium* infections in soybean roots. Can. J. Bot. 62, 2375-2384.
- Carroll, B.J., McNeil, D.L., and Gresshoff, P.M. (1985). A supernodulation and nitrate tolerant symbiotic (*nts*) soybean mutant. Plant Physiol. 78, 34-40.
- Carroll, B.J., McNeil, D.L., and Gresshoff, P.M. (1985). Isolation and properties of soybean mutants which nodulate in the presence of high nitrate concentrations. Proc. Natl. Acad. Sci. USA 82, 4162-4166.
- Carroll, B.J., and Mathews, A. (1990). Nitrate inhibition of nodulation in legumes. In Molecular biology of symbiotic nitrogen fixation, P.M. Gresshoff, ed (Boca Raton, FL: CRC Press), pp. 159-180.
- Carter, N.P. (1994). Cytogenetic analysis by chromosome painting. Cytometry 18, 2-10.
- Chang, C., Bowman, J.L., DeJohn, A.W., Lander, E.S., and Meyerowitz, E.M. (1988). Restriction fragment length polymorphism map for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA 85, 6856-6860.
- Chang, C., Kwok, S.F., Bleecker, A.B., and Meyerowitz, E.M. (1993). *Arabidopsis* ethylene-response gene *ETR1*: similarity of product to two-component regulators. Science 262, 539-544.
- Chasan, R. (1993). Notes from underground. Plant Cell 5, 140-144.
- Chaubet, N., Clement, B., and Gigot, C. (1992). Genes encoding a Histone H3.3-like variant in *Arabidopsis* contain intervening sequences. J. Mol. Biol. 225, 569-574.
- Chen, L., Kosslak, R., and Atherly, A.G. (1994). Mechanical shear of high molecular weight DNA in agarose. Biotechniques 16, 228-229.
- Chen, Z., Silva, H., and Klessig, D.F. (1993). Active oxygen species in the induction of plant systemic acquired resistance by salicylic acid. Science 262, 1883-1886.

- Cheon, C.I., Lee, N.G., Siddique, A.B.M., Bal, A.K., and Verma, D.P.S. (1993). Roles of plant homologs of Rab1p and Rab7p in the biogenesis of the peribacteroid membrane, a subcellular compartment formed de novo during root nodule morphogenesis. *EMBO* 12, 4125-4135.
- Christou, P., McCabe, D.E., Swain, W.F., and Russell, D.R. (1993). Legume Transformation. In *Control of Plant Gene Expression*, D.P.S. Verma, ed (CRC Press, Inc.), pp. 547-564.
- Chu, G., Vollrath, D., and Davis, R.W. (1986). Separation of large DNA molecules by contour-clamped homogeneous electric fields. *Science* 234, 1582-1585.
- Civardi, L., Xia, Y., Edwards, K.J., Schnable, P.S., and Nikolau, B.J. (1994). The relationship between genetic and physical distances in the cloned *al-sh2* interval of the *Zea mays* L. genome. *Proc. Nat. Acad. Sci. USA* 91, 8268-8272.
- Coleman, M., Németh, A.H., Campbell, L., Raut, C.P., Weissenbach, J., and Davies, K.E. (1994). A 1.8 Mb YAC contig in Xp11.23: identification of CpG islands and physical mapping of CA repeats in a region of high gene density. *Genomics* 21, 337-343.
- Connelly, C., McCormick, M.K., Shero, J., and Hieter, P. (1991). Polyamines eliminate an extreme size bias against transformation of large yeast artificial chromosome DNA. *Genomics* 10, 10-16.
- Cooper, J.B., and Long, S. (1994). Morphogenetic rescue of *Rhizobium meliloti* nodulation mutants by *trans*-zeatin secretion. *Plant Cell* 6, 215-225.
- Creusot, F., Seignac, M., and Dron, M. (1992). Large-scale mapping using high-molecular-weight weight from *P. vulgaris*. *Plant Mol Biol Rep* 10, 207-218.
- Day, D.A., Carroll, B.J., Delves, A.C., and Gresshoff, P.M. (1989). The relationship between autoregulation and nitrate inhibition of nodulation in soybeans. *Physiol. Plant.* 75, 37-42.
- De Block, M. (1993). The cell biology of plant transformation: Current state, problems, prospects, and the implications for the plant breeder. *Euphytica* 71, 1-14.
- Deckert, J., Taranenko, N., and Gresshoff, P.M. (1994). Cell Cycle Genes and their Plant Homologues. In *Current Topics in Plant Molecular Biology*, P.M. Gresshoff, ed (Boca Raton, Florida: CRC Press, Inc.), pp. 169-193.
- Del-Favero, J., Vauterin, M., Weyens, G., Edwards, K.E., and Jacobs, M. (1994). Construction and characterization of a yeast artificial chromosome library containing five haploid sugarcane (*Beta vulgaris* L.) genome equivalents. *TAG* 88, 449-453.
- Delves, A.C., Mathews, A., Day, D.A., Carter, A.S., Carroll, B.J., and Gresshoff, P.M. (1986). Regulation of the soybean-*Rhizobium* symbiosis by shoot and root factors. *Plant Physiol.* 82, 588-590.
- Delves, A.C., Carroll, B.J., and Gresshoff, P.M. (1988). Genetic analysis and complementation studies on a number of mutant supermodulating soybean lines. *J. of Genetics* 67, 1-8.
- Delves, A.C., Higgins, A., and Gresshoff, P.M. (1992). Shoot apex removal does not alter autoregulation of nodulation in soybean. *Plant, Cell, and Environment* 15, 249-254.

- Demmer, L.A., and Chaplin, D.D.** (1993). Simultaneous transfer of four functional genes from the HLA class II region into mammalian cells by fusion with yeast spheroplasts carrying an artificial chromosome. *J. Immunol.* **150**, 5371.
- Denton, P.H., Cullen, J.B., Loeb, D., Lucas, A., and Nunes, K.** (1994). Partitioned PFGE-PCR: a new method for pulsed-field mapping of SRS and microsatellites. *Nucleic Acids Res.* **22**, 1776-1777.
- Doggett, N.A., Smith, C.L., and Cantor, C.R.** (1992). The effect of DNA concentration on mobility in pulsed field gel electrophoresis. *Nucleic Acids Res.* **20**, 859-864.
- Duc, G., and Messenger, A.** (1989). Mutagenesis of pea (*Pisum sativum* L.) and the isolation of mutants for nodulation and nitrogen fixation. *Plant Science* **60**, 207-213.
- Dunford, R., and Rogner, U.C.** (1991). Construction of yeast artificial chromosomes cotaining barley DNA and the identification of clones carrying copies of the repeated element BIS-1. *Hereditas* **115**, 133-138.
- Dunford, R., Vilagelio, L., and Moore, G.** (1993). Stabilization of a yeast artificial chromosome containing plant DNA using a recombination-deficient host. *Plant Mol Biol* **21**, 1187-1189.
- Edwards, K.J., Thompson, H., Edwards, D., de Saizieu, A., Sparks, C., Thompson, J.A., Greenland, A.J., Evers, M., and Schuch, W.** (1992). Construction and characterization of a yeast artificial chromosome library containing three haploid maize genome equivalents. *Plant Mol Biol* **19**, 299-308.
- Evers, M., Edwards, K., and Schuch, W.** (1992). Construction and characterization of a yeast artificial chromosome (YAC) library containing two haploid *Beta vulgaris* L. genome equivalents. *Gene* **121**, 195-202.
- Ferrin, L.J., and Camerini-Otero, D.** (1991). Selective cleavage of human DNA: RecA-assisted restriction endonuclease (RARE) cleavage. *Science* **254**, 1494-1497.
- Ferrin, L.J., and Camerini-Otero, D.** (1994). Long-range mapping of gaps and telomeres with Rec-assisted restriction endonuclease (RARE) cleavage. *Nature Genet.* **6**, 379-383.
- Flavell, R.** (1980). The Molecular Characterization and Organization of Plant Chromosomal Sequences. In *Annual Review of Plant Physiology*, W.R. Briggs, P.B. Green, and R.L. Jones, eds (Palo Alto, CA: Annual Reviews Inc.), pp. 569-596.
- Foote, S.** Construction of YAC libraries with large inserts. In *Current Protocols in Human Genetics*, E. Green, ed. In press.
- Foote, S., Vollrath, D., Hilton, A., and Page, D.C.** (1992). The Human Y Chromosome: DNA Clones Spanning the Euchromatic Region. *Science* **258**, 60-66.
- Funke, R.P., Kolchinsky, A., and Gresshoff, P.M.** (1993). Physical mapping in a region of the soybean genome containing duplicated sequences. *Plant Mol Biol* **22**, 437-446.
- Funke, R.P., Kolchinsky, A., and Gresshoff, P.M.** (1994). High EDTA concentrations cause entrapment of small DNA molecules in the compression zone of pulsed field gels, resulting in smaller than expected insert sizes in YACs prepared from size selected DNA. *Nucleic Acids Res.* **22**, 2708-2709.
- Gaensler, K.M.L., Kitamura, M., and Kan, Y.W.** (1993). Germ-line transmission and developmental regulation of a 150 kb yeast artificial chromosome containing the human beta-globin locus in transgenic mice. *Proc. Nat. Acad. Sci. USA* **90**, 11381-11385.

- Gamborg, O.L., Davis, B.P., and Stahlhut, R.W. (1983).** Cell division and differentiation in protoplasts from cell cultures of *Glycine* species and leaf tissue of soybean. *Plant Cell Reports* **2**, 213-215.
- Ganal, M.W., Young, N.D., and Tanksley, S.D. (1989).** Pulsed field gel electrophoresis and physical mapping of large DNA fragments in the *Tm-2a* region of chromosome 9 in tomato. *Mol. Gen. Genet.* **215**, 395-400.
- Gibson, S., and Somerville, C. (1993).** Isolating plant genes. *Trends in Biotechnology* **11**, 303-313.
- Gibson, S.I., and Somerville, C. (1992).** Chromosome walking in *Arabidopsis thaliana* using Yeast Artificial Chromosomes. In *Methods in Arabidopsis Research*, C. Koncz, N.H. Chua, and J. Schell, eds (Singapore: World Scientific), pp. 119-143.
- Gorman, S.W., Roberts-Oehlschlager, S.L., and Cullis, C.A. (1992).** Producing a P1 bacteriophage library from pine: Isolation and cloning of very high molecular weight DNA. *BioTechniques* **12**, 722.
- Gottlob-McHugh, S.G., Lévesque, M., MacKenzie, K., Olson, M., Yarosh, O., and Johnson, D.A. (1990).** Organization of the 5s rRNA genes in the soybean *Glycine max* (L.) Merrill and conservation of the 5s rDNA repeat structure in higher plants. *Genome* **33**, 486-494.
- Green, E.D., and Olson, M.V. (1990).** Systematic screening of yeast artificial-chromosome libraries using the polymerase chain reaction. *Proc. Nat. Acad. Sci. USA* **87**, 1213-1217.
- Gremaud, M.F., and Harper, J.E. (1989).** Selection and initial characterization of partially nitrate tolerant nodulation mutants of soybean. *Plant Physiol.* **89**, 169-173.
- Gresshoff, P.M., Krotzky, A., Mathews, A., Day, D.A., Schuller, K.A., Olsson, J., Delves, A.C., and Carroll, B.J. (1988).** Suppression of symbiotic supernodulation symptoms of soybean. *J. Plant Physiol.* **132**, 417-423.
- Gresshoff, P.M., and Caetano-Anollés, G. (1992).** Systemic regulation of nodulation in legumes. In *Current Topics in Plant Molecular Biology*, P.M. Gresshoff, ed (Boca Raton, FL: CRC Press), pp. 87-100.
- Gresshoff, P.M. (1993).** Molecular genetic analysis of nodulation genes in soybean. *Plant Breeding Reviews* **11**, 275-318.
- Gresshoff, P.M. (1993).** Plant function in nodulation and nitrogen fixation in legumes. In *New Horizons in Nitrogen Fixation*, R. Palacios, J. Mora, and W.E. Newton, eds (Norwell, MA: Kluwer Academic Publishers), pp. 31-42.
- Gresshoff, P.M., and Landau-Ellis, D. (1994).** Molecular Mapping of Soybean Nodulation Genes. In *Current Topics in Plant Molecular Biology*, P.M. Gresshoff, ed (Boca Raton, Florida: CRC Press), pp. 97-112.
- Grierson, D., and Covey, S.N. (1988).** *Plant Molecular Biology*, (New York: Chapman and Hall).
- Grill, E., and Somerville, C. (1991).** Construction and characterization of a yeast artificial chromosome library of *Arabidopsis* which is suitable for chromosome walking. *Mol. Gen. Genet.* **226**, 484-490.
- Gruenbaum, Y., Naveh-Man, T., Cedar, H., and Razi, A. (1981).** Sequence specificity of methylation in higher plant DNA. *Nature* **292**, 860-862.

- Gullis, C.A., Rivin, C.J., and Walbot, V. (1984). A rapid procedure for the determination of the copy number of repetitive sequences in eukaryotic genomes. *Plant Mol Biol Rep* 2, 24-31.
- Gurley, W.B., Hepburn, A.G., and Key, J.L. (1979). Sequence organization of the soybean genome. *Biochim. Biophys. Acta* 561, 167-183.
- Guzman, P., and Ecker, J.H. (1988). Development of large DNA methods for plants: molecular cloning of large segments of *Arabidopsis* and carrot DNA into yeast. *NAR* 16, 11091-11105.
- Hadley, H.H., and Hymowitz, T. (1973). In *Soybeans: Improvement, Production, and Uses*, B.E. Caldwell, ed (Madison, Wi: American Society of Agronomy), pp. 97-114.
- Handberg, K., and Stougaard, J. (1992). *Lotus japonicus*, an autogamous, diploid legume species for classical and molecular genetics. *The Plant Journal* 2, 487-496.
- Helentjaris, T., Weber, D., and Wright, S. (1988). Identification of the genomic locations of duplicate nucleotide sequences in maize by analysis of restriction fragment length polymorphisms. *Genetics* 118, 353-363.
- Herridge, D.F., and Betts, J.H. (1985). Nitrate tolerance in soybean: variation between genotypes. In *Nitrogen Fixation Research Progress*, H.J. Evans, P.J. Bottomly, and W.E. Newton, eds (Dordrecht: Martinus Nijhoff), p. 32.
- Hieter, P., Connelly, C., Shero, J., McCormick, M.K., Antonarakis, S., Pavan, W., and Reeves, R. (1990). Yeast Artificial Chromosomes: Promises Kept and Pending. In *Genome Analysis Volume I: Genetic and Physical Mapping*, (NY: Cold Spring Laboratory Press), pp. 83-119.
- Hinchee, M.A.W., Connor-Ward, D.V., Newell, C.A., McDonnell, R.E., Sato, S.J., Gasser, C.S., Fischhoff, D.A., Re, D.E., Fraley, R.T., Horsch, R.B. (1988). Production of transgenic soybean plants using *Agrobacterium*-mediated DNA transfer. *BioTechnology*, 915-922.
- Honeycutt, R.J. (1991). Physical mapping of *Rhizobium meliloti* and *Glycine max* L. (Merr.) using pulsed field gel electrophoresis. Ph. D. dissertation, Iowa State University, Ames, IA.
- Honeycutt, R.J., Sobral, B.W.S., McClelland, M., and Atherly, A.G. (1992). Analysis of large DNA from soybean (*Glycine max* L. Merr.) by pulsed-field gel electrophoresis. *The Plant Journal* 2, 133-135.
- Hwang, I., Kohchi, T., Hauge, B.M., Goodman, H.M., Schmidt, R., Cnops, G., Dean, C., Gibson, S., Iba, K., Lemieux, B., Arondel, V., Danhoff, L., and Somerville, C. (1991). Identification and map position of YAC clones comprising one-third of the *Arabidopsis* genome. *The Plant Journal* 1, 367-374.
- Hymowitz, T., and Singh, R.J. (1987). Taxonomy and Speciation. In *Soybeans: improvement, production, and uses*, J.R. Wilcox, ed (Madison, Wis.: ASA, CSSA, and SSSA), pp. 23-48.
- Keim, P., Shoemaker, R.C., and Palmer, R.G. (1989). Restriction fragment length polymorphism diversity in soybean. *Theor. Appl. Gen.* 77, 786-792.
- Kesseli, R., Witsenboer, H., Stanghellini, M., Vandermark, G., and Michelmore, R. (1993). Recessive Resistance to *Plasmopara lactucae-racidis* Maps by Bulk Segregant Analysis to a Cluster of Dominant Disease Resistance Genes in Lettuce. *MPMI* 6, 722-728.

- Ketner, G., Spencer, F., Tugendreich, S., Connelly, C., and Hieter, P. (1994). Efficient manipulation of the human adenovirus genome as an infectious yeast artificial chromosome clone. *Proc. Nat. Acad. Sci. USA* **91**, 6186-6190.
- Kleine, M., Michalek, W., Graner, A., Herrmann, R.G., and Jung, C. (1993). Construction of a barley (*Hordeum vulgare* L.) YAC library and isolation of a *Hor1*-specific clone. *Mol Gen Genet* **240**, 265-272.
- Kolchinsky, A., Funke, R., and Gresshoff, P.M. Dissecting molecular mechanisms of nodulation: taking a leaf from *Arabidopsis*. *Plant. Mol. Biol.* In press.
- Kolchinsky, A., Funke, R., and Gresshoff, P.M. (1993). DAF-amplified fragments can be used as markers for DNA from pulse field gels including yeast artificial chromosomes. *Biotechniques* **14**, 400-403.
- Kolchinsky, A., and Gresshoff, P.M. (1994). Plant Telomeres as Molecular Markers. In *Current Topics in Plant Molecular Biology*, P.M. Gresshoff, ed (Boca Raton, Florida: CRC Press), pp. 97-112.
- Koncz, C., Schell, J., and Rédei, G.P. (1992). T-DNA transformation and insertion mutagenesis. In *Methods in Arabidopsis Research*, C. Koncz, N.H. Chua, and J. Schell, eds (Singapore: World Scientific), pp. 119-143.
- Kosslak, R.M., and Bohlool, B.B. (1984). Suppression of nodule development of one side of a root system of soybeans caused by prior inoculation of the other side. *Plant Physiol.* **75**, 125-130.
- Kouchi, H., and Hata, S. (1993). Isolation and characterization of novel cDNAs representing genes expressed at early stages of nodule development. *Mol Gen Genet* **238**, 106-119.
- Kovács, L.G., and Pueppke, S.G. (1993). The chromosomal background of *Agrobacterium tumefaciens* Chry5 conditions high virulence on soybean. *Mol. Plant-Microbe Interact.* **6**, 601-608.
- Kurata, N., Moore, G., Nagamura, Y., Foote, T., Yano, M., Minobe, Y., and Gale, M. (1994). Conservation of Genome Structure between Rice and Wheat. *Bio/technology* **12**, 276-278.
- Lai, E., Birren, B.W., Clark, S.M., Simon, M.I., and Hood, L. (1989). Pulsed Field Gel Electrophoresis. *BioTechniques* **7**, 34-42.
- Lamb, B.T., Sisodia, S.S., and Lawler, A.M. (1993). Introduction and expression of the 400 kilobase precursor amyloid protein gene in transgenic mice. *Nature Genet.* **5**, 22-26.
- Landau-Ellis, D., Angermüller, S., Shoemaker, R., and Gresshoff, P.M. (1991). The genetic locus controlling supernodulation in soybean (*Glycine max* L.) co-segregates tightly with a cloned molecular marker. *Mol. Gen. Genet.* **228**, 221-226.
- Landau-Ellis, D., and Gresshoff, P.M. (1992). Supernodulating soybean mutant alleles, nts382 and nts1007, show no recombination with the same restriction fragment length polymorphism marker. *Mol. Plant-Microbe Interact.* **5**, 428-429.
- Landau-Ellis, D., and Gresshoff, P.M. (1994). The RFLP marker closely linked to the supernodulation locus of soybean contains three inserts. *Mol. Plant-Microbe Interact.* **7**, 432-433.
- Lapitan, N.L.V. (1992). Organization and evolution of higher plant nuclear genomes. *Genome* **35**, 171-181.

- Larin, Z., Monaco, A.P., and Lehrach, H. (1991). Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc. Natl. Acad. Sci. USA* **88**, 4123-4127.
- Lark, K.G., Weisemann, J.M., Matthews, B.F., Palmer, R., Chase, K., and Macalma, T. (1993). A genetic map of soybean (*Glycine max* L.) using an intraspecific cross of two cultivars: 'Minsoy' and 'Noir'. *Theor Appl Genet* **86**, 901-906.
- Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13**, 1095-1107.
- Lee, J.S., and Verma, D.P.S. (1984). Structure and chromosomal arrangement of leghemoglobin genes in kidney bean suggest divergence in soybean leghemoglobin gene loci following tetraploidization. *Euro. Mol. Biol. Org. J.* **3**, 2745-2752.
- Lee, N.G., Stein, B., Suzuki, H., and Verma, D.P.S. (1993). Expression of antisense nodulin-35 RNA in *Vigna aconitifolia* transgenic root nodules retards peroxisome development and affects nitrogen availability to the plant. *Plant J.* **3**, 599-606.
- Leung, J., Bouvier-Durand, M., Morris, P.-C., Guerrier, D., Cheddor, F., and Giraudat, J. (1994). *Arabidopsis* ABA response gene *ABI1* features of a calcium-modulated protein phosphatase. *Science* **264**, 1448-1451.
- Leyser, H.M.O., Lincoln, C.A., Timpste, C., Lammer, T., and Estelle, M. (1993). *Arabidopsis* auxin-resistance gene *AXR1* encodes a protein related to ubiquitin-activating enzyme E1. *Nature* **364**, 161-163.
- Liang, P., and Pardee, A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967-970.
- Lin, W., Odell, J.T., and Schreiner, R.M. (1987). Soybean protoplast culture and direct gene uptake and expression by cultured soybean protoplasts. *Plant Physiology* **84**, 856-861.
- Lisitsyn, N., Lisitsyn, N., and Wigler, M. (1993). Cloning the difference between two complex genomes. *Science* **259**, 946-951.
- Liu, Y.-G., and Whittier, R.F. (1994). Rapid preparation of megabase plant DNA from nuclei in agarose plugs and microbeads. *Nucleic Acids Res.* **22**, 2168-2169.
- Marchuk, D., and Collins, F.S. (1988). pYAC-RC, a yeast artificial chromosome vector for cloning DNA with infrequently cutting restriction endonucleases. *Nucleic Acids Res* **16**, 7743.
- Martin, G., Williams, J., and Tanksley, S.D. (1991). Rapid identification of markers linked to a *Pseudomonas* resistance gene by using random primers and near-isogenic lines. *Proc. Natl. Acad. Sci. USA* **88**, 2336-2340.
- Martin, G.B., Ganai, M.W., and Tanksley, S.D. (1992). Construction of a yeast artificial chromosome library of tomato and identification of cloned segments linked to two disease resistance loci. *Mol Gen Genet* **233**, 25-32.
- Martin, G.B., de Vicente, M.C., and Tanksley, S.D. (1993a). High-Resolution Linkage Analysis and Physical Characterization of the *Pto* Bacterial Resistance Locus in Tomato. *MPMI* **6**, 26-34.
- Martin, G.B., Brommonschenkel, S.H.S.H., Chungwongse, J., Frary, A., Ganai, M.W., Spivey, R., Wu, T., Earle, E.D., and Tanksley, S.D. (1993b). Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* **262**, 1432-1436.

- Matallana, E., Bell, C.J., Dunn, P.J., Lo, M., and Ecker, J.R.** (1992). Genetic and physical linkage of the *Arabidopsis* genome: Methods for anchoring Yeast Artificial Chromosomes. In *Methods in Arabidopsis Research*, C. Koncz, N.H. Chua, and J. Schell, eds (Singapore: World Scientific), pp. 144-169.
- Mathews, A., Carroll, B.J., and Gresshoff, P.M.** (1989). Development of *Bradyrhizobium* infections in supernodulating and non-nodulating mutants of soybean (*Glycine max*[L.] Merrill). *Protoplasma* **150**, 40-47.
- Mathews, A., Carroll, B.J., and Gresshoff, P.M.** (1990). The genetic interaction between non-nodulation and supernodulation in soybean: an example of developmental epistasis. *Theor. Appl. Genet.* **79**, 125-130.
- McClelland, M.** (1983). The frequency and distribution of methylatable DNA sequences in leguminous plant protein coding sequences. *J Mol Evol* **19**, 346-354.
- McClelland, M., Jones, R., Patel, Y., and Nelson, M.** (1987). Restriction endonucleases for pulsed field mapping of bacterial genomes. *Nucleic Acids Research* **15**, 5985-6005.
- McClelland, M., and Nelson, M.** (1988). The effect of site-specific DNA methylation on restriction endonucleases and DNA modification methyltransferases-a review. *Gene* **74**, 291-304.
- McCormick, M.K., Antonarakis, S.E., and Hieter, P.** (1990). YAC cloning of DNA embedded in an agarose matrix. *Genetic Analysis Techniques and Applications* **7**, 114-118.
- McCormick, M.K., Campbell, E., Deaven, L., and Moyzis, R.** (1993). Low-frequency chimeric yeast artificial chromosome libraries from flow-sorted human chromosomes 16 and 21. *Proc. Natl. Acad. Sci. USA* **90**, 1063-1067.
- Meyer, K., Leube, M.P., and Grill, E.** (1994). A protein phosphatase 2C involved in ABA signal transduction in *Arabidopsis thaliana*. *Science* **264**, 1452-1454.
- Meyerowitz, E.** (1987). *Arabidopsis thaliana*. *Ann. Rev. Genet.* **21**, 93-111.
- Miao, G.-H., and Verma, D.P.S.** (1993). Soybean nodulin-26 gene encoding a channel protein is expressed only in the infected cells of nodules and is regulated differently in roots of homologous and heterologous plants. *Plant Cell* **5**, 781-794.
- Michelmore, R.W., Paran, I., and Kesseli, R.V.** (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA* **88**, 9828-9832.
- Monaco, A.P., and Larin, Z.** (1994). YACs, BACs, PACs and MACs: artificial chromosomes as research tools. *TIBTECH* **12**, 280-286.
- Moore, G., Gale, M.D., Kurata, N., and Flavell, R.B.** (1993). Molecular analysis of small grain cereal genomes: current status and prospects. *Bio/technology* **11**, 584-589.
- Moore, G., Abbo, S., Cheung, W., Foote, T., Gale, M., Koebner, R., Leitch, A., Leitch, I., Money, T., Stancombe, P., Yano, M., and Flavell, R.** (1993). Key features of cereal genome organization as revealed by the use of cytosine methylation-sensitive restriction endonucleases. *Genomics* **15**, 472-482.
- Nam, H.-G., Giraudat, J., Boer, D.B., Moonan, F., Loos, W.D.B., Hauge, B.M., and Goodman, H.M.** (1989). Restriction fragment length polymorphism map of *Arabidopsis thaliana*. *Plant Cell* **1**, 699-705.

- Nap, J.-P., and Bisseling, T. (1988). Nodulin function and nodulin gene regulation in root nodule development. In *Molecular biology of symbiotic nitrogen fixation*, P.M. Gresshoff, ed (Boca Raton, FL: CRC Press), pp. 181-230.
- Negorov, D.G., Macina, R.A., Spais, C., Ruthig, L.A., Hu, X.-L., and Riethman, H.C. (1994). Physical mapping of the terminal 270 Kb from human chromosome 1q. *Genomics* 22, 569-578.
- Neil, D.L., Villasante, A., Fisher, R.B., Vetrie, D., Cox, B., and Tyler-Smith, C. (1990). Structural instability of human tandemly repeated DNA sequences cloned in yeast artificial chromosome vectors. *NAR* 18, 1421-1428.
- Newell, C.A., and Luu, H.T. (1985). Protoplast culture and plant regeneration in *Glycine canescens* F.J. Herm. *Plant Cell Tissue Organ Culture* 4, 145-149.
- Okagaki, R.J., Neuffer, M.G., and Wessler, S.R. (1991). A deletion common to two independently derived waxy mutations of maize. *Genetics* 128, 425-431.
- Olsson, J.E., Nakao, P., Bohlool, B.B., and Gresshoff, P.M. (1989). Lack of systemic suppression of nodulation in split root systems of supernodulating soybean (*Glycine max* [L.] Merr.) mutants. *Plant Physiol.* 90, 1347-1352.
- Paterson, A.H., and Wing, R.A. (1993). Genome mapping in plants. *Current Opinion in Biotechnology* 4, 142-147.
- Pearce, G., Strydom, D., Johnson, S., and Ryan, C.A. (1991). A polypeptide from tomato leaves induces wound-inducible proteinase inhibitor proteins. *Science* 253, 895.
- Pena, S.D.J., Barreto, G., Vago, A.R., De Marco, L., Reinach, F.C., Dias Neto, E., and Simpson, A.J.G. (1994). Sequence-specific 'gene signatures' can be obtained by PCR with single specific primers at low stringency. *Proc. Nat. Acad. Sci. USA* 91, 1946-1949.
- Peterson, K.R., Zitnik, G., and Huxley, C. (1993). Use of YACs for studying control of gene expression: correct regulation of the genes of a human beta-globin locus after transfer to a mouse erythroleukemia line. *Proc. Nat. Acad. Sci. USA* 90, 11207-11211.
- Pierce, M., and Bauer, W.D. (1983). A rapid regulatory response governing nodulation in soybean. *Plant Physiol.* 73, 286-290.
- Poehlman, J.M. (1979). *Breeding Field Crops*, (Westport, Ct: The Avi Publishing Company, Inc.).
- Potterill, J., Robson, F., Lee, K., and Coupland, G. (1993). Chromosome walking with YAC clones in *Arabidopsis*; isolation of 1700 kb of contiguous DNA on chromosome 5, including a 300 kb region containing the flowering-time *CO*. *MGG* 239, 145-157.
- Prabhu, R.R., and Gresshoff, P.M. Mendelian, maternal, and paternal inheritance of polymorphic markers generated by short single arbitrary oligonucleotide primers in soybean. *Plant Mol Biol* In press.
- Raskin, I. (1992). Salicylate, a new plant hormone. *Plant Physiol.* 99, 799-803.
- Rhodes, P.R., and Vodkin, L.O. (1988). Organization of the *Tgm* Family of Transposable Elements in Soybean. *Genetics* 120, 597-604.
- Riley, J., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J.C., and Markham, A.F. (1990). A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *NAR* 18, 2887-2890.

- Rosendahl, L., Vance, C.P., Miller, S., and Jacobsen, E. (1989). Nodule physiology of a supernodulating pea mutant. *Physiologia Plantarum* **77**, 606-612.
- Roth, L.E., Jeon, K., and Stacey, G. (1988). Homology in Endosymbiotic Systems: the term "Symbiosome". In *Molecular Genetics of Plant-Microbe Interactions*, R. Palacios, and D.P.S. Verma, eds (St. Paul, MN: The American Phytopathological Society), pp. 220-225.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). In *Molecular Cloning: A Laboratory Manual*, (New York: Cold Spring Harbor Laboratory Press).
- Sanjuan, J., Carlson, R.W., Spaink, H.P., Bhat, U.R., Barbour, W.M., Glushka, J., and Stacey, G. (1992). A 2-*O*-methylfucose moiety is present in the lipo-oligosaccharide nodulation signal of *Bradyrhizobium japonicum*. *Proc. Natl. Acad. Sci. USA* **89**, 8789-8793.
- Sayavedra-Soto, L.A., Angermüller, S.A., Prabhu, R., and Gresshoff, P.M. Polypeptide patterns in leaves of *Glycine max* L. Merr. cv. Bragg and its supernodulating mutant during nodulation. *Physiology and Molecular Biology of Plants* In press.
- Schmidt, R., and Dean, C. (1993). Towards Construction of an Overlapping YAC library of the *Arabidopsis thaliana* genome. *BioEssays* **15**, 63-69.
- Schubert, K.R. (1986). Products of biological nitrogen fixation in higher plants: Synthesis, Transport, and Metabolism. *Ann. Rev. Plant. Physiol.* **37**, 539-574.
- Schwartz, D.C., Saffran, W., Welsh, J., Haas, R., Goldberg, M., and Cantor, C. (1982). New techniques for purifying large DNAs and studying their properties and packaging. *Cold Spring Harbor Symp. Quant. Biol.* **47**, 189-195.
- Schwartz, D.C., and Cantor, C.R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67-75.
- Segal, G., Sarfatti, M., Schaffer, M.A., Ori, N., Zami, D., and Fluhr, R. (1992). Correlation of genetic and physical structure in the region surrounding the *I₂* *Fusarium oxysporum* resistance locus in tomato. *Mol. Gen. Genet.* **231**, 179-185.
- Shero, J.H., McCormick, M.K., Antonarakis, S.E., and Hieter, P. (1991). Yeast artificial chromosome vectors for efficient clone manipulation and mapping. *Genomics* **10**, 505-508.
- Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* **89**, 8794-8797.
- Shoemaker, R.C., Guffy, R.D., Lorenzen, L.L., and Specht, J.E. (1992). Molecular genetic mapping of soybean: map utilization. *Crop Sci.* **2**, 1091-1098.
- Singh, R.J., and Hymowitz, T. (1988). The genomic relationship between *Glycine max* (L.) Merr. and *Glycine soja* (Sieb. and Zucc.) as revealed by pachytene chromosome analysis. *Theor. Appl. Genet.* **76**, 705-711.
- Southard-Smith, M., Pierce, J.C., and MacDonald, R.J. (1994). Physical mapping of the rat tissue kallikrein family in two gene clusters by analysis of P1 bacteriophage clones. *Genomics* **22**, 404-417.
- Staswick, P.E. (1992). Jasmonate, Genes, and Fragrant Signals. *Plant Physiol.* **99**, 804-807.

- Sternberg, N. (1990). Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc. Natl. Acad. Sci. USA* **87**, 103-107.
- Stougaard, J., Petersen, T.E., and Marcker, K.A. (1987). Expression of a complete soybean leghemoglobin gene in root nodules of transgenic *Lotus corniculatus*. *Proc. Nat. Acad. Sci. USA* **84**, 5754-5757.
- Stougaard, J., Sandal, N.N., Gron, A., Kuhle, A., and Marcker, K.A. (1987). 5' Analysis of the soybean leghaemoglobin Lbc3 gene: regulatory elements required for promoter activity and organ specificity. *EMBO J.* **6**, 3565-3569.
- Streeter, J. (1988). Inhibition of legume nodule formation and N₂ fixation by nitrate. *CRC Critical Reviews in Plant Sciences* **7**, 1-23.
- Tanksley, S.D., Ganai, M.W., Prince, J.P., De Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovanonni, J.J., Grandillo, S., Martin, G.B., Messeguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Roder, M.S., Wing, R.A., Wu, W., and Young, N.D. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132**, 1141-1160.
- Tanner, J.W., and Anderson, I.C. (1963). An external effect of inorganic nitrogen in root nodulation. *Nature* **198**, 303-304.
- Truchet, G., Barker, D.G., Camut, S., de Billy, F., Vasse, J., and Huguet, T. (1989). Alfalfa nodulation in the absence of *Rhizobium*. *Mol Gen Genet* **219**, 65-68.
- Udvardi, M.K., Price, G.D., Gresshoff, P.M., and Day, D.A. (1988). A dicarboxylate transporter on the peribacteroid membrane of soybean nodules. *FEBS Letters* (1), 36-40.
- van Wordragen, M., Bonnema, G., Verkerk, R., Hontelez, J., Peytavi, R., Grimsley, N., and Zabel, P. (1994). YAC-Transfer to Plant Cells. 4th International Congress of Plant Molecular Biology June.
- Verdoes, J.C., Calil, M.R., Punt, P.J., Debets, F., Swart, K., Stouthamer, A.H., and van den Hondel, C.A.M.J.J. (1994). The complete karyotype of *Aspergillus niger*: the use of induced electrophoretic mobility variation of chromosomes for gene assignment studies. *Mol Gen Genet* **244**, 75-80.
- Verma, D.P.S., Hu, C.A., and Zhang, M. (1992). Root nodule development: origin, function, and regulation of nodulin genes. *Physiol. Plant.* **85**, 253-265.
- Vincent, J.M. (1980). Factors controlling the legume-*Rhizobium* symbiosis. In *Nitrogen Fixation*, W.E. Newton, and W.H. Orme-Johnson, eds (Baltimore: Univ. Park Press), pp. 103-129.
- Wada, M., Abe, K., Okamura, K., Taguchi, H., Kohno, K., Imamoto, F., Schlessinger, D., and Kuwano, M. (1994). Chimeric YACs were generated at unreduced rates in conditions that suppress coligation. *Nucleic Acids Res.* **22**, 1651-1654.
- Walsh, K.B., and Carroll, B.J. (1991). A reappraisal of nitrate inhibition of nitrogenase in A317, a nitrate reductase-deficient mutant of pea (*Pisum sativum*). *Physiol. Plant.* **84**, 283-291.
- Walter, M.A., Surti, U., Hofker, M.H., and Cox, D.W. (1990). The physical organization of the human immunoglobulin heavy chain gene locus. *EMBO J.* **9**, 3303-3313.
- Warburton, P.E., and Willard, H.F. (1990). Genomic analysis of sequence variation in tandemly repeated DNA in the human genome. *J. Mol. Biol.* **216**, 3-16.

- Ward, E.R., and Jen, G.C.** (1990). Isolation of single-copy-sequence clones from a yeast artificial chromosome library of randomly-sheared *Arabidopsis thaliana* DNA. *Plant Mol Biol* 14, 561-568.
- Wicking, C., and Williamson, B.** (1991). From linked marker to gene. *Trends in Genet.* 7, 288-293.
- Wing, R.A., Rastogi, V.K., Zhang, H.B., Paterson, A.H., and Tanksley, S.D.** (1993). An improved method of plant DNA megabase isolation in agarose microbeads which is suitable for physical mapping and YAC cloning. *Plant J.* 4, 893-898.
- Wu, Z.-W., Bogre, L., Vincze, É., Kiss, G.B., and Dudits, D.** (1988). Isolation of an alfalfa histone H3 gene: structure and expression. *Plant Mol Biol* 11, 641-649.
- Xu, S.J., Singh, R.J., and Hymowitz, T.** (1994). Research on Primary Trisomics in Soybean. *Proceedings of the 5th Biennial Conference on the Molecular and Cellular Biology of the Soybean.*
- Yalpani, N., Silverman, P., Wilson, T.M.A., Kleier, D.A., and Raskin, I.** (1991). Salicylic acid is a systemic signal and an inducer of pathogenesis-related proteins in virus-infected tobacco. *Plant Cell* 3, 809-818.
- Young, N., Menancio-Hautea, D., Fatokun, C., Komar, L., Danesh, D., and Shoemaker, R.** (1992). Parallel RFLP mapping in soybean (*Glycine max*), mungbean (*Vigna radiata*) and cowpea (*V. unguiculata*). *IV Biennial Conference on Molecular and Cellular Biology of Soybean* 31.
- Young, N.D., Zamir, D., Ganai, M.W., and Tanksley, S.D.** (1988). Use of isogenic lines and simultaneous probing to identify DNA markers tightly linked to the *Tm-2a* gene in tomato. *Genetics* 120, 579-585.

APPENDICES

APPENDIX 1

GLOSSARY AND LIST OF ABBREVIATIONS

BSA- Bulked Segregant Analysis. A method for identifying molecular markers linked to a trait that utilizes pooled individuals from a population that is segregating for that trait (Michelmore *et al.*, 1991).

CHEF- Contour-clamped Homogeneous Electrical Field. Refers to a geometrical arrangement of independently controlled electrodes that creates homogeneous electrical fields during PFGE. It results in straight lanes, and is used in the popular instrument manufactured by BioRad Laboratories.

cM- centiMorgan. The unit of crossing over, equal to 1% crossing over.

CZ- Compression Zone. Zone on pulsed field gels a short distance below the well where DNA molecules larger than a specific size (as determined by the switch interval) are focused in a narrow band.

DAF- DNA Amplification Fingerprinting. A method for producing diagnostic PCR amplification patterns for single DNA molecules, individuals, cultivars or different species using short single primers (Caetano-Anollés *et al.*, 1991).

Kb- Kilo base. One thousand bases or base pairs in single-standed or double-stranded nucleic acids.

Mb- Mega base. A unit of length equal to one million bases or base pairs in single-standed or double stranded nucleic acids. One thousand kilo bases.

NILs-Near Isogenic Lines. Genetic lines that have been backcrossed to the same parent repeatedly until they differ only for a particular trait (e.g., disease resistance), and are presumed to differ only at the chromosomal loci controlling that trait.

PFGE- Pulsed Field Gel Electrophoresis. A method for electrophoretically separating very large DNA molecules based on their different size-dependent relaxation times as they migrate under the influence of an alternating electrical field (see Chapter 5 and Appendix 5).

PMSF- Phenyl Methyl Sulfonyl Fluoride. An inhibitor of proteinase K used to prepare agarose DNA blocks for restriction digestion.

RFLP-Restriction Fragment Length Polymorphism. DNA polymorphism between two species or cultivars caused by a different distribution of restriction sites around

homologous sequences in their genomes. Different-sized restriction fragments from each species or cultivar carrying the sequence are detected by autoradiography.

RIL- Recombinant Inbred Line. Line derived from one cross through single-seed descent over many generations. RILs are practically homozygous at most loci, and are used for the efficient creation of genetic and molecular maps (Lark *et al.*, 1993).

SCAR-Sequence Characterized Amplified Region. Small sequenced segment of the genome for which specific primers are available that permit PCR amplification of that region.

YAC- Yeast Artificial Chromosome. A large segment of DNA adapted with all the requisite elements, including a centromere and telomeres, for propagation as an extra chromosome in yeast (see Chapters 7 and 8).

APPENDIX 2

DISSECTING MOLECULAR MECHANISMS OF NODULATION: TAKING A LEAF FROM *ARABIDOPSIS**

The model plant *Arabidopsis* is providing increasing numbers of biologists with exciting opportunities to study aspects of plant metabolism and differentiation. However, *Arabidopsis* cannot be used to tackle one of the most complicated and important differentiation processes: nodulation. Although legume research has progressed, the molecular mechanisms triggering and controlling the differentiation of nitrogen-fixing nodules remain rather obscure.

However, mechanistic aspects on the bacterial side of this process have been significantly unveiled in recent years. Attracted by flavonoids secreted by the plant roots, bacteria release richly decorated oligosaccharides that are able to induce the nodulation response, beginning with root hair curling and cortical cell divisions. While the synthesis of the mitogenic factor is relatively clear, the molecular details of the signal-response pathways that lead from the first nodulation signal to a differentiated nodule remain obscure. Current ideas revolve around key regulatory elements of the cell cycle. With this in view, several researchers have investigated the expression of cell-cycle proteins, such as p34 protein kinase (*cdc2*), GTP-binding proteins, and proliferating cell nuclear antigen (Cheon *et al.*, 1993; Deckert *et al.*, 1994). Although good correlation has been found between *cdc2*, cyclin expression, and nodule emergence, one would not expect it to be a specific differentiation signal.

There is good evidence to suggest that early events are determined by interplay of plant hormones. First, early nodulation can be induced by so-called auxin transport inhibitors (ATI) in the absence of any other signals (Verma, 1992). Second, hormones can supplement mutant non-nodulating rhizobacteria and enable them to induce nodules (Cooper and Long, 1994). Furthermore, some early nodule-specific gene expression such as for the hydroxy-proline-rich cell wall protein ENOD2 are induced by cytokinin.

The legume host has a major role in regulating the overall process of nodulation and nitrogen fixation (Caetano-Anollés and Gresshoff, 1991). In both alfalfa and white clover, plants were discovered which formed nodules in the absence of *Rhizobium*. This demonstrated that the entire genetic complement needed for nodule development is plant-based. The plant determines the nodule morphology and contributes mechanisms for carbon and nitrogen assimilation as well as oxygen metabolism (Gresshoff, 1993).

*In Press. Kolchinsky, A., Funke, R.P., and Gresshoff, P.M. *Plant Mol Biol.*

Systemic feedback of nodule initiation, called autoregulation, emanates from cell division foci in the root (Caetano-Anollés and Gresshoff, 1991; Gresshoff, 1993). A further regulatory mechanism restricts the number of nodules in the presence of nitrate. Interestingly, inhibition of nodulation and nitrogen fixation by nitrate differs, as demonstrated by genetic analysis.

As an approach to understanding molecular events in the legume host, many genes with specific or significantly increased expression in nodules have been isolated (for review, see Nap and Bisseling, 1988, and Verma, 1992). A number of late nodulins, such as uricase and glutamine synthase, are nodule-specific duplicates of genes with known metabolic functions in other plant organs. The temporal and spatial patterns of expression of early nodulins suggested potential functions. The 'earliest' expressed nodulin, GmN#36, is detected 72h after inoculation but is transcribed in stem tissue as well (Kouchi and Hata, 1993). In contrast, cell-cycle-related genes, such as those for cyclin and p34 kinase, show increased mRNA 24h after inoculation Deckert *et al.*, 1994).

The expression of some nodulins has been studied in heterologous transgenic plants, including the legume *Lotus corniculatus* and the nonlegume tobacco (e.g., Miao and Verma, 1993; Stougaard *et al.*, 1987a,b). *L. corniculatus* has been used as a heterologous host to study nodulins isolated from soybean and other legumes, because it permits fast and reliable production of numerous transgenic nodules. However, transgenic homologous plants that have been recalcitrant to molecular transfection are starting to be used (e.g. Miao and Verma, 1993). Chimeric gene fusions and deletion analysis have identified both conserved and non conserved potential *cis*-acting elements responsible for nodule-specific expression and general transcriptional control of different nodulins. However, no specific *trans*-acting factors have been identified. The isolation and characterization of increasingly early nodulins by novel methods, such as differential display, may elucidate the relationship between bacterial nodulation factors and the transcriptional activation of early nodulin genes. Knowledge of the regulation of specific genes at the point where nitrate inhibition and feedback suppression of nodule formation are active should shed light on the mechanism of autoregulation.

However, almost all currently known nodulins represent genes that are expressed in already committed cells, necessary for morphological development of symbiotic tissue and nitrogen fixation. Genes that trigger or regulate nodule development could escape detection in differential screens, or may be expressed in another part of the plant.

One approach to the cloning of regulatory genes is based on mutations of these genes by transposons or other elements capable of integrating in the genome (gene tagging). Vast collections of such mutations have been generated in *Arabidopsis* by T-DNA insertional mutagenesis and by transferred maize transposons. Recently, a model legume, *Lotus japonicus*, has been proposed by Handberg and Stougaard as an 'Arabidopsis equivalent' plant model in nodulation research (Handberg and Stougaard, 1992). This plant has many advantages for molecular studies, among them small genome size, relatively easy transformation, and short generation time. One goal of Stougaard's group is to establish a transposon tagging system in this convenient leguminous plant. However, even with all the advantages of this model legume, thousands of transformants would have to be screened to find mutations relevant to nodulation. An additional complication of this otherwise promising project is scarce data on the biology and genetics of nodulation in *L. japonicus* as compared with other legumes.

A second approach that has been used successfully over the past 2-3 years is positional cloning of genes known by phenotype. The idea of positional cloning is based on the availability of detailed genetic maps consisting of hundreds of molecular markers, on the one hand, and libraries of large DNA fragments cloned in YACs or other suitable vectors, on the other. When molecular markers tightly linked to the gene of interest are found, a contig of overlapping clones has to be built. The search for the gene is based on physical mapping of rare recombinants and detection of expressed sequences in the contig. Eventual confirmation comes with transformation of the mutant with the putative gene to restore the wild-type phenotype.

Recent examples show that the cloning of potential regulatory genes leads to new proteins of utmost importance. Thus, cloning of the *ETR1* gene of *Arabidopsis* (*ETR1* mutants are not sensitive to ethylene) produced an element of a novel two component regulatory mechanism in plants (Chang *et al.*, 1993). In another work, the tomato gene for resistance to *Pseudomonas* infection turned out to be a protein kinase (Martin *et al.*, 1993).

We consider soybean (*Glycine max* L.) as an alternative model legume. A rich collection of nodulation-related mutants has been generated in soybean (Caetano-Anollés and Gresshoff, 1991; Gresshoff, 1993). The universal nature of these mutations is supported by the existence of similar mutations in other legumes, primarily in pea. First, there exist many different non-nodulating mutants, some apparently insensitive to the nodulation signal. These map to several loci and differ in their ability to block nodulation at different stages or to develop nodules with some, but not other, strains, of *Bradyrhizobium*. Second, there are

the so-called supernodulating mutants that develop many times more nodules than wild-type plants. They were initially described as mutants capable of nodulating in the presence of nitrate, which inhibits nodulation on wild-type plants. Several known mutations all map to a single locus.

Several detailed molecular maps of soybean were constructed, and, in contrast to pea, its genome has a reasonable size for cloning in YACs (1000 Mb) per haploid genome. We have begun the construction of a YAC library of soybean, and currently have about 10% of a haploid genome stored in clones ranging in size from 50 to 900 kb, with an average size of 160 kb. The clones are stable to repeated amplification and passage through yeast, and analysis of ends of YACs containing high and low proportions of repetitive DNA showed that about half of them represented unique sequences that might be useful for the identification of overlapping clones. About 7% of the clones represent chloroplast DNA.

Promising protocols for transformation of soybean roots with *Agrobacterium rhizogenes* (Cheon *et al.*, 1993; Lee *et al.*, 1993; Miao and Verma, 1993) and the discovery of *A. tumefaciens* strains that can infect soybeans (Kovács and Pueppke, 1993) have been reported recently. There is hope that within 1-2 years soybean transformation will be suitable for the studies of presumptive nodulation genes in transgenic plants.

In conclusion, two systems seem to be the most attractive for the identification of host regulatory genes involved in nodulation. The first is positional cloning of nodulation-related genes identified by mutations from a YAC library of soybean; and the second is transposon tagging of *L. japonicus* genes related to nodulation. The success of the model plant approach using *Arabidopsis* inspires new developments in nodulation research.

APPENDIX 3

METHODS FOR YAC LIBRARY CONSTRUCTION OF SOYBEAN

In summary, a partial *EcoRI* digest of soybean DNA is size selected by PFGE to remove molecules smaller than the desired insert size and cloned into the *EcoRI* site of the vector YAC4. The ligation mixture is used to transform spheroplasts of the yeast host strain AB1380.

A. Partial digestion with *EcoRI*, size selection by PFGE, and ligation to YAC4 vector

1. Solutions

- 10x ligation buffer

660 mM Tris.HCl, pH 7.5 (6.6 ml 1 M Tris.HCl per 10 ml)

50 mM MgCl₂ (0.5 ml 1 M MgCl₂)

300 mM NaCl (0.6 ml 5 M NaCl)

10x polyamines (1 ml of 100x polyamines stock)

- 100x polyamines stock

0.75 mM spermidine trihydrochloride

0.3 mM spermine tetrahydrochloride

dissolve 0.19 g of spermidine and 0.1 g of spermine in 10 ml of sterile water. These stocks should be kept at -20 °C.

2. Preparation of high molecular weight DNA in plugs

Protoplasts are prepared from young leaves and embedded in LMP agarose to produce DNA plugs, as described (Funke *et al.*, 1993; Chapter 5 of this thesis).

The plugs should be handled in sterile conditions with glass cover slips that have been washed with 70% alcohol.

The quality and intactness of the DNA in the plugs is variable, probably because of variation in the state of the starting fresh leaf material. When subjected to PFGE, a large

proportion of the DNA remains in the plug under conditions that resolve yeast chromosomes. Some of the DNA is focused in the compression zone (> 2Mb), with more severely degraded material being smeared along the lane. Sometimes there is a bimodal distribution, with very high molecular weight DNA being concentrated in the well and the compression zone, and a second peak in the range of 50 -150 kb.

For cloning purposes, it is very important to start with high molecular weight DNA of uniform quality. Uniformity makes it easier to titrate the *EcoRI* partial digestion conditions for each new batch of plugs. This is because the number of clonable molecules of a certain size produced after partial digestion is determined by the average size of the starting material. This average size should be much larger than the expected insert size.

3. *Pre-electrophoresis to remove degraded DNA*

Because the high EDTA salt concentration in stored plugs affects the mobility of DNA and causes entrapment of small molecules (Funke *et al.*, 1994), DNA plugs are washed before pre-electrophoresis. The plugs are rinsed for 15 minutes with sterile TE buffer (10 mM Tris.HCl, pH 7.5, 1 mM EDTA). Then they are rinsed two times for 5 minutes with TE buffer containing 1 mM PMSF (prepared fresh as a 100 mM stock in acetone or isopropanol) at 50°C. The plugs are rinsed one more time in TE buffer.

The plugs are loaded in a long well on a 1% agarose gel cast in 0.5x TBE. They are subjected to PFGE at 6 V/cm (200 V), 60 s pulse time, for 4 hours at 14°C in 0.5x TBE (as in Edwards *et al.*, 1992).

Running buffers are autoclaved and the electrophoresis chamber is washed with 70% ethanol before use.

Figure A3.1 shows the effect of pre-electrophoresis on the contents of the plug. Degraded DNA is removed, and what remains is a suitable substrate for partial digestion with *EcoRI*.

4. *Partial digestion of DNA*

Plugs containing high molecular weight DNA are retrieved from the gel.

The plugs are rinsed two times for 15 minutes with TE to dialyze the running buffer.

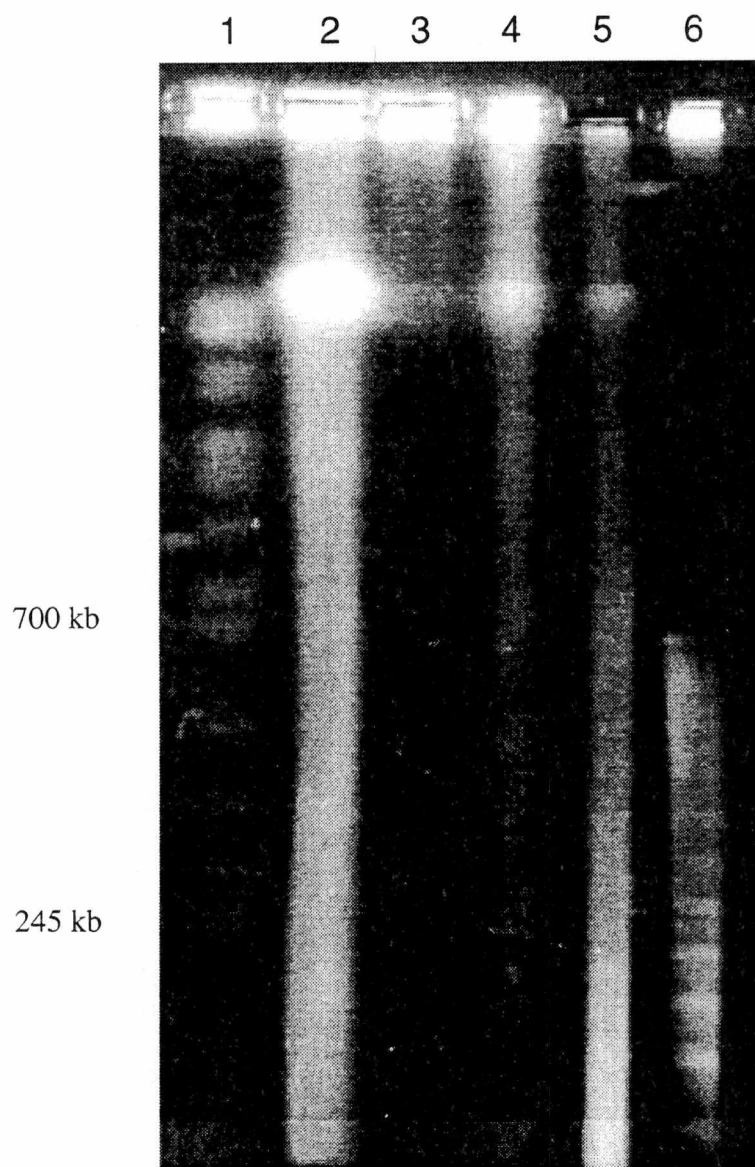


Figure A3.1. Pre-electrophoresis and partial digestion of soybean agarose DNA plugs. Lane 1, yeast chromosome size standard; lane 2, original plug of soybean DNA; lane 3, plug after removal of degraded DNA by pre-electrophoresis; lane 4, plug incubated with *Eco*RI digestion buffer at 37°C for 1 hour as mock digestion; lane 5, plug incubated with 0.2 unit *Eco*RI, 1 hour; lane 6, λ ladder size standard.

The plug is placed in an Eppendorf tube and weighed. Treating the plug as void volume, the contents of the tube are made up to 1x *Eco*RI buffer with sterile water and a 10x concentrate from the supplier (NEB). The tube is left on ice for 30 minutes to equilibrate.

The plug is cut into quarters and each piece of 40 μ l is placed in a separate tube containing 50 μ l restriction enzyme buffer. To titrate the conditions of partial digestion, 0.05, 0.2, 2, and 10 units of *Eco*RI are added to each piece using dilutions of the 20 unit/ml glycerol stock from the supplier (NEB).

The tubes are left on ice for 30 minutes to allow the enzyme to diffuse into the plugs before it reaches full activity.

Partial digestion is carried out at 37°C for one hour.

Digestions are quenched by adding 100 μ l of 100 mM EDTA containing 1 mg/ml proteinase K to each tube. Incubation at 37°C is continued for a further 30 minutes before returning the tubes to ice.

5. Titration of partial digestion by PFGE

The plugs are loaded on a 1% agarose mini gel (gel volume = 40 ml) and sealed in the wells with some residual molten agarose. Appropriate size standards (yeast and λ concatemers) are loaded on either side of the plugs.

To check the efficiency of partial digestion, the gel is usually run for 16 hours at 200 V with a switch interval ramped from 20-70 seconds.

The gel is stained with ethidium bromide and destained thoroughly by rinsing in running water for one hour.

With increasing units of enzyme, DNA is increasingly shifted down from the well and compression zone, and the average size of the fragments produced is smaller. With 20-40 units of enzyme distinct bands begin to be seen, showing that complete digestion has been achieved. Some material usually remains in the compression zone, and characteristic high molecular weight bands indicate that *Eco*RI sites are not randomly distributed in the genome (see Figure A3.1; Figure 8.2).

6. PFGE size selection of partial digestion products

The point at which digestion is just evident (usually about 0.2 units/40 µl plug) is used for a scaled up partial digestion . Scaling up is done simply by setting a larger number of digests as described above with the chosen number of units of enzyme.

Plugs containing partially digested DNA are loaded in a long slot on a 1% LMP gel. The gel volume is about 40 ml, and the plugs are sealed in the slot using some residual molten agarose. Size standards may be included on the side.

The running buffer for size selection is 0.5x TBE.

The efficiency of size selection is variable. That is, usually some DNA molecules smaller than the apparent cut-off size are entrapped or retained in the compression zone. These small DNA molecules preferentially transform yeast. In the construction of plant YAC libraries, this effect is usually overcome by employing switching regimes that are much longer than apparently necessary. Thus a switching regime of 30 s, which apparently focuses molecules larger than about 400 kb, resulted in clones with average inserts about half this size (eg. Martin *et al.*, 1992). Alternatively, a second PFGE size selection after ligation can be used to bring the median insert size closer to that expected from the conditions of size selection (Albertsen *et al.*, 1990). This has the disadvantage that DNA recovery from the compression zone is not quantitative, and some dilution takes place. In experiments with soybean DNA, some transformations using a 3 s pulse time resulted in median insert sizes of exactly 160 kb, as expected, while in others many clones with much smaller insert sizes were produced. If a single size selection gel is to be run, the conditions used are a compromise between the desired insert size and the desired number of clones, since more stringent selection also results in loss of material. This should be determined empirically.

The gel is run at 200 V for 12-16 hours, at the desired switch time (between 3 and 30 s).

While the gel is still in the chamber, strips which include the edges of the loading slot are cut from each side. The gel should be handled with gloves only, and glass cover slips washed with 70% ethanol are used for cutting. The strips are stained with ethidium bromide and viewed on a UV light table. The edges of the compression zone where large partial digestion products have been focused are marked with notches.

The strips are returned to the chamber and aligned with the central part of the gel. The

compression zone is cut out of the gel using the notches in the strips as a guide.

Figure A3.2 shows an example of a preparative gel from which the compression zone has been removed.

7. Preparation of vector DNA

The vector pYAC4 (described by Burke *et al.*, 1987) has an *EcoRI* cloning site. The telomere sequences of the vector are subject to rearrangement in some strains of *E. coli*. Upon digestion with *HindIII*, four bands should be produced: 3.5, 3.0, 1.9, and 1.5 kb (the band at 1.5 kb is a doublet containing the telomere sequences of each chromosome arm). Once the integrity of the telomeres in a particular plasmid preparation has been checked, the plasmid should be maintained in a *recA* strain such as DH5 α or HB101.

The plasmid preparation should be free of contaminating bacterial chromosomal DNA to avoid cloning artefacts. It should be purified over two CsCl gradients.

The plasmid is cut to completion with *Bam*HI in NEB buffer #3. 40 μ g may be digested in a volume of 400 μ l. The completeness of *Bam*HI digestion should be checked by electrophoresis. Bands of 1.7 and 9.4 kb should be seen. *Eco*RI is added to the reaction mixture (buffer #3 is compatible with both of these restriction enzymes). Complete digestion with *Bam*HI and *Eco*RI leaves a 1.7 kb HIS3 fragment, a centric chromosome arm (6.0 kb), and an acentric arm (3.7 kb).

Phosphatase (shrimp alkaline phosphatase, or SAP, supplier USB) is added at 0.1 unit/mg plasmid. After incubation at 37°C for one hour, the phosphatase is heat inactivated at 68°C for 20 minutes. One μ l of the plasmid preparation may be checked for the ability of the ends to be ligated in the presence and absence of kinase. The ligations are done in a volume of 10 μ l for one hour at room temperature or 16°C, and the entire reaction can be loaded on an agarose gel. An example of a good plasmid preparation with appropriate controls is shown in Figure A3.3. The vector is extracted sequentially with phenol-chloroform and chloroform. One half volume of 6 M ammonium acetate and two and a half volumes of ethanol are added to precipitate the DNA. To obtain quantitative yields, precipitation is accelerated by immersing the tube in a small amount of liquid nitrogen. After centrifugation for 15 minutes in the cold, the pellet is rinsed with 70% ethanol, and the vector is redissolved to a concentration of 0.5 -1.0 mg /ml.

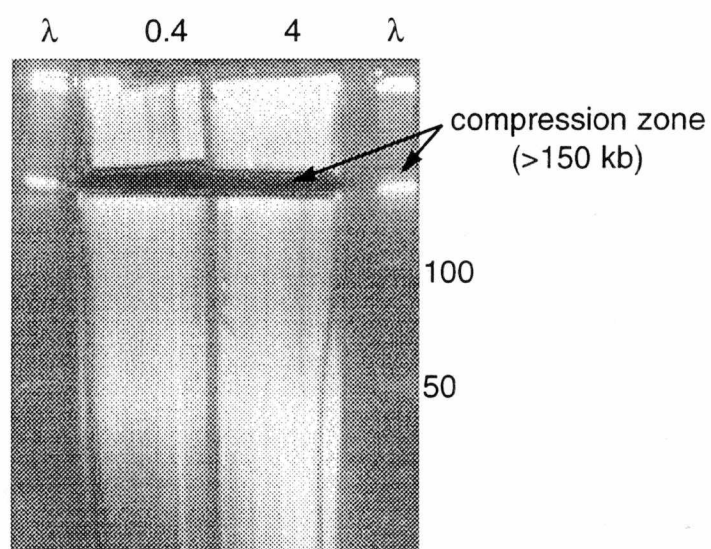


Figure A3.2. Preparative gel from which the compression zone, containing partial digestion products, has been removed. Running conditions were a 3 second switch interval, 0.25xTBE running buffer, 0.9% LMP agarose, 200V, for 12 hours. The number of units of *Eco*RI used per 40 μ l DNA plug is indicated.

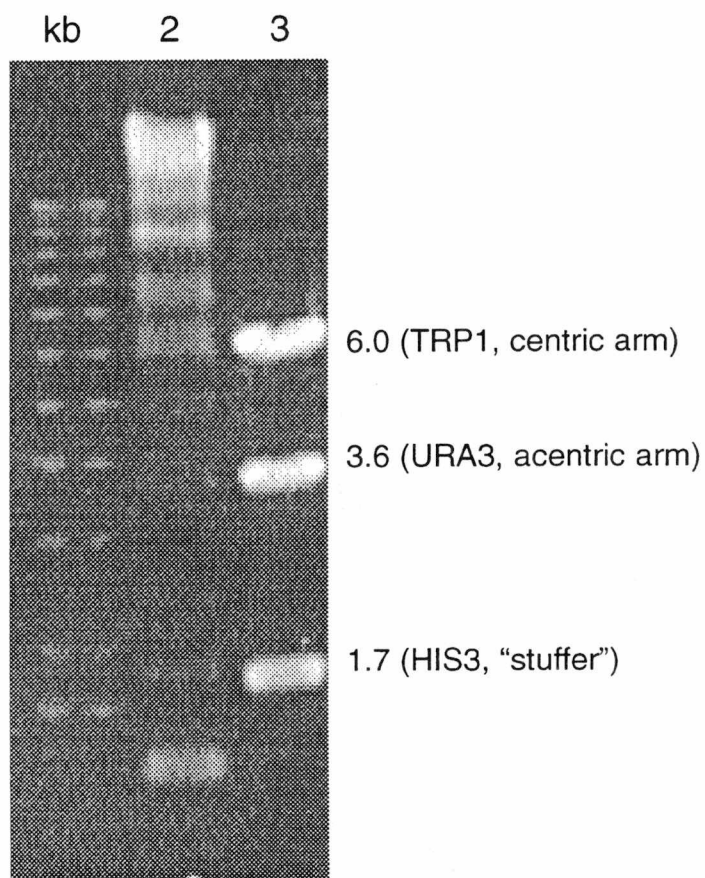


Figure A3.3. YAC4 vector preparation. The plasmid was digested to completion with BamHI and EcoRI, and treated with 0.05 units/ μ g shrimp alkaline phosphatase. 100 ng of treated vector was incubated with ligase (lane 3), and with ligase plus kinase (lane 1). The vector can ligate to itself only in the presence of kinase, showing that the dephosphorylation was successful.

8. Ligation reaction

The gel slice containing focused partial digestion products is washed two times with TE buffer.

The slice is equilibrated with ligation buffer twice for 30 minutes at room temperature.

The NaCl and polyamines in this buffer have been determined to protect high molecular weight DNA from degradation during melting of the LMP and subsequent manipulations of the DNA (Larin *et al.*, 1991).

A compression zone prepared from four 40 μ l plugs contains 2 to 3 μ g of DNA in a volume of about 300 μ l. The equilibrated gel slice is sliced into 2 to 3 mm sections and placed in an Eppendorf tube. An amount of vector equal in mass to the DNA in the slice is added, and the contents of the tube are weighed.

The agarose is melted at 68°C for 20 minutes. Meanwhile, a 10x stock of ligation buffer (NEB) is warmed to 37°C. The tube containing the melted agarose is placed at 37°C and one tenth volume of the ligation buffer concentrate is added. The buffer may be mixed by gently inverting and rotating the tube a few times. After 10 minutes, 4 μ l T4 DNA ligase is added to the reaction (NEB, 400 units/ml). This is also mixed through the molten agarose by inverting and rotating the tube. The reaction is left at 37°C for 30 to 45 minutes, and ligation is then left to proceed overnight at room temperature.

About one hour before the time of transformation (at the time when the cells are harvested), the ligation mixture is melted at 68°C for 20 minutes. The tube is transferred to a 45°C water bath and after 10 minutes 1 μ l of gelase is added to digest the LMP agarose. The DNA solution should be aqueous by the time of transformation.

B. Yeast transformation

This procedure is based on the high-efficiency spheroplast transformation protocol of Burgers and Percival (1987).

1. Solutions

- YPD

Autoclave 20 g peptone and 10 g yeast extract in 900 ml water. Separately autoclave 20 g dextrose in 100 ml water, and mix

- 1 M sorbitol- 182 g sorbitol/liter

- SPEM- 1 M sorbitol

10 mM Na Phosphate, pH 7.5 (1:50 of 0.5 M stock)

10 mM EDTA (1:50 of 0.5 M stock)

30 mM β -mercaptoethanol (1:500 of 14 M solution)

Prepared fresh immediately before use

- lyticase stock, 5 mg/ml in

50 % glycerol (1 ml)

50 mM NaCl (20 ml)

0.1 mM EDTA (0.5 ml)

10 mM Na Phosphate (40 ml)

30 mM β -mercaptoethanol (4 μ l)

H₂O to 2 ml

Store at -20°C in 200 μ l aliquots

STC

1 M sorbitol

10 mM Tris.HCl, pH 7.5

10 mM CaCl₂

Prepare from stocks immediately before using

- carrier DNA

5 mg/ml herring sperm DNA in sterile water. Shear through an 18G 1 1/2 needle. Store in small aliquots at -20°C.

- PEG solution, 20% PEG 8000 (Sigma), 10 mM Tris.HCl, pH 7.5, 10 mM CaCl₂

Autoclave a solution of 20% PEG, 10 mM Tris.HCl. Add 1:100 volume of 1 M CaCl₂ immediately before use

- SOS

1 M sorbitol

6.5 mM CaCl₂ (660 ml/100 ml)

0.25% yeast extract (250 mg/100 ml)

0.5% bactopectone (500 mg/ 100 ml)

Supplement with uracil and tryptophan to 20 mg/liter

pH 5.8, filter sterilize and store in 10 ml aliquots at -20 °C

- TOP agar (also used to pour plates)

Autoclave 182 g sorbitol in 400 ml water

Autoclave 20 g agar in 500 ml water

Autoclave 20 g dextrose in 100 ml water

Mix sorbitol and dextrose solutions, and dissolve 0.8 g -URA dropout

powder (BIO101, Inc.), 1.7 g yeast nitrogen base (YNB) w/o amino acids and ammonium sulfate (Difco), 5 g ammonium sulfate

Mix in the agar solution, cool down to 45 °C before pouring plates and plating spheroplasts

2. Preparation of spheroplasts

Streak the host strain AB1380 out to single colonies on YPD plate.

Use a single colony to inoculate 10 ml of YPD. Allow it to grow overnight to saturation (5×10^7 - 10^8 /ml) at 30°C, 250 rpm. Aliquots of this culture may be used for a week to start larger volumes for transformation.

Inoculate 200 ml of YPD in a one liter flask with a few hundred ml of the overnight culture. When the cells are harvested for transformation, their density should be 3×10^7 /ml (counted on a hemocytometer). Depending on what time the transformation is to be started the next day, the inoculum may be calculated assuming a doubling time of one and a half hours.

When the cells have reached the right concentration, they are harvested by spinning in 50 ml conical centrifuge tubes at 1500 rpm for 4 minutes. New tubes are used for each transformation, because traces of detergent and other chemicals that might be present in tubes that have been washed and autoclaved can affect the transformation adversely.

The cells in each tube (from 50 ml of culture) are rinsed with 20 ml sterile water and pelleted again.

A final wash with 20 ml 1 M sorbitol is performed before spheroplasting.

The cells are resuspended in 20 ml SPEM. Set amounts of lyticase are added to each tube, in the range of 10 to 30 ml.

Spheroplasting is carried out at 30°C for 15-20 minutes with very gentle shaking (lay the tubes on their side in the platform shaker).

During this time, 800 µl of water and 800 µl of 1M sorbitol are added to plastic cuvettes, one pair for every point in the lyticase titration. The spectrophotometer is zeroed in the absorbance mode at a wavelength of 800 nm using SPEM as a blank .

At the end of spheroplasting, 200 µl of cells from each tube are quickly added to a cuvette

containing water and to one containing sorbitol. The remaining cells are pelleted gently at 100xg (900 rpm) for 4 minutes. From this point forward in the procedure, the spheroplasts should be handled with care when they are being resuspended or pipetted.

While the cells are in the centrifuge, the absorbance of the samples in the cuvettes is measured. In water, spheroplasts will burst due to hypoosmotic shock, and the absorbance will decrease in proportion to the degree of spheroplasting. The absorbance of the samples in water and sorbitol is compared. From the range of lyticase treatments, the tube representing 60-70% spheroplasting is chosen for transformation:

$$\% \text{ spheroplasting} = 1 - ([A]_{\text{water}}/[A]_{\text{sorbitol}})$$

The SPEM is decanted from the pelleted spheroplasts, and they are gently resuspended in 20 ml 1 M sorbitol containing 10 mM Na Phosphate, pH 7.5.

The spheroplasts are pelleted for 4 minutes at 100xg.

The cells are gently resuspended in 1.5 ml of STC, to achieve a concentration of $3\text{--}5 \times 10^8/\text{ml}$. They are now ready for transformation.

3. Transformation

Add 1/100 volume of carrier DNA, and 1/100 volume of 100x polyamines. The use of polyamines has been reported to eliminate an extreme bias against large molecules in transformation (Connelly *et al.*, 1991). All solutions that come into contact with the DNA from this point forward in the procedure should contain 1x polyamines. The reason for the elimination of size bias may be compaction of the DNA molecules, which could have the effect of protecting longer molecules from shear. Other groups have reported that the use of polyamines did not improve the average insert size of their YAC libraries, and reduced the efficiency of transformation (*eg.* Kleine *et al.*, 1993). In our hands, addition of polyamines does not decrease efficiency of transformation, and reduces the proportion of clones with smaller insert sizes. **However**, polyamines definitely have an adverse effect on the efficiency of transformation with the control plasmid. So if polyamines are to be included, the control transformation set in parallel with plasmid should not contain polyamines. This makes the procedure a little more awkward. The procedure from here on is described for transformation with the ligation mixture, and it is understood that all steps are the same for the control plasmid transformation, except that polyamines are left out.

The spheroplasts are added 300 μl at a time into new 17x100 mm sterile culture tubes. To

the control tube, 10 ng of plasmid DNA are added. The plasmid we use is pSR316. It is used at a concentration of 1-2 ng/μl in 0.2M NaCl.

Using a cut yellow tip, 50 μl of ligation mixture is gently added to each tube. The cells and DNA are mixed by swirling the tube a few times.

The tubes are left at room temperature for 10 minutes.

3 ml of PEG solution containing 1x polyamines are added to each tube using a 5 or 10 ml graduated glass pipette. The solution is added with just enough force to disturb the cells and mix them with the PEG.

After 10 more minutes at room temperature, the spheroplasts are pelleted at 100xg for 4 minutes and resuspended in 450 μl of SOS. The cells are incubated at 30°C for 20 minutes without shaking before they are plated.

The plating agar should be 45-48°C, not warmer. The plating volume is about 10 ml, and the cells are mixed with the agar by inverting the tube two or three times before spreading on the plate.

When the agar has solidified, the plates are turned upside down and incubated at 30°C.

C. Analysis of results of transformation

1. Solutions

- -URA-TRP double selection plates

Autoclave 20 g agar in 900 ml water

Autoclave 20 g dextrose in 100 ml water

Dissolve 0.7 g -URA-TRP dropout powder (BIO101, Inc.), 1.7 g YNB w/o amino acids and ammonium sulfate (Difco), and 5 g ammonium sulfate in the dextrose solution.

Mix with the agar and pour plates.

- 2xAHC

Autoclave in 900 ml: 3.4 g YNB, 10 g ammonium sulfate, 20 g casein-hydrolysate acid, 40 mg adenine hemisulfate, 100 ml conc. HCl. Add 100 ml autoclaved 40% dextrose.

This medium is used at 1.5x strength to grow YAC-containing yeast (e.g., to make 300 ml use 225 ml 2x medium and 75 ml water).

Supplement with 100 mg/ml ampicillin.

- SE

0.9 M sorbitol, 50 mM EDTA

- SEM

0.9 M sorbitol, 50 mM EDTA, 30 mM β -mercaptoethanol

- YPD w/ 20% glycerol

- ESP

0.5 M EDTA, pH 8.0, 0.5% N-Lauroyl-sarcosine, 0.5 mg/ml proteinase K.

Repicking transformants

Many colonies should be visible on the -URA transformation plates after 4 or 5 days. Theoretically, two-thirds of these should be genuine YAC-containing colonies. In practice this fraction is lower, because of some background caused by the excess vector that is not removed.

YAC-containing clones are characterized by their ability produce strong red patches of growth on -URA-TRP medium. For unknown reasons, the growth rate of clones on both -URA and double selective medium is variable. Therefore, it can be useful to continue repicking colonies from the primary transformation plates for as long as ten days.

Colonies are patched onto double selection plates using sterile toothpicks (Forster square-center round). After 2 or 3 days a number of these colonies (20-30%) should regrow with a red phenotype.

3. Preparation of YAC DNA plugs

10-20 of clones from each transformation should be checked by PFGE to determine their average insert size.

Some cells are scraped up from each patch into 10 ml of 1.5x AHC containing 100 µg/ml ampicillin. This can be done by dropping toothpicks into 50 ml Corex tubes. The clones are grown for two days at 30°C, 250 rpm.

The cells are transferred to 15 ml plastic culture tubes and harvested by spinning in a table top centrifuge at 1500 rpm.

The cells are rinsed one time in SE and pelleted again.

The pellet from a 10 ml of culture is resuspended in 200 µl of SEM containing 10 µl 5 mg/ml lyticase. Spheroplasting is carried out at 37°C for one hour.

During this time, a 5 ml solution of 2% LMP agarose in SE is prepared and kept at 50°C.

The progress of spheroplasting is monitored by adding 10 µl of the suspension to 200 µl of sorbitol and 200 µl of sorbitol containing 1% SDS. SDS will clarify the suspension if spheroplasting is complete.

When spheroplasting is complete, 200 µl of the molten agarose is added to each tube. The mixture is pipetted up and down rapidly a few times and pipetted into a prechilled plug mold.

Each plug is extruded into a separate eppendorf tube. 1 ml of ESP is added and the plug is incubated at 50°C for at least 4 hours. The ESP is replaced with fresh buffer and the incubation continued for another 4 hours. The change in buffer is necessary to obtain good plugs.

NOTE: While this procedure gives good results that will allow a transformation to be evaluated by Southern hybridization, degraded DNA can make it hard to see the extra chromosomes on an ethidium bromide-stained gel. Better plugs are obtained if the cells are spheroplasted inside the agarose blocks.

To do this, embed the pellet of yeast cells in 150 µl 1% LMP agarose in SE. Just before adding the agarose to the pellets, when it has cooled to 37°C, add lyticase and β-mercaptoethanol to it in the amounts indicated above.

The plugs are incubated for 4 hours in sheroplasting solution, which is then replaced with ESP as described above.

4. PFGE analysis of YACs

Small pieces of the plugs (3-4 mm high) are loaded in the wells of a 1% gel cast in 0.5xTBE. The high concentration of salt in the plug will influence the mobility of the yeast chromosomes relative to size standards. To avoid this the gel can be left to equilibrate in the chamber for 30 minutes before the run, or the plugs can be equilibrated with running buffer before loading on the gel.

Running conditions can be varied according to the expected size range of the inserts. A ramped pulse time of 20-70 seconds provides good resolution across a wide range of sizes. To avoid running small inserts off the bottom of the gel, running times can be kept short (12-16 hours), or long gels can be employed.

After thorough staining in 0.5-1 µg/ml ethidium bromide and destaining for one hour, the yeast chromosomes and YACs can be visualized. Whether or not YACs can be distinguished depends on their size and the quality of the DNA plug. Degradation of DNA will give rise to background in the lanes which will make smaller YACs hard to see. Although the YACs are supposedly maintained in the same ratio as the host's native chromosomes, sometimes even relatively large YACs are somewhat faint. In the best case, extra chromosomes as small as 40 kb can be seen.

5. Analysis of YACs by hybridization

In some cases it may be desirable to transfer the gel and probe the membrane it with total soybean DNA, pBR322 plasmid, or a specific probe corresponding to a particular sequence in the genome.

pBR322 has sequence homology with the vector and will hybridize to all the YACs. Total soybean DNA will hybridize to most YACs with variable intensity (see Figure 8.7).

The gel and autoradiograms will help to evaluate the results of the transformation. On this basis it can be decided whether or not to store the transformants.

6. Storage of transformants

Colonies showing the correct phenotype are scraped from double selection plates into the wells of a 96-well plate containing 200 μ l of YPD w/ 20% glycerol.

Recovery plates are duplicated by aliquoting 50 μ l of the contents of each well into the wells of a second plate containing 100 μ l of YPD w/ 20% glycerol. The use of a 12-channel pipette will speed this step up and reduce errors (Costar Corp., Cambridge, MA).

The duplicate plates are sealed around the edges with parafilm and stored at -70°C in separate freezers.

7. Preparation of DNA from recovery plates

For the purpose of screening the library using PCR primers, plugs can be prepared containing DNA from all the YACs in a recovery plate. This allows all the YACs from a single plate to be screened for the presence of a specific sequence.

A plate from -70°C is thawed out and 10 μ l from each well is used to inoculate the corresponding wells of a second culture plate containing 150 μ l 1.5xAHC.

The plate is sealed with parafilm to prevent evaporation and incubated at 30°C for two to three days. The YACs show slight variation in their rate of growth. After three days a pink precipitate should be seen on the bottom of most wells.

The contents of the wells are resuspended and combined in one volume using a 12-channel pipettor (Costar Corp., Cambridge, MA).

Plugs are prepared from the plates using the procedure described previously. 200 μ l spheroplasting volume is sufficient for two plugs of DNA from each recovery plate. The quality of the plugs is not important for PCR applications.

8. PCR analysis of plate DNA plugs

The plugs are rinsed with a large excess of TE buffer two times at room temperature. Then they are dialyzed two at 50°C in TE with 1mM PMSF. Finally, one more rinse at room temperature with a large excess of buffer is performed.

A small piece of a plug is melted at 65°C for 20 minutes. 4 volumes of warm TE are added and the tube is vortexed to disperse the DNA and molten agarose. 1-2 μ l of this should be

sufficient for a PCR reaction.

The conditions for amplification are as follows: (95°C 3 min), 35x(95°C 5 sec, 60°C 30 sec, 72°C 2 min), 3x(95°C 3 min). Primer concentrations are 0.1 μ M, diluted 1:10 from a 1 mM stock into a reaction volume of 15 μ l.

10x PCR buffer: 500 mM KCl

100 mM Tris.HCl pH 8.3

15 mM MgCl₂

0.2% Triton X-100

0.2% Tween 20

100 mg/ml gelatin

2 mM dNTPs (diluted from 50 or 100 mM stock in 50 mM Tris)

Reaction conditions may be modified for particular products.

To check the quality of the DNA as a substrate for PCR, primers for the amplification of yeast actin genes are used. Two strong products between 500 and 1000 bp are seen. These primers can be used to ascertain that the plugs were washed thoroughly enough, and that the DNA concentration after dilution is adequate.

Actin primers:

Act1- GCC GGT TTT GCC GGT GAC GAC

21mer, 64°C, 33% AT

Act2- GGA AGA TGG AGC CAA AGC GGT G

22mer, 60°C, 41% AT

D. Amplification of insert ends using ligation to vectorette*

The idea of vectorette-assisted amplification is based on the ligation of the “bubbled” duplex to completely digested DNA of a particular clone with subsequent specific amplification. It aims at amplification of any sequence next to a known sequence (in this case the left and right arms of the YAC vector) to step aside from the known area. Because of the bubble and of the sequence of the vectorette universal primer, amplification of random fragments bearing ligated vectorette on both ends is prevented (Riley *et al.*, 1990).

We designed the vectorette for ligation with *Csp6I* ends (4-cutter) and alternatively two other 6-cutters with complementary ends, *AseI* and *NdeI*.

5'-TACAAGGAGAGGACGCTGTCTGTCTCGAAGGTAAGGAACGGACGAGAGAAGGGAGAG

3'- GTTCTCTCCTGTCGCTAAGAGCATGCTTGCCAATGCTAAGCTCTTCCCTCTC

3'-TCGCTAAGAGCATGCTTGCCAATGCTAAGC-5'

The lower primer is designated 'universal vectorette primer', and is used together with the YRP and YLP primers (see Figure A3.4) first for amplification and then for sequencing of end products.

1. Annealing of vectorette

To anneal the two strands of vectorette, mix corresponding oligonucleotides in equimolar amounts at 1 μ M final concentration in 50 mM NaCl, 1 mM EDTA, 10 mM Tris pH8.0, heat for 3 min at 90°C in a beaker and leave on the bench to cool down slowly.

2. Isolation of yeast DNA with glass beads

Grow 10 ml of yeast to saturation. Spin down the cells, resuspend in 0.5 ml water, transfer in an Eppendorf tube and spin down again. Resuspend cells in 200 μ l breaking

*This section of the procedures written by Alexander Kolchinsky

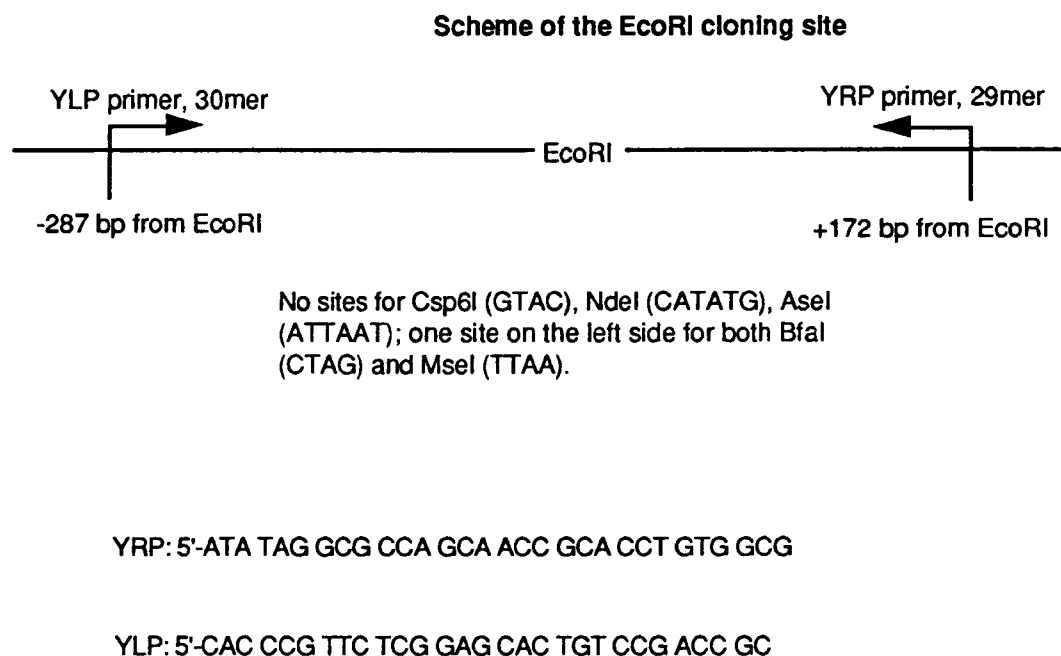


Figure A3.4. Scheme of the *Eco*RI cloning site of YAC4. The sequences of the YLP and YRP primers are shown, as well as their location relative to the cloning site.

buffer (2% Triton X-100, 1% SDS, 0.1 M NaCl, 10 mM Tris 8.0, 1 mM EDTA). Add 0.3 g glass beads (~0.2 ml) and 200 µl phenol/chloroform/isoamyl alcohol. Vortex at top speed for 3 min. Add 200 µl TE and vortex briefly. Centrifuge for 5 min, transfer aqueous layer in a fresh tube and add 1 ml 100% EtOH. Mix. Centrifuge for 3 min. Wash pellet with cold 70% EtOH and dissolve in 0.4 ml TE. Add 30 µl RNase (1 mg/ml), incubate 5-10 min at 37°C, add 10 µl 4M NH₄CH₃COO and 1 ml EtOH. Mix, centrifuge for 3 min, wash pellet with 70% EtOH, dry and dissolve in 100 µl TE. Expected yield 20 µg.

3. Digestion of yeast DNA

Digest 1 µg (5 µl) with 10 units of the enzyme of choice. Treat with phenol/chloroform to inactivate. When precipitating DNA, use 20 µg glycogen as carrier.

For ligation mix 12.5 µl digested DNA (250 ng), 1 µl 2 µM annealed vectorette, 5 µl 10x ligation buffer, water to 50 µl and 1 µl ligase (NEB) and incubate at r.t. 3-16 h. Add water to 250 µl and store in a freezer.

4. Primary amplification

Assemble PCR reaction in 20 µl final volume. Use 2 µl of ligated DNA preparation, 1 unit of Amplitaq-polymerase, final concentration of primers 0.1 µM.

Run mock samples (no DNA, no ligation). Run samples with combinations of primers YLP+vectorette and YRP+vectorette separately. The cycles should be:

Melting - one cycle of 3 min at 95°C

35 cycles of 5 sec at 95°C, 30 sec at 60°C and 2 min at 72°C

One cycle of 5 min at 72°C

Load samples on 1.5-2% agarose gel.

5. Secondary amplification

After staining and taking pictures, cut prominent band from each lane, freeze, thaw and use 1 µl for reamplification in desired volume. Secondary amplification products can be sequenced without cloning or used for labeling (see Chapter 8 of this thesis).

APPENDIX 4

HIGH EDTA CONCENTRATIONS CAUSE ENTRAPMENT OF SMALL DNA MOLECULES IN THE COMPRESSION ZONE OF PULSED FIELD GELS, RESULTING IN SMALLER THAN EXPECTED INSERT SIZES IN YACS PREPARED FROM SIZE SELECTED DNA*

Many procedures for the preparation of yeast artificial chromosome (YAC) libraries use size selection of insert DNA by pulsed field gel electrophoresis (PFGE) to remove small molecules and increase the average size of clones (Anand *et al.*, 1989). In most cases, the compression zone containing the focused partial digestion products is used for ligation to artificial chromosome arms. In some cases, particularly in the construction plant YAC libraries, electrophoresis is used to remove short molecules from the DNA plugs themselves, which are then retrieved from the gel and used for ligation and transformation (e.g., Edwards *et al.*, 1992; Evers *et al.*, 1992). Both approaches have been used successfully, although the average insert size with the latter procedure is shorter.

Here we demonstrate that the method used to stop the partial digestion reaction, and in particular the level of EDTA in the DNA plugs before loading them on a gel, drastically influences the size distribution of molecules compressed in the limiting mobility band. This is reflected in the insert sizes of YAC clones prepared from the focused DNA.

Chromosomal soybean DNA was embedded in agarose plugs and prepared for restriction digestion as described (Funke *et al.*, 1993). Partial digestion with *EcoRI* was carried out using 0.4 to 4 units of enzyme per 40 μ l plug in a final volume of 100 μ l. The reactions were quenched by adding 100 μ l of 100 mM EDTA (to produce a final EDTA concentration of 50 mM) or 100 μ l of 500 mM EDTA as in (Evers *et al.*, 1992). The reaction mixtures were incubated at 37°C with 1 mg/ml proteinase K. The plugs were then loaded on a 0.9% LMP agarose gel and electrophoresed for 12h at 200V, 3s switch interval, in 0.25 x TBE. Partially digested DNA was recovered from the compression zone and ligated to dephosphorylated YAC4 vector. The ligation mixture was used directly for transformation of yeast strain AB1380 as described (Burgers and Percival, 1987).

Transformants showing strong patches of red growth on URA⁻TRP⁻ medium were picked and analysed by PFGE. YACs prepared from partial digestion reactions that were quenched with 50 mM final EDTA concentration were between 100 and 200 kb in size. Clones from reactions that were terminated with 500 mM EDTA were all shorter than 100 kb. This is in

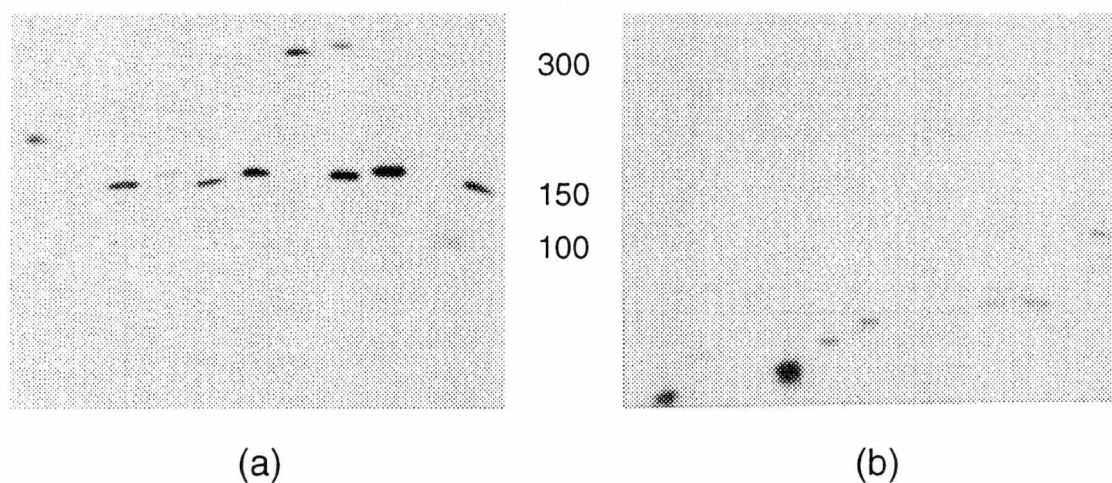
*Funke, R.P, Kolchinsky, A., and Gresshoff, P.M. (1994). *Nucleic Acids Res.* 22, 2708-2709.

spite of the fact that the same switch interval was used to compress the partial digestion products in both cases (Figure A3.1).

To determine whether the procedure for quenching the reaction could be responsible for this discrepancy, plugs containing soybean chromosomal DNA were equilibrated with running buffer (0.25 x TBE), 50 mM EDTA, or 500 mM EDTA. The plugs were then loaded on an LMP focusing gel and subjected to electrophoresis as described above. The compression zones from each plug were excised from the gel and loaded on a second gel to determine the size distribution of the focused DNA molecules. This gel was run to resolve DNA molecules up to about 1 Mb in size. The gel was transferred and probed with total nick translated soybean DNA (Figure A3.2).

Both the ethidium bromide fluorescence and hybridization showed that rinsing the plugs in 500 mM EDTA before PFGE fractionation dramatically affects the size distribution of molecules in the compression zone. Fragments longer than 1 Mb (focused in the limiting mobility band of the second gel) are less numerous in the compression zone from these plugs as compared to plugs rinsed in running buffer or 50 mM EDTA. In addition, the compression zone contains a sizable fraction of molecules much shorter than the cut-off size of the focusing gel (150 kb). The smear in the second gel extends to less than 50 kb and the average size of molecules in the smear, judged from fluorescence and hybridization, is lower compared to the other two treatments. Shorter molecules will ligate to vector at higher frequency than long partial digestion products, and will subsequently transform yeast more efficiently.

These observations explain our experience with compression zones from plugs prepared from partial digestions that were quenched with 500 mM EDTA. In published procedures where high EDTA concentrations were used prior to PFG size selection, the original plugs and not the compression zone were retrieved from the gel and used for ligation and transformation (Eyers *et al.*, 1992). Interestingly, in these experiments, a switching regime was used under which molecules longer than about 300 kb would be expected to leave the plug and be focused in the compression zone (10s, 200V, 3h run time). The results presented here suggest that the level of EDTA in the plugs at the time of loading is critical in determining how molecules are partitioned between the compression zone and



FigureA4.1. Soybean YACs from compression zones focused from partial digestion reactions quenched with different concentrations of EDTA. a) Clones from reactions terminated with a 50 mM final EDTA concentration. b) Clones from reactions quenched by addition of an equal volume of 500 mM EDTA. Electrophoretically separated clones were transferred to membranes and probed with pBR322. Sizes are in kb. The clone on the far right in panel (b) is a standard from a previous transformation experiment.

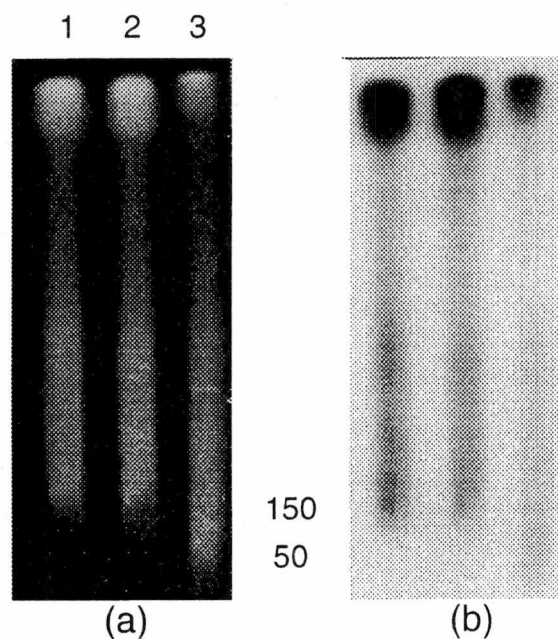


Figure A4.2. The effect of EDTA concentration in DNA plugs on the size distribution of DNA fragments focused in the compression zone. Soybean DNA plugs were equilibrated with running buffer (0.25xTBE; 1), 50 mM EDTA (2), and 500 mM EDTA (3) before subjecting them to electrophoresis in 0.9% LMP agarose, 3 s switch interval, at 200 V for 12 h in 0.25xTBE. a) The compression zones from this gel were removed and their contents analyzed by subjecting them to a second electrophoresis in 1% agarose, 40 s switch interval, at 200 V for 20 h in 0.25xTBE. b) The gel was transferred and probed with nick-translated total soybean DNA. Size standards were λ ladders (BioRad) and yeast chromosomes from *Saccharomyces cerevisiae* strain AB1380.

the plugs remaining in the well. The combination of high EDTA concentrations and the short run time may cause longer molecules to remain in the well. If the compression zone is to be used eventually for cloning, plugs should be rinsed in running buffer or no more than 50 mM EDTA in order to ensure that the focused partial digestion products are of the size anticipated from the cut-off pulse time.

APPENDIX 5

PULSED FIELD GEL ELECTROPHORESIS*

Abstract

Pulsed field gel electrophoresis extends the size range of resolution of DNA molecules to many megabases. This bridges a gap that existed in techniques for the structural analysis of large genomes. The ability to analyze large molecules of DNA has enabled the development of cloning vectors with large capacities such as yeast artificial chromosomes, and permitted applications including the transfer of artificial chromosomes containing intact gene clusters to mammalian cells. It has also brought the creation of artificial chromosomes for mammals and plants into the realm of feasibility.

INTRODUCTION

Pulsed field gel electrophoresis (PFGE) resolves molecules of DNA in the range of 10 kilobases (kb) to several megabases (Mb). This cannot be achieved by conventional electrophoretic techniques. PFGE was developed twelve years ago by Schwartz *et al.* (1982). Since then, both protocols and equipment have become routine in almost every laboratory working with DNA, and a comprehensive practical manual has been published (Birren and Lai, 1993). This book covers virtually every aspect of PFGE, contains 900 references, and is written in an extremely user-friendly manner. Another useful manual was published by Burmeister and Ulanovsky (1992). Therefore, we will discuss only very recent publications in light of the most important applications of the technique and will present a protocol based on our experience that can be easily accommodated for other tissues and purposes.

DNA molecules separated in agarose under a constant electric field are retarded by the agarose matrix. This retention correlates inversely with the size of the molecule: DNA molecules adopt a loose compact conformation, and the longer the molecule, the larger the globule. However, longer DNA molecules adopt a stable rod-like conformation and their cross-section and retention by the agarose matrix are not determined by their size anymore. As a result, all DNA molecules above a particular cut-off size have the same mobility. The cut-off size is determined by the concentration of agarose and conditions of electrophoresis, but hardly exceeds 50 kb.

*In press. Kolchinsky, A. and Funke, R.P. in "Molecular Biology: Current Innovations and Future Trends", (Hugh Griffin, ed.). Horizon Scientific Press, Norfolk, England.

every change, the random coils of DNA molecules change their conformation and orientation. This process is called relaxation. The time needed by individual molecules to realign themselves with the new field before recommencing their movement in that direction depends on their size. The molecules follow a zigzag path, and their net movement down the gel is again inversely correlated with their size.

DNA resolved by PFGE has a limit of resolution which is determined primarily by the pulse time, and to a lesser extent by other parameters of the run. All fragments above a specific size move together forming the so-called compression zone (CZ), or zone of limiting resolution. For instance, at a 2 sec pulse time in 0.25 x TBE the CZ comprises fragments longer than approximately 70 kb, at 10 sec - 300 kb, at 1 min. - 1 Mb, at 10 min. - 3 Mb.

There are many designs of gel boxes for PFGE based on different geometries of electric fields, and they are described in the reference mentioned above (Birren and Lai, 1993). The simplest of these is the field-inversion gel electrophoresis (FIGE) and the most popular is the contour-clamped homogenous gel electrophoresis (CHEF).

FIGE can be performed in any conventional horizontal chamber. The electric field is periodically inverted and forward movement of DNA is achieved either by longer pulses or higher voltage applied in one direction. This system can be built in any electric shop by inserting a switching device between the power pack and the chamber. Since the angle of the electric field in FIGE does not change, the lanes are straight and parallel. Its disadvantages are that it gives somewhat fuzzy bands above 500-750 kb and generates artifactual inversion of mobility for some size ranges, with fragments moving slower than expected.

CHEF exploits hexagonal geometry of electrodes (Figure A5.1), and the lanes are straight and parallel (see below). In the manufactured device CHEF-DRIII made by BioRad Laboratories, the voltage applied to each of 24 electrodes is adjusted in order to provide even better uniformity across the gel. The angle between the fields is maintained at 120° in an older device, CHEF-DRII, and is adjustable in newer systems CHEF-DRIII and CHEF-MAPPER to optimize separation of very large fragments.

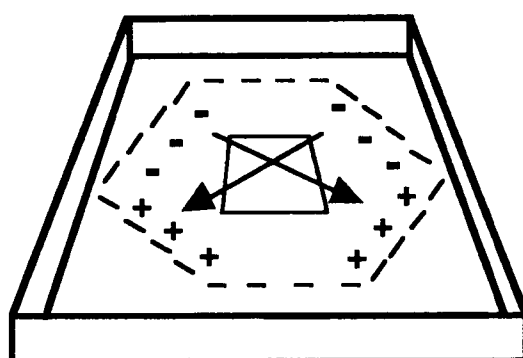


Figure A5.1. Scheme of the gelbox for CHEF. Electrodes are located in a hexagonal configuration, the gel is shown as a rectangle in the middle. Alternating electric fields are shown by crossed arrows.

It was mentioned that the range of resolution of PFGE is determined primarily by pulse time. To improve the linearity of resolution, a gradually changing (ramped) pulse time throughout the run is often used. This also solves the problem of size reversal that is sometimes encountered in pulsed field migration.

PFGE: CURRENT APPLICATIONS

PFGE became an indispensable means to bridge the gap between three levels of structural studies of eukaryotic genomes: genetic, cytogenetic, and molecular. Optical resolution of metaphase chromosomes usually corresponds to several megabases of DNA, while conventional electrophoresis resolves only fragments up to 50 kb. The resolution of genetic analysis in higher eukaryotes rarely goes beyond 1-2 centimorgans, which again corresponds to megabase(s) of DNA. Thus, PFGE enables researchers to construct physical maps with resolutions spanning these extremes.

A technique that was enabled by PFGE and became a major tool in genome research was cloning of DNA in yeast artificial chromosomes (YACs; for review see Monaco and Larin, 1994). Prior to the introduction of YACs the largest fragments amenable to cloning did not exceed 40 kb; YACs can hold 100 kb-2 Mb of foreign DNA. Naturally, the analysis of these cloned inserts was unthinkable without PFGE. Later, cloning in YACs was supplemented with alternative procedures for cloning large DNA fragments (Monaco and Larin, 1994). These cloning techniques covered the gap between the emerging power of physical mapping and limited size of cloned DNA fragments.

Integration of genetic and physical maps enables one to undertake positional cloning of a gene of interest. The prerequisites for positional cloning include a well defined gene that confers the phenotype of interest, a set of molecular markers tightly linked to the gene, a physical map of the corresponding segment with molecular markers on the map, and a library of large DNA inserts cloned, for instance, in YACs. Because of the resolution range of PFGE, a detailed genetic map around the gene can be correlated with the physical map, and the genomic region encompassing the gene can be cloned in YACs.

With this in view, it is quite important to determine the correlation between physical and genetic distances in the genome to estimate the approximate number of linked molecular markers required for complete coverage, and length of DNA to be cloned in search of a gene starting from linked molecular markers. A preliminary estimate can be produced simply by dividing the total physical size of a haploid genome by its genetic size deduced

from its detailed genetic map, if available. However, real mapping shows that this value varies enormously throughout genome. First, it was found recently that telomeric segments of chromosomes are highly enriched in genes (see, for example, Moore *et al.*, 1993). The term gene is used here in its traditional meaning, an elementary unit of recombination. In other words, recombination frequency near the ends of chromosomes is increased. In addition, fluctuations of recombination rate along chromosomes were estimated for many organisms. For example, physical and genetic mapping of a 470 kb segment of maize DNA containing two genes showed that recombination frequencies in the immediate vicinities of the genes are increased as compared with intergenic intervals (Civardi *et al.*, 1994). The distances are 217 kb/cM (centiMorgan, the elementary unit of genetic recombination) for a 1 kb segment inside the locus, and 1560 kb/cM for the intergenic region.

Physical mapping in chromosome areas with the lowest recombination frequencies, near centromeres, is hampered by a shortage of DNA markers in those regions. Centromeric areas consist of arrays of tandem repeats with little variability. In some cases molecular markers for these areas have to be found by the analysis of large fragments of DNA resolved by PFGE (Warburton and Willard, 1990; Kolchinsky and Gresshoff, 1994).

In some cases physical mapping goes to the margin of PFGE resolution. Although resolution of fungal chromosomes up to 10-12 Mb long has been reported, it is difficult to resolve fragments longer than 2-2.5 Mb obtained by restriction digestion and revealed by hybridization to molecular probes. The longest restriction fragments used in mapping, that we were able to find, were published recently by Ambrose *et al.* (1994) and reached 5.7 Mb. The electrophoresis took six days because at this resolution, the applied voltage gradient must be very low. The authors were able to locate an important gene for ataxia-telangiectasia on a 3 Mb fragment between two molecular markers.

Physical mapping of large genomes encounters several major problems. First, ordering of sites for the same restriction enzyme requires partial digestion of DNA. In many cases, this is achieved naturally, because of incomplete and/or variable methylation of the restriction sites. Most rare cutting restriction enzymes are sensitive to methylation of the sequences CpG found in animals and CpG and CpXpG found in plants. In addition, different sources of high molecular weight DNA used for mapping, like cell lines in mammals and varieties in plants, may show variable methylation patterns. We found marked differences even in different batches of soybean leaves, probably because of minor variations in growth conditions (unpublished). On the one hand, this complicates the interpretation of the data. On the other hand, it can be very useful in the analysis of partial digests, as, for example,

in a paper on the detailed mapping of kallikrein genes in rat genome (Southard-Smith *et al.*, 1994).

A more difficult task is to establish colinearity of cloned products of partial digests with their genomic counterpart. This can be achieved by so-called RARE mapping (Ferrin and Camerini-Otero, 1991). First, DNA around the *EcoRI* sites that are to be mapped is sequenced and a complementary oligonucleotide is synthesized. In the presence of the RecA protein, it replaces one strand in the target site and protects it from methylation, while the rest of DNA is methylated with *EcoRI*-methylase. Then the oligonucleotide and proteins are washed away and the DNA is treated with *EcoRI*, which is expected to introduce just one cut at the non-methylated site. By this approach Ferrin and Camerini-Otero (1994) and Negorev *et al.* (1994) were able to establish colinearity of cloned fragments to their sources.

Often markers from molecular maps detect a mixture of monomorphic and polymorphic bands in restriction digests used to score polymorphisms. When the same marker shows multiple bands in rare-cutter restriction digests, it is difficult to determine which band corresponds to the polymorphic copy of the marker and whether some of the bands represent partial digestion products. By digesting the entire lane of PFGE resolved DNA with the enzyme that revealed the polymorphic band and separating the products on a conventional gel perpendicular to the direction of the first electrophoresis, it is possible to assign copies of the marker to individual bands on the PFGE Southern. This innovation was originally used to produce a physical map of the immunoglobulin heavy chain variable region genes of the human genome (Walter *et al.*, 1990), and has been adapted for genomic analysis of tandemly repeated DNA (Warburton and Willard, 1990) and physical mapping of homoeologous chromosomal regions of the soybean genome (Funke *et al.*, 1993)

Partial digests become more important in the mapping of YAC inserts. There is a growing consensus that so-called CpG islands are the best landmarks for genes in eukaryotic DNA (Larsen *et al.*, 1992). The mapping procedure can be illustrated by work recently published by Coleman *et al.* (1994). After finding YACs in the area of interest, the authors generated partial digests with restriction enzymes that are enriched for CpG sequences in their recognition sites, such as *NotI*, *BssHIII*, *EagI* and *SacI*. CpG islands show coincidence of at least three different restriction sites within a very short distance. The result looks like a band of partial products that spans several lanes on the Southern blot.

Occasionally, PFGE fragments are used as molecular markers. It applies to simple tandem

repeats that cannot be converted into shorter informative hybridizing fragments. One application of these markers to map centromeres was mentioned above. It proved very useful in the mapping of plant telomeres as well (see review by Kolchinsky and Gresshoff, 1994). Plant telomeres consist of 7 bp telomeric repeats that form arrays up to 60 kb long and subtelomeric tandem satellite repeats that can occupy up to several hundred kb. Both of these tandem repeats were successfully used in several plant species to generate molecular markers and put them on existing maps (Kolchinsky and Gresshoff, 1994).

Sometimes PFGE is used to produce electrophoretic karyotypes of DNA from organisms with small genomes. Single chromosomes of bacteria or eukaryotic organelles are restricted by a rare cutting enzyme and the resulting fragments resolved by PFGE. Use of partial digests and/or several enzymes makes it feasible to construct a physical map of the molecule of interest. The resolution of individual chromosomes of eukaryotes is usually more challenging because of their size. For example, authors of a recent paper were able to resolve all eight chromosomes of *Aspergillus niger* measuring up to 6.6 Mb (Verdoes *et al.*, 1994). Because some of the chromosomes have the same mobility on PFGE, the authors constructed strains in which they artificially enlarged individual chromosomes by integrating multiple copies of the *glaA* gene. As a result they were able to differentiate the chromosomes by PFGE and accurately align the genetic and physical maps by Southern hybridization. One PFGE run took six days, as in a physical mapping case mentioned above.

Although the great majority of physical mapping by PFGE uses hybridization to reveal fragments complementary to the probe, there is an alternative approach based on PCR. Pieces of pulsed-field gels in the desired size range can be used in PCR reactions directly, after melting and dilution of agarose to prevent solidification, or after treatment with agarase. These fractions can be amplified with arbitrary primers for further cloning (Kolchinsky *et al.*, 1993) or for localization of a fragment corresponding to the probe of interest (Denton *et al.*, 1994).

FUTURE TRENDS

Recently, it became possible to transfer large segments of DNA cloned in YACs into mammalian cells. This is important for several reasons. First, the expression of mammalian genes in transgenic cells depends on flanking sequences (Gaensler *et al.*, 1993; Peterson *et al.*, 1993), and incorporation of a large segment surrounding the gene of interest alleviates these fluctuations. Second, some mammalian genes are so large that functional copies can

be cloned and handled in YACs or other vectors of similar capacity only (see, for example Lamb *et al.*, 1993). Third, some important genes form large clusters and must be transferred together for proper functioning and further studies (see, for example, transfer of HLA genes reported by Demmer and Chaplin, 1993). Eventually, the transfer of large segments of DNA and assembly of functional artificial chromosomes of higher plants and animals will be one of the most challenging applications of PFGE in genetic research.

The creation of vectors for gene therapy will make use of PFGE technology. In a recent work, a combination of an adenovirus and a YAC was obtained that could be a promising vector for genetic manipulations inside the yeast cells and for the delivery of genes into mammalian cells for gene therapy (Ketner *et al.*, 1994).

EXPERIMENTAL SECTION

We are going to describe a mapping experiment by PFGE. It will include all major steps for the preparation of DNA for physical mapping or cloning and two-dimensional electrophoresis.

Preparation of megabase DNA in plugs

DNA molecules in solution are extremely sensitive to shearing, and isolation procedures that involve pipetting and ethanol precipitation usually yield DNA with an average molecular weight 20-50 kb. To avoid handling DNA in liquid, suspensions of live cells are mixed with low melting point agarose and embedded in plugs or blocks. The embedded cells are then lysed with detergent in a high-EDTA buffer. Nucleases and cellular debris are removed by diffusion in the presence by proteinase K, leaving high molecular weight chromosomal DNA essentially intact and accessible to restriction enzymes.

This procedure was first described by Schwartz and Cantor (1984), who used it to produce the first electrophoretic karyotype of yeast. It has been adapted to study DNA from many organisms. Plants pose a special problem because of the cell wall which must be removed. High nuclease activities result in a certain amount of degradation of the DNA, and deposits of starch or phenolic compounds in the agarose blocks may interfere with the diffusion or activity of restriction enzymes. Procedures have been developed for the isolation of megabase DNA from nuclei. This is reported to be an effective method with the added advantage that chloroplast and mitochondrial DNA are not included in the DNA plugs or microbeads (Wing *et al.*, 1993; Liu and Whittier, 1994). The following protocol was designed for the preparation of megabase DNA starting from soybean leaf tissue, but may

be readily adapted for other organisms. The density at which the cells are embedded should be adjusted depending on the nuclear DNA content of the experimental organism and on the desired final DNA concentration in the plugs. For accurate sizing of high molecular weight fragments, DNA concentration should not be too high (80-120 µg per ml; when DNA concentration is 200 µg per ml and more, the apparent size of fragments is significantly greater than their true size (Doggett *et al.*, 1992)).

1. Young expanding trifoliate leaves (1 cm or less in length) or primary leaves are harvested for the preparation of protoplasts. Healthy appearance of plants is extremely important for successful isolation. Surface sterilized 1-2 mm leaf strips are digested with 2% driselase (Sigma Chemicals), 0.2% pectolyase Y-10 (Sigma Chemicals), and 1% cellulase Onozuka R-10 (Serva) for 5 hours with gentle shaking in a buffer (PB) containing 0.4 M mannitol, 10 mM CaCl₂, 5 mM MES, and 5 mM ascorbic acid, pH 5.6.

2. Following digestion the mixture is filtered through a 43 mm nylon mesh and protoplasts are collected by centrifugation at 100xg for 10 minutes. The cells are rinsed once with PB. Cells are counted and after a second centrifugation step gently resuspended in PB. The suspension is then mixed with 1.5% agarose in PB kept at 45°C, and immediately cast into a chilled mould. Moulds can be designed to suit to individual preferences, and our DNA plugs have a volume of 150-200 µl depending on the tooth size of the comb we plan to use. The final concentration of cells should be 2 to 3 x 10⁷ per ml. Assuming that the diploid nucleus of soybean contains about 5 pg of DNA, this yields plugs with a nuclear DNA concentration of about 100 to 150 µg per ml. The total DNA concentration in the plugs is somewhat higher if the plastid and mitochondrial genomes are taken into consideration.

3. Plugs are transferred to three times their combined volume of ESP (0.5M EDTA, 1% sarkosyl, 1 mg per ml proteinase K) at 50°C while being gently agitated at 50 rpm. After 24 hours the solution is replaced with fresh ESP and left for a further 24 hours, at which point the plugs can be prepared for restriction digestion.

Restriction digestion and electrophoresis

1. Proteinase K is inactivated by washing the plugs twice for one-half hour in 10-20 volumes of TE buffer containing 1 mM PMSF (100 µl of 100 mM freshly prepared acetone stock per 10 ml of TE) at 50°C. The plugs are rinsed three more times with fresh changes of TE without PMSF.

2. One-third to one-quarter of a 200 µl plug (about 5-10 µg of DNA) is weighed in a tared

microcentrifuge tube. Treating the plug as void volume, water and 10x restriction buffer are added to make 100 μ l of 1x buffer and the tube is placed on ice for one half hour. The liquid is removed and replaced with 50 μ l 1x buffer and 10-20 units restriction enzyme per μ g DNA. The mixture is left at 4°C for an additional half hour to allow the enzyme to diffuse into the plugs before becoming fully active. Digestions are carried out at the recommended temperature for at least 6 hours, and terminated with the addition of one ml of ice cold TE. Plugs are sealed into the wells of a gel with a small amount of melted agarose.

3. Conditions for pulsed field gel resolution will depend on the size range of fragments produced by the restriction enzymes and the desired range of resolution. Some examples of restriction digests of embedded soybean chromosomal DNA resolved under different conditions of electrophoresis are shown in Figure A5.2. Switch time, temperature, buffer concentration, agarose concentration, and salt concentration in the sample block are among the factors that influence resolution of DNA fragments in pulsed field gel electrophoresis (Birren and Lai, 1993, Funke *et al.*, 1994).

4. It is important to include on the gel size standards appropriate to the range of electrophoretic resolution. In the range 200 kb to 2.2 Mb, it is convenient to use chromosomes of *Saccharomyces cerevisiae*; phage lambda concatemers are suitable size standards for the range 50 - 1000 kb. These may be purchased or prepared. We prefer standards from Sigma (St. Louis, MO), which come in convenient syringes. Before preparing yeast standards from cells, one should make sure that the size of chromosomes of the strain in use is well characterized.

Hybridization

The conditions for transfer and hybridization of pulsed field gels are similar to those for conventional gels, with the provision that high molecular weight DNA fragments can be nicked efficiently and consistently. Different laboratories use combinations and modifications of procedures that involve UV irradiation and treatment with 0.25 M HCl. The protocol we supply below works well in our hands, but conditions should be optimized for each laboratory. To do this, it is convenient to use lanes of resolved yeast chromosomes from a single gel. The DNA in each lane is nicked using different procedures (different lengths of time of HCl treatment, dose of UV irradiation, or combinations thereof), the gel is put back together and transferred as a whole. The whole gel is probed with labeled yeast chromosomal DNA and the signal quality obtained with each treatment

can be directly compared.

1. Stain the gel for 30 minutes with 1 $\mu\text{g/ml}$ ethidium bromide. It is important for ultraviolet cleavage of large DNA that all the fragments throughout the whole thickness of the gel is stained, since only DNA bound by ethidium bromide will be nicked. For clear photographs, destain the gel for at least one hour.
2. The DNA is nicked with 1500 μJ UV in a UV StratalinkerTM 1800, followed by depurination with 0.25M HCl for 20 minutes.
3. After briefly rinsing the gel with distilled water, the DNA is denatured for 30 minutes with 0.4 M NaOH.
4. The DNA is vacuum-transferred to membranes (BioTrans, ICN or Zeta-Probe, BioRad) in 10 x SSC for 2 hours at 40 cm H₂O.
5. Following transfer the membranes are rinsed in 2 x SSC and the DNA is crosslinked to the membrane by irradiation with 1200 μJ UV light.

Figure A5.2 shows an example of hybridization of a PFGE Southern blot probed with a soybean molecular marker.

Pre-electrophoresis

Highest quality megabase DNA is necessary for both cloning and long range mapping (fragments larger than 1 Mb). However, plugs almost always contain a fraction of relatively low molecular weight DNA (50-300 kb). The presence of this sheared DNA in plugs will yield a large number of fragments with at least one end that is not suitable for cloning. In mapping, it generates undesirable smearing of the signal. The degraded DNA may be removed by pre-electrophoresis prior to complete or partial digestion and cloning, leaving only very high molecular weight DNA in the plugs (Edwards *et al.*, 1992). The efficiency of the procedure can be improved by equilibrating the plugs with low ionic strength buffer before loading (Funke *et al.*, 1994).

1. DNA plugs are treated with TE buffer and PMSF as described above.
2. The plugs are placed in a long slot of a 1% gel.
3. Running conditions are 60 s pulse time, 4 hours, in 0.5x TBE, 14°C.

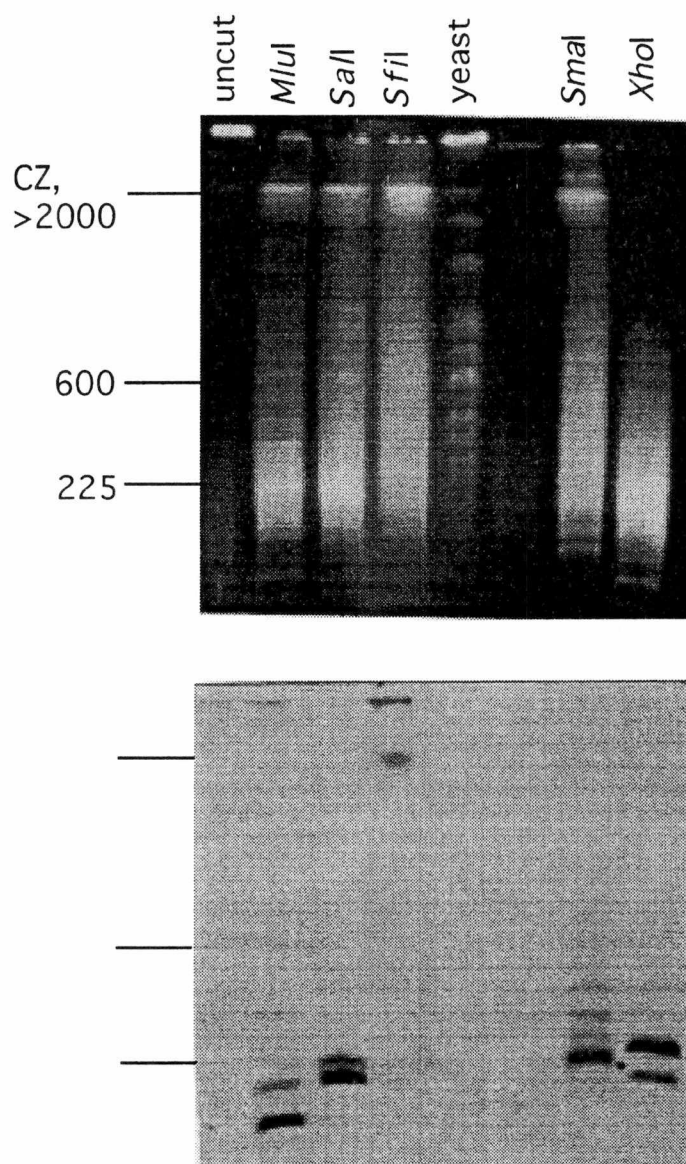


Figure A5.2. Digestion of agarose embedded soybean DNA with several restriction enzymes, and Southern hybridization with a marker from the soybean RFLP map. Conditions of electrophoresis were a 50s pulse time for 15 hours followed by a 90s pulse time for 9 hours in 0.5 x TBE, 200V, at 14°C. A CHEF-DRII apparatus was used in all experiments. The ethidium bromide stained gel shows that the majority of DNA that is not cut with a restriction enzyme remains in the well or in the compression zone. The appearance of characteristic discrete bands in the other lanes is evidence of complete digestion. Hybridization to the well and the compression zone of the lane containing the *Sfi*I digest shows that the *Sfi*I restriction fragments detected by the marker are too large to be resolved under these conditions of electrophoresis.

4. The DNA plugs are retrieved from the well and rinsed two times with TE prior to partial digestion and ligation. Modifications of conditions for partial digestion exist. One may use limiting amounts of enzyme for a set length of time, or perform a time course for incubation with a set amount of enzyme. The amount of Mg^{2+} may be titrated after incubating the plugs in restriction buffer lacking Mg^{2+} for 4 hours on ice (Albertsen *et al.*, 1990). Finally, different ratios of *EcoRI* and *EcoRI* methylase may be employed in a compromise buffer to achieve the desired degree of digestion (Larin *et al.*, 1991). A typical result with soybean DNA is shown in Figure A5.3. In this case, the different number of units of *EcoRI* were allowed to diffuse into the plugs (30-40 μ l each, in 50 μ l *EcoRI* buffer containing Mg^{2+}) for 30 minutes on ice. The digests were then placed at 37°C and terminated after one hour by the addition of 100 μ l 50 mM EDTA containing 1 mg/ml proteinase K. The removal of degraded DNA from the original plugs permits a more accurate estimation of the appropriate conditions for partial digestion.

Second dimension

1. For two dimensional electrophoresis, twice as much DNA (at the same concentration) is used in the first dimension PFGE to improve the signal intensity on hybridization.
2. After staining with ethidium bromide and photographing, lanes of fractionated DNA are cut out of the gel and equilibrated on ice in slender glass tubes with the buffer appropriate for the enzyme to be used in the second dimension for one hour.
3. About 200 units of the enzyme are added and allowed to diffuse into the slice for an additional hour before placing the digests at 37°C overnight.
4. The slices are briefly (15-20 minutes) equilibrated with running buffer.
5. They are then placed in a casting tray perpendicular to the direction of electrophoresis and embedded in a large gel. A normal plug of DNA not fractionated by PFGE and digested with the same enzymes is loaded alongside the slice in order to be able to identify individual copies of markers, members of a gene family, or repeated elements.
6. Following electrophoresis, the gel is transferred and blotted using conventional procedures.

A typical result showing how second dimension electrophoresis was used to assign individual copies of a marker in the soybean genome to separate bands produced by *SfiI* digestion on a PFGE Southern is shown in Figure A5.4.

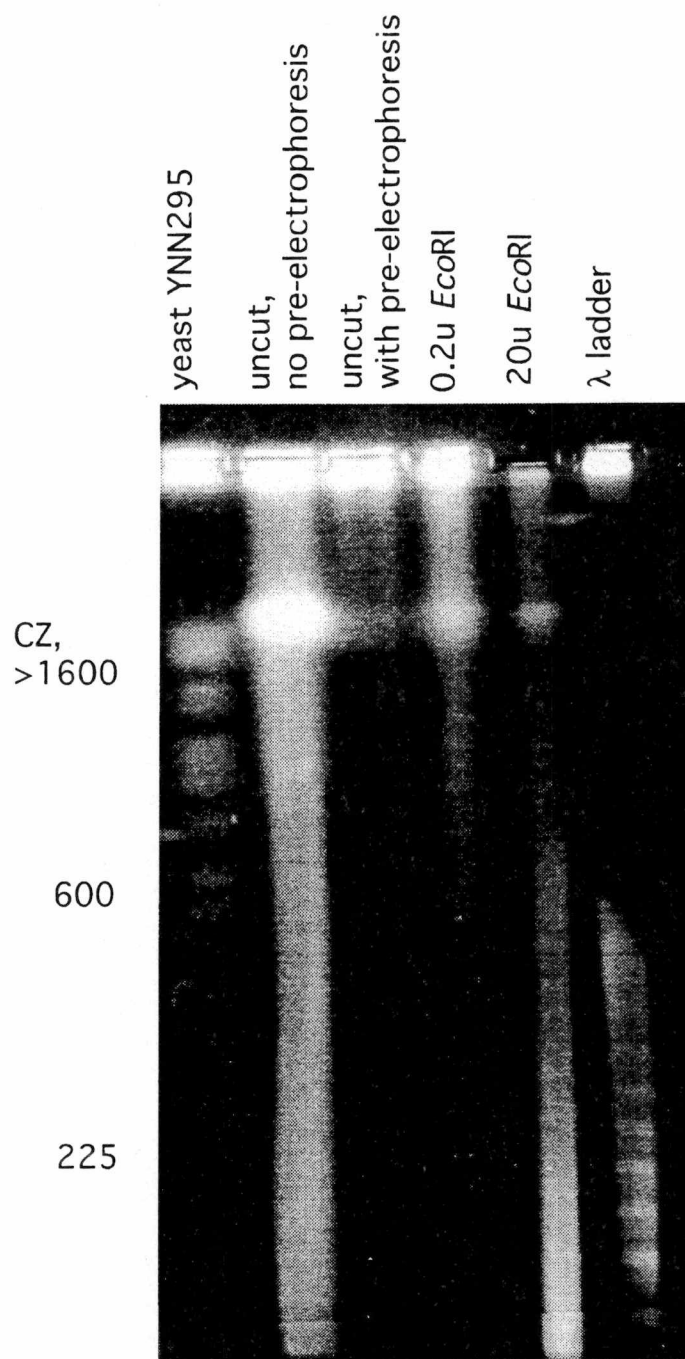
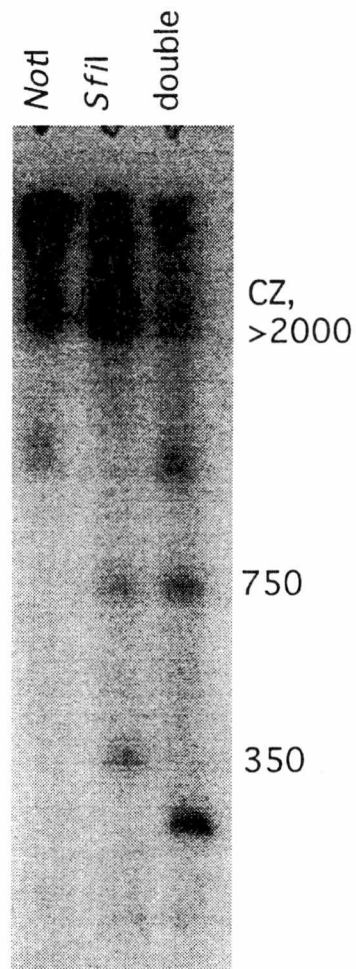
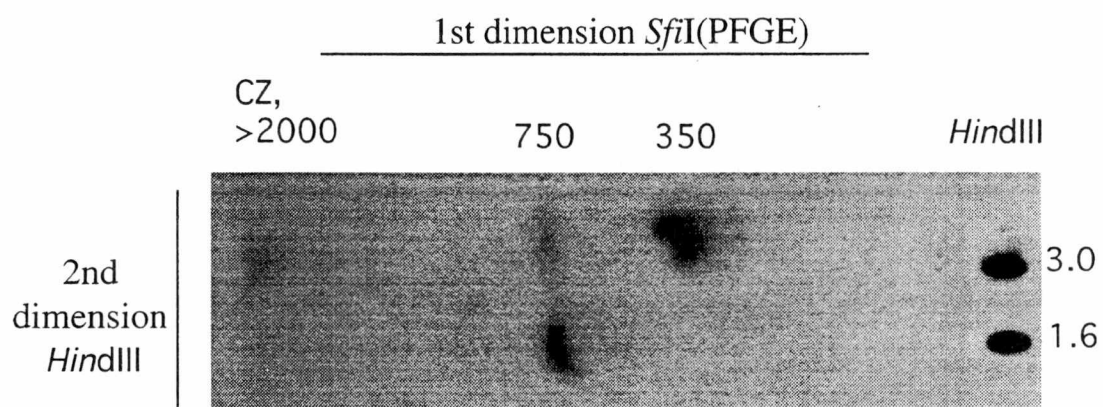


Figure A5. 3. Efficiency of pre-electrophoresis for removal of degraded DNA from plugs, and partial digestion. Without pre-electrophoresis, this batch of plugs contained a considerable amount of degraded DNA which trails along the whole lane. After pre-electrophoresis, this fraction is almost completely removed as demonstrated by the retention of most of the remaining DNA in the well. The sensitivity of the DNA to partial digestion is shown by the fact that as little as 0.2u of *EcoRI* is sufficient to produce visible digestion. Conditions for pre-electrophoresis are described in the text. Conditions for the gel shown were 20-70s switch interval ramped over 20 hours in 0.25 x TBE at

Figure A5.4. Two dimensional electrophoresis of soybean DNA. A soybean molecular marker which has two copies in the genome, one polymorphic and one monomorphic, detected two distinct bands in *Sfi*I digests, as well as hybridizing to the compression zone (a). To assign the monomorphic and polymorphic copies of the marker (corresponding to 3.0 and 1.6 kb *Hind*III fragments, respectively) to the *Sfi*I bands, a second dimension electrophoresis was performed as described in the text (b). The result showed that only the monomorphic copy of the marker resides on a 350 kb *Sfi*I fragment, as well as on a partial product at 750 kb. The polymorphic copy of the marker is situated on a separate 750 kb *Sfi*I fragment.



a. Southern blot of PFGE resolved digests of soybean DNA



b. Southern blot of second dimension electrophoresis

Preparative gels

Preparative PFGE should be run in low-melting point agarose which gives slightly better resolution than regular agarose. If DNA is intended for cloning, precautions should be taken not to stain the bands with ethidium. Instead, side slices of the gel are stained and the location of the desired band is marked by making notches in the gel. If the preparation will be labeled, the whole gel can be stained. Figure A5.5 shows an example of preparative isolation of several YACs containing soybean DNA for fluorescent in situ hybridization (FISH).

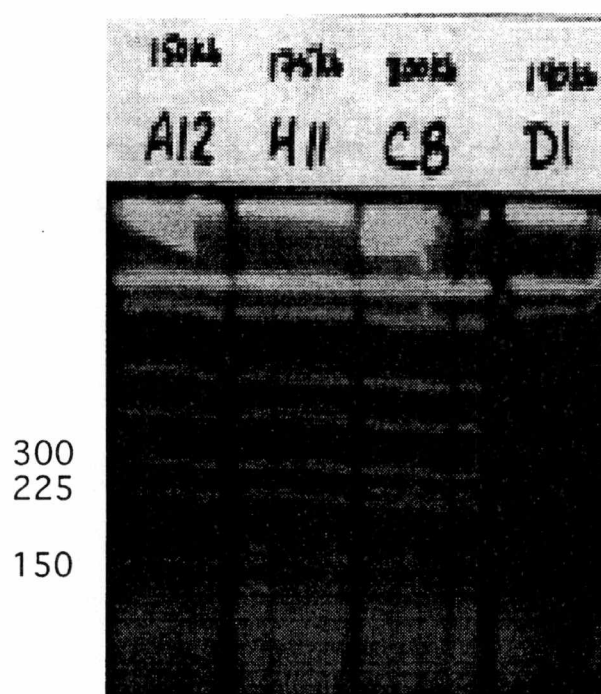


Figure A5.5. Preparative electrophoresis of soybean YACs. YACs were produced and plugs containing chromosomal DNA prepared as described (Appendix 2). 150 ml plugs were loaded on a 1% low melting point agarose gel, and a 40s pulse time was applied for 20 hours in 0.5 x TBE at 14°C. The entire gel was stained in this case because it was not necessary for the DNA to be intact for the application for which it was intended.

VITA

Roel Peter Funke was born in Alkhobar, Saudi Arabia on November 10, 1966. He is a citizen of the Netherlands who attended elementary schools in Penang, Malaysia, Jakarta, Indonesia, and Kobe, Japan. In Singapore, he earned an International Baccalaureate Diploma from the United World College of South East Asia.

He studied Biophysics and Zoology at the University of Leeds in England, and first came to Tennessee in 1987 to do research on muscle function in swimming carp, receiving a Master of Science degree in 1988.

Having earned his Ph.D. in 1994 analyzing and cloning large segments of DNA relevant to systemic communication in soybean, he will be moving from the hills of Tennessee to the plains of Nebraska to concentrate on a single-celled plant, *Chlamydomonas*.