ANISSA LŸBAERT

SEQUENCE AND GENE EXPRESSION VARIABILITY IN CULTIVARS OF OAT (*Avena sativa L.*)

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy

DEPARTMENT OF PLANT SCIENCE
MCGILL UNIVERSITY
MONTREAL, CANADA

JUNE 2006

Your file Votre référence
ISBN: 978-0-494-32211-6
Our file Notre référence
ISBN: 978-0-494-32211-6

Résumé

Ph.D. Anissa Lÿbaert Sciences végétales

L'amélioration génétique des plantes cultivées est compliquée par le fait que de nombreux traits d'intérêt économique sont quantitatifs. Des études récentes indiquent que la variabilité génétique se manifeste souvent par des variations du contrôle de l'expression génique plutôt que par des variations structurales des protéines. Les taux en lipides et protéines du grain sont des caractéristiques importantes de l'avoine cultivée (*Avena sativa* L.*)*. Pour la première des quatre études présentées ici, nous avons obtenu des séquences partielles de huit gènes impliqués dans la biosynthèse lipidique et protéique, provenant de 10 cultivars d'avoine ayant des contenus en lipides et en protéines différents. L'analyse phylogénétique montre un regroupement des séquences en familles représentant hypothétiquement des gènes homéologues. Des polymorphismes de séquences propres aux familles et cultivars ont été identifiés. La deuxième étude présente un échantillonnage de gènes exprimés de façon différentielle entre les cultivars Kanota et Ogle dans le jeune grain. Seule une minorité des 195 contigs obtenus ont des similitudes avec des séquences identifiées, mais le regroupement des séquences en catégories ontologiques montre des profils d'expression différents pour les deux cultivars. La troisième étude teste une méthode de transformation de données de macropuces, consistant à diviser le signal d'un échantillon par le bruit de fond médian de la puce. Cette transformation diminue l'effet des différences d'exposition entre puces. Pour la quatrième étude, nous avons considéré le niveau d'expression de gènes exprimés de façon différentielle entre les parents comme un paramètre quantitatif ségréguant dans la population Kanota x Ogle. La majorité des 33 QTLs d'expression détectés est localisée sur le groupe 29_43, identifiant un point névralgique potentiel de la régulation de l'expression génique chez l'avoine.

Abstract

Ph.D. Anissa Lÿbaert Plant Science

Many traits of economic importance in crop plants are quantitative, complicating the selection for desirable characteristics. Recent studies suggest a complex relationship between genotype and phenotype, with genetic variability often appearing as differences in gene expression rather than structural changes in proteins. In oat (*Avena sativa* L.), lipid and protein content are economically important traits. In the first of four studies reported here, partial sequences for eight genes involved in lipid or protein biosynthesis were obtained from ten oat cultivars with varying lipid and protein content. Phylogenetic analysis showed that these sequences clustered into families possibly corresponding to homeologous genes. Some cultivar- and family-specific polymorphisms were identified. In the second study, we surveyed differential gene expression between developing kernels of cultivars Kanota and Ogle by constructing reciprocal subtractive libraries. Of the 195 contig sequences obtained, only a minority had homology to characterized sequences. Grouping these sequences in categories based on gene ontology of their BLAST hits showed different profiles of expression for each cultivar. In the third study, we tested a method for transforming macroarray data consisting of dividing spot signal by the median array background. This reduced variation due to array exposure time. In the fourth study, gene expression levels were considered as quantitative traits in the Kanota x Ogle mapping population. Macroarrays featuring oat clones differentially expressed between both parents were hybridized with cDNA from the population lines. Among the 33 significant expression quantitative trait loci detected, most clustered to linkage group 29_43, a possible "hot-spot" of gene expression regulation.

Acknowledgements

I would like to thank Dr. Diane Mather who supervised me throughout this Ph.D. I have greatly benefited from her advice and guidance, and have appreciated the promptness with which she has provided logistical and financial support throughout this study. I am very grateful for the trust she put in me, giving me the freedom to explore new directions in my research.

My thanks go to Dr. Stephen Molnar, who co-supervised my Ph.D. and welcomed me in the oat genomics lab at ECORC, giving me access to excellent laboratory settings and equipment. I have appreciated his moral support and his help in career advice. I have also benefited financially and in knowledge from the research assistant contracts he provided.

I am very thankful to Dr. Nicholas Tinker for being a member of my committee and contributing so much of his time. I have benefited from his constructive advice, and for his patience during our many brainstorming sessions when my mind was racing to catch up with his train of thought. I am also thankful to Dr. Tinker for encouraging me to start a Ph.D. and introducing me to Dr. Mather. I never regretted having followed his suggestion.

I am also grateful for the time Drs. Mather, Molnar and Tinker have spent thoroughly reviewing my thesis and the speed with which I have received their feedback and answers to my many questions.

I wish to express my sincere appreciations to the other members of my committee, Drs. Robin Beech and Donald Smith for their constructive advice and patient guidance throughout my study.

I am grateful for the services provided by the Department of Plant Science administrative staff, most particularly Carolyn Bowes, Roslyn James and Louise Mineau.

I would like to thank Dr. Howard Rines, for providing phenotypic data from the Kanota x Ogle population and allowing me to use this data in my research.

I wish to thank Nicholas Cheff and Dulce Reyes for their excellent assistance in the laboratory, and Kathy Upton for taking good care of my plants. Without their help, the scale of the experiments presented in this thesis would have had to be reduced. My thanks

also go to Drs. Martina Stromvik, Judith Frégeau-Reid, Laurian Robert and Jas Singh for giving me access to their laboratories and some essential equipment.

I wish to thank Philippe Couroux for his help; his patience and good humour were precious when my sequence data seemed to highlight every possible bug in the programs I used. Very special thanks go to Jean Gerster for her help managing my data, and particularly for the scripts she wrote that allowed me to sleep blissfully while my data was being processed. Thanks also to Hai Pham for introducing me to some tools and tricks that made my life easier.

I am indebted to Ghislaine Allard, Charlene Wight, Titus Tao, Helene Rocheleau, Martin Charette, as well as Drs. Johan Schernthaner, Therese Ouellette and Gopal Subramaniam for useful discussions. Their input has been precious and helped me solve a few lab mysteries.

As the past five years have involved many overnight stays in Montreal and Ste-Anne-de-Bellevue, I want to express my gratitude to Yella and Mooshie, Diane and Rob, Steve Allen and Louise Cournoyer who have repeatedly offered me hospitality and very pleasant company. I keep fond memories of those campus trips.

Warm thanks go to the oat research group and the Bioproducts, Bioprocesses and Bioinformatics division at ECORC, as well as to the members of Dr. Mather's lab at Macdonald, for the joyful and dynamic working atmosphere and the pleasant social agenda of parties, hallway conferences, thunderous coffee breaks and Thursday beers.

Special thanks go to Mike Burvill and Dorothy Sibbitt for their precious friendship and support. Many people have helped me during these demanding years, and I can not name them all, but I would like to extend my thanks to all for their kindness and encouragements.

My love and gratitude go to my partner Luc Pelletier who has been patient and caring, and for bearing my moments of frustration. I thank him also for reminding me now and again that there was a life away from my pipettes and computer.

Finally, I would like to acknowledge the Natural Sciences and Engineering Research Council of Canada, Quaker Foods and Beverages and Quaker Tropicana Gatorade of Canada for their support of this research, and the Frederick Dimmock Memorial Fellowship for providing financial assistance.

Aan mijn grootmoeder Rosanna,
die graag naar school ging.
Met veel liefs, altijd.

Table of Contents

List of Tables

List of Figures

Glossary

Allele: "One of the variant forms of a gene at a particular locus on a chromosome. Different alleles produce variation in inherited characteristics [...]. When "genes" are considered simply as segments of a nucleotide sequence, allele refers to each of the possible alternative nucleotides at a specific position in the sequence". (The NCBI Handbook Glossary, 2006)

Anthesis: "The phase of a flower when pollen is presented and/or the stigma is receptive" or " The stage at which any flower(s) on the plant are open" (Plant Ontology Consortium, http://dev.plantontology.org/docs/growth/growth.html)

BLAST: "Basic Local Alignment Search Tool (Altschul et al., 1990). A sequence comparison algorithm that is optimized for speed and used to search sequence databases for optimal local alignments to a query". (The NCBI Handbook Glossary, 2006)

Contig: "A contiguous segment of the genome made by joining overlapping clones or sequences. A clone contig consists of a group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. A sequence contig is an extended sequence created by merging primary sequences that overlap. A contig map shows the regions of a chromosome where contiguous DNA segments overlap. Contig maps provide the ability to study a complete and often large segment of the genome by examining a series of overlapping clones, which then provide an unbroken succession of information about that region". (The NCBI Handbook Glossary, 2006)

Cultivar: "A variety of plant produced through selective breeding by humans and maintained by cultivation". (Paran and Zamir, 2003)

EST: "Expressed Sequence Tag. ESTs are short (usually approximately 300–500 base pairs), single-pass sequence reads from cDNA. Typically, they are produced in large batches. They represent the genes expressed in a given tissue and/or at a given developmental stage. They are tags (some coding, others not) of expression for a given cDNA library. They are useful in identifying full-length genes and in mapping". (The NCBI Handbook Glossary, 2006)

Gene: "The functional and physical unit of heredity passed from parent to offspring. Genes are pieces of DNA, and most genes contain the information for making a specific protein". (National Human Genome Research Institute, 2006)

Gene ontology: "A controlled vocabulary used to describe the biology of a gene product in any organism. There are 3 independent sets of vocabularies, or ontologies, that describe the molecular function of a gene product, the biological process in which the gene product participates, and the cellular component where the gene product can be found". (Hyperdictionnary, 2006)

Genetic map: "(Also known as a linkage map) A chromosome map of a species that shows the position of its known genes and/or markers relative to each other, rather than as specific physical points on each chromosome". (National Human Genome Research Institute, 2006)

Genome: All the DNA contained in an organism or a cell, which includes both the chromosomes within the nucleus and the DNA in organelles. (modified from National Human Genome Research Institute, 2006)

Genomics: "The study of genes and their function. Genomics aims to understand the structure of the genome, including the mapping genes and sequencing the DNA. Genomics examines the molecular mechanisms and the interplay of genetic and environmental factors [...]. Genomics includes:

Functional genomics: the characterization of genes and their mRNA and protein products.
Structural genomics: the dissection of the architectural features of genes and chromosomes.
Comparative genomics: the evolutionary relationships between the genes and proteins of different species.
Epigenomics (epigenetics): DNA methylation patterns, imprinting and DNA packaging". (MedicineNet, 2006)

Genotype: "The genetic identity of an individual that does not show as outward characteristics. The genotype refers to the pair of alleles for a given region of the genome that an individual carries". (The NCBI Handbook Glossary, 2006)

Indel: "Insertion/deletion event. Some sequences (DNA or amino acid) do not just evolve by point mutation process, but also by expansion and/or contraction of the length of the sequence". (Molecular Systematics and Evolution, 2006)

Metabolome: "All native metabolites, or small molecules, that are participants in general metabolic reactions and that are required for the maintenance, growth and normal function of a cell". (Hyperdictionnary, 2006)

Phenotype: "The observable traits or characteristics of an organism [...]. Phenotypic traits are not necessarily genetic". (National Human Genome Research Institute, 2006)

Proteome: "The complete set of proteins expressed and modified following their expression by the genome. The analysis of the proteome is proteomics". (MedicineNet, 2006)

QTL: Quantitative Trait Locus.

"A locus at which segregation contributes to the variation of a quantitative character." (Lynch and Walsh, 1997)

Recombinant inbred lines: A population of homozygous individuals that is obtained by repeated selfing from an F_1 hybrid.

SNP: Single Nucleotide Polymorphism. "A SNP is a single base-pair site within the genome at which more than one of the four possible base pairs is commonly found in natural populations". (The NCBI Handbook Glossary, 2006)

Trait: "In genetics, a trait refers to any genetically determined characteristic". (MedicineNet, 2006)

Transcriptome: "The complete set of RNA transcripts produced by the genome at any one time. The transcriptome is dynamic and changes under different circumstances due to different patterns of gene expression. The study of the transcriptome is termed transcriptomics". (MedicineNet, 2006)

Contribution of coauthors

This thesis has been written in the manuscript-based format. Chapters 3 to 6 have been modified from manuscripts prepared for publication.

For Chapter Three, I selected the target genes, analyzed the template sequences, designed the primers and tested them in the laboratory, cloned the PCR products, carried the phylogenetic analysis and prepared the manuscript. For Chapter Four, I designed the experiment, grew the plants, collected the material, performed the subtractive suppressive hybridization and built the libraries, analysed the sequences and prepared the manuscript. For Chapter Five, I designed the experiment, performed the laboratory work, the data analysis and prepared the manuscript. For Chapter Six, I designed the experiment, grew the plants, collected the material, designed the macroarrays, carried out the hybridizations and the image analysis, analysed the data and prepared the manuscript.

Dr. Diane Mather contributed to the research and all the manuscripts by suggesting research directions, providing laboratory supplies, funding and supervision, and giving guidance and careful corrections of the manuscripts.

Dr. Stephen Molnar is a coauthor of all the manuscripts and his contributions have involved providing laboratory settings and equipment, providing constructive suggestions and by giving detailed guidance and corrections during the writing of the manuscripts.

Dr. Nicholas Tinker contributed to the research and all the manuscripts by suggesting research directions, solving problems, providing access to bioinformatic tools and facilities, providing critical comments and corrections for the manuscripts.

Chapter 1
General Introduction

1.1 Background

Most agronomic traits in crop plants are quantitative, exhibiting continuous variation, which makes their inheritance and characteristics difficult to study. Underlying phenotypic variation are genetic and environmental factors interacting via complex networks to define the morphology, behaviour and fitness under selection of an individual. Despite significant progress in knowledge of the physiology, biochemistry and molecular biology of plants and animals, large gaps remain in our understanding of the path between genotype and phenotype.

Phenotypic diversity can result from differences in the sequence of genes that can alter protein function, or from variations in regulatory elements influencing gene expression. The latter variations include differences in sequences, but also changes in DNA methylation or chromatin structure. Two types of loci involved in the control of gene expression can be distinguished. The first resides in the vicinity of the gene it regulates (*cis* regulation), and can be located in the promoter of the gene or in regions of the gene affecting the stability of its transcripts. The second type is located away from its target gene (*trans* regulation) and is more difficult to study without prior information about its location and/or effects.

Molecular markers and the constitution of genetic maps have made it possible to detect genomic regions affecting quantitative traits, called quantitative trait loci (QTLs). Each of these loci presumably contains one or more undefined genes or regulatory regions affecting the trait. A limited number of QTLs with large effects explaining most of the variation for a particular trait are often detected in QTL studies (Prioul et al., 1997). Although these QTLs may explain a large portion of the phenotypic variance that is observed in a given environment, there are probably many additional QTLs that are not detected because their effects are small, or because they affect genetic changes that are not critical in the environment(s) where data is collected.

Several recent studies have examined variations in the transcriptome as quantitative traits, and loci responsible for such variations, named eQTLs, have been mapped in different

organisms (Wayne and McIntyre 2002; Schadt et al., 2003; Yvert et al., 2003; Bystrykh et al., 2005; Chesler et al., 2005; DeCook et al., 2005; Hubner et al., 2005). This type of approach allows the detection of both *cis* and *trans* regulatory factors, as long as they are segregating in the population studied. Recent results suggest that the regulation of complex traits is mainly driven by factors acting in *trans* on the primary physiological pathways involved (Chagnon et al., 2003; Pomp et al., 2004).

Among the important end-use quality characteristics of oat (*Avena sativa* L.) are lipid and protein content, both of which exhibit quantitative variation. Metabolic pathways for lipid and protein biosynthesis in plants have been intensely studied (Larkins et al., 1982; Ohlrogge and Browse, 1995; Ohlrogge and Jaworski, 1997; Voelker and Kinney, 2001; Hills, 2004), and several steps in these pathways have been identified as potentially limiting. Identifying the number, location and impact on the phenotype of genes, or at least genomic region, involved in the trait could greatly facilitate the selection, assisted by molecular markers, of such traits of economic interest.

1.2 Hypotheses

The following hypotheses were formulated in order to study the genetic variation underlying different known phenotypic characters of oat cultivars:

- Candidate genes coding for products involved in lipid and protein biosynthesis exhibit sequence variation among oat cultivars differing in their lipid and/or protein content, and further, that this sequence variation would be sufficient to support the development of gene based markers for oat.
- Loci in the oat genome that influence the level of expression of genes (eQTLs) can be identified, and some of these loci may co-locate with QTLs for known phenotypic characteristics of oat. Identifying the genes whose transcription is affected by loci where QTLs and eQTLs co-locate could lead to hypotheses about genetic mechanisms affecting quantitative traits.

1.3 Objectives

The first objective of the work presented here was to obtain oat sequences for selected genes implicated in lipid and protein biosynthesis from 10 oat cultivars contrasting in lipid and protein content. The degree of sequence variability would then be

assessed from the analysis of these sequences. This information would then be used to design gene-based markers.

A second objective was to isolate, sequence and identify genes differentially expressed between young kernels of Kanota and Ogle.

A third objective was to develop a protocol to survey differences in the expression of selected genes across the Kanota and Ogle genetic population.

A fourth objective was to detect and map gene expression-level polymorphisms in the Kanota x Ogle population, and compare their map locations with known QTLs for oat traits

Chapter 2
Literature Review

2.1 The genetics of phenotypic variation

2.1.1 Genotype *versus* phenotype

The study of phenotypic variation is central to many disciplines involving living organisms, from evolutionary biology to medicine and agriculture. Phenotype is the level at which natural or human-driven selection acts, and the response to selection depends on the heritable components of phenotype present in a population. Phenotype being the result of genotype, environment and their interaction, the existence of genetic diversity in a population is essential to maintain the potential for some individuals to adapt to a changing environment. The relationships between genotype, phenotype and environment have interested geneticists for many years (Schmalhausen, 1949; Wolff, 1955; Kidwell, 1963), and have often sparked controversy when dealing with human traits such as intelligence or aggressiveness (Scarr-Salapatek, 1971; Selmanoff et al., 1975). Traits such as body weight in animals and yield in crop plants have been intensely studied, but we are still far from being able to identify the genetic variability underlying those traits.

2.1.2 Resources and strategies to study the impact of genotype on phenotype

Our understanding of genotype/phenotype interactions remains partial, complicated by the fact that genetic diversity does not always have an impact on phenotype, and when it does, it is often expressed in quantitative rather than qualitative variation. Several new resources are now available and research strategies are being implemented that should increase our knowledge of the impacts of genetic variation.

2.1.2.1 Available resources

Sequence databases

The GenBank sequence database (http://www.psc.edu/general/software/packages/genbank/genbank.html), an annotated

collection of all publicly available DNA sequences, has grown 400% since 2002 and contains 52,016,762 sequences composed of 56,037,734,462 base pairs (March 7, 2006). This database also includes genome sequences, complete or in progress, of 1366 organisms, including 30 mammals, and 30 land plants (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html, January 2007). Among the whole genome sequences available or in progress, some belong to closely related organisms. The genome of rice (*Oryza sativa* L.) cultivar group *japonica* is available, as will soon be that of cultivar group *indica*. Similarly, the genome of *Arabidopsis lyrata* (L.), a close relative of *Arabidopsis thaliana* (L.) is currently being sequenced (DOE Joint Genome Institute, http://www.jgi.doe.gov/sequencing/why/CSP2006/AlyrataCrubella.html; Arabidopsis Genome Initiative, http://www.arabidopsis.org/info/agi.jsp). Those projects will provide information on intra-genus sequence variability that will complement genetic mapping studies (McCouch and Doerge, 1995; Yoshimura et al., 1997; Kuittinen et al., 2004) and phenotypic data.

Phenotype databases

Many phenotype databases are available, collecting data from projects involving reverse genetics or mutation studies, mainly organism-specific databases for well-studied organisms such as *Homo sapiens*, *Caenorhabditis elegans*, Drosophila (*Drosophila melanogaster*), mouse (*Mus musculus*) and Arabidopsis. Several of these have now been merged into PhenomicDB (http://www.phenomicdb.de/), a multi-organism phenotype-genotype database. This database also includes gene information from NCBI (http://www.ncbi.nlm.nih.gov/) and orthologous gene information from HomoloGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene), which allows the comparison of phenotypes associated with a gene over many organisms.

Mutant collections and databases

Mutant collection databases are available for model organisms, but are also becoming more common for crop plants. The Maize Gene Discovery Project (http://www.maizegdb.org/rescuemu-phenotype.php) made available a collection of

maize (*Zea mays* L.) lines mutated by modified transposon RescueMu as well as information on the sequences flanking the site of insertion. Collections of barley (*Hordeum vulgare* L.) (http://germinate.scri.sari.ac.uk/barley/mutants/index.php?option=com_frontpage&Itemid=1) and rice mutants (http://www.iris.irri.org/IRFGC/ir64.jsp) are also available.

2.1.2.2 Research strategies

Some leading approaches for investigating the genetic sources of phenotypic variation are mentioned here, and are listed and discussed in several recent review articles (Cheung and Spielman, 2002; Paran and Zamir, 2003; Vasemagi and Primmer, 2005).

"Bottom-up" strategies involve looking for meaningful relationships between allelic information, and the phenotypes of individuals contrasting for the trait of interest. They require prior hypothesis about the involvement of certain alleles in the phenotype. This category includes forward and reverse genetic approaches, comparative studies of sequence information, and gene and protein expression studies. A second category of strategies can be labelled "top-down": starting from phenotypic variation in a population or group of individuals, it aims at establishing the genetic basis of this variation.

Reverse genetics

Reverse genetics aims at mutating or silencing known genes in order to study the resulting phenotypes. Gene knockout and silencing have been used with success (Galli-Taliadoros et al., 1995), but these methods require at least partial sequence information for the target gene. A technique applicable to organisms that lack well-developed genetic tools is Targeting Induced Local Lesions IN Genomes (TILLING). It consists of using chemical mutagenesis followed by screening for single-nucleotide changes to discover induced mutations that alter protein function (McCallum et al., 2000). High-throughput TILLING has produced hundreds of point-mutated genes in Arabidopsis (Till et al., 2003), and is now being applied to crop plants such as maize and wheat (*Triticum aestivum* L.) (Slade and Knauf, 2005).

QTL mapping

QTL mapping requires the existence of a mapping population derived from a suitable cross, a genetic linkage map as well as phenotypic data for the traits of interest across the population. It allows the identification of regions in the genome influencing the trait of interest within the population, defined by the association between the genotype at every marker of the map and the value of the trait studied (Lander and Botstein, 1989; Zeng, 1993 and 1994). QTL studies have been performed in many species, and in some case the underlying genes have been identified. For example, genes responsible for variation in flowering time have been isolated in *Arabidopsis thaliana*, maize and rice, and genes for fruit shape and weight cloned in tomato (*Lycopersicon esculentum* Mill.) (El-Din-El-Assal et al., 2001; Thornsberry et al., 2001; Takahashi et al., 2001; Kojima et al., 2002; Frary et al., 2000; Liu et al., 2002). QTL studies conducted across several environments can also locate regions responsible for interactions between the genotype and the environment (Jansen, 1992; Jansen et al., 1995; Tinker and Mather, 1995).

Gene expression polymorphisms

Gene expression polymorphism is a new approach in which mRNA abundance is treated as a quantitative trait. QTL mapping strategies are used to identify particular regions of the genome that are associated with variations in gene expression in the population studied. It has been conducted on a few candidate genes at a time using Northern hybridization (Consoli et al., 2002), but also using microarrays to investigate thousands of genes simultaneously in Drosophila, baker's yeast (*Saccharomyces cerevisiae*), mouse, human, rat (*Rattus norvegicus*), maize, and Arabidopsis (Wayne and McIntyre 2002; Schadt et al., 2003; Yvert et al., 2003; Bystrykh et al., 2005; Chesler et al., 2005; DeCook et al., 2005; Hubner et al., 2005). Results have shown that mapping of loci affecting transcript abundance can be a successful strategy for dissecting complex traits and identifying candidate genes. Screening microarrays of mouse and maize genes, Schadt et al. (2003) identified thousands of significant expression QTLs (eQTLs), and reported some genomic regions where eQTLs seemed to cluster into what appear to be expression regulation hot-spots. Similar clustering was reported by DeCook et al. (2005) in *Arabidopsis thaliana*.

Linkage disequilibrium-based association mapping

Linkage disequilibrium-based association mapping is derived from the analysis of linkage disequilibrium, the non-random association between two alleles at two or more loci (Gupta et al., 2005). Such loci may be represented by genetic markers, known candidate genes, QTL, or any combination thereof. Association mapping tests whether a certain genotype is significantly associated with the trait of interest within populations or groups of individuals. Unlike QTL mapping, it does not require random samples of progeny and has therefore been extensively used for studying complex traits in mammals, and particularly hereditary human diseases such as cardiovascular diseases (Crawford et al., 2005; Morton 2005). Association mapping has been successfully used in plants (Kraakman et al., 2004; Grattapaglia, 2004; Ingvarsson, 2005), and is an attractive alternative to QTL mapping in natural populations or species, such as tree species, for which the development of mapping population is challenging.

2.1.3 A complex relationship between genotype and phenotype

A current hypothesis to explain the relationship between genotype and traits is that many genes affecting complex traits may be regulatory factors involved in *trans* regulation of physiological pathways rather than structural genes (Pomp et al., 2004). This might in part explain why many QTLs seem specific to particular genetic backgrounds, or are not stable across environments (Quarrie et al., 2005).

An example of the complex interactions between the genome and the environment is the phenomenon called "canalization". It involves the repression of the expression of genetic variations in a population in favour of a "robust", relatively invariant phenotype (Flatt, 2005). The cryptic genetic variability accumulated in a population can be released after "decanalization" events, such as a drastic change in environment. "Canalization" is another indication that genotype might not exert direct control on phenotype, but that complex regulation processes, involving responses to the environment, are in place to control how much of the genotype will contribute to the phenotype.

Furthermore, it has been shown that the amount of noise occurring in gene expression is not always negligible, and that stochastic events might also have a significant influence on the phenotype. Noise in gene expression can be divided into

"intrinsic" noise, and "extrinsic" noise (Swain et al., 2002; Raser and O'Shea, 2005). The first category, caused by the random nature of biochemical processes inside the cell, would affect in different ways two genes in the same cell. The second category, caused by either differences in the internal states of cells in the tissue or population studied or small environmental changes, would affect several genes inside a cell equally. In prokaryotes, the dominant source of noise in protein levels seems to be the transcription system (Elowitz et al., 2002; Ozdubak et al., 2002). A study by Raser and O'Shea (2004) in baker's yeast indicates that noise in gene expression seems to be predominantly gene specific. In the case of one particular yeast gene, the source of noise was identified as being due to random chromatin remodelling causing the interconversion of a promoter between active and inactive states. Several mechanisms seem to be in place in living organisms to minimize the impact of noise in gene expression, one example of which being a high gene copy number (Kellis et al., 2004). Raser and O'Shea (2005) formulate the hypothesis that those mechanisms can be regulated, particularly in case of stress, to produce more phenotypic diversity.

Results of the research on the interactions between genotype and phenotype show a complex relationship (Fig. 2.1) that combines the effects of variations in DNA sequence, the regulation of transcription and translation, as well as stochastic effects and interactions between the genotype and the environment.

2.2 Oat as a crop

2.2.1 The genus *Avena*

The genus *Avena* is a small genus of about 27 species, growing primarily in temperate areas. It was domesticated over 2000 years ago around the Mediterranean, but is now cultivated all around the world. North American taxa have been introduced from Europe and the Mediterranean area. One species (*A. sativa* ssp *sativa* L.) is an important cereal crop, but some *A. sativa* ssp *byzantina* C. Koch. is also cultivated. Other species are mainly weeds, including *Avena fatua* (L.), a major disturbance species around the world. The plants are annuals and herbaceous, with hermaphrodite florets that are mainly self-pollinating. The basic chromosome number of the genus is x=7, and diploid, tetraploid and hexaploid species exist.

2.2.2 The crop plant

Although originating in the Middle East (Baum, 1977), oat has adapted well to cold climates, where most of it is produced today. Winter and spring varieties of oat exist and they are grown as a multipurpose crop, including livestock feed, human food and industrial applications such as cosmetics (Schrickel 1986). Oat straw is used for livestock bedding, and oat is also grown as a forage crop, particularly in the Southern Hemisphere and in the United States. High yield and resistance to disease (mainly rust, barley yellow dwarf virus and smut) have been the main breeding objectives for decades. Other breeding objectives include:

- Winter hardiness
- Reduced lodging
- Good germination, no pre-harvest sprouting
- Appropriate maturity to avoid frost or drought and maximize kernel filling
- Hullessness ("naked oat")
- Less trichomes on naked oat varieties
- Groat percent, kernel shape and ease of dehulling
- Kernel quality traits: high or low lipid content, high protein, high or low ß-glucans, and reduced kernel damage.

2.2.3 Chemical constitution of the oat grain

2.2.3.1 Lipids

Oat lipids, like other cereal lipids, are nutritionally important because of their high level of unsaturation, and their abundance of linoleic acid (18:2) (Frey and Hammond, 1975; Schipper and Frey, 1991). In mammals, this fatty acid is a precursor in the synthesis of prostaglandins, a class of compounds implicated in the action of several hormones, the contraction of smooth muscles and in inflammatory reactions. With a range of 2 to 16.2% of the total groat weight oat has also the highest lipid content among cereals. Lipid content is highest in the embryo, but because the embryo is a relatively

small component of the oat seed, over 90% of the kernel lipids are located in the starchy endosperm and the bran.

Storage lipids in plants consist mainly of triglycerides. Depending on the method of extraction, the estimation of the triacylglycerol content in total oat lipids varies from 41 to 80% (Youngs, 1986). Similarly, the estimates of the proportion of other lipids are variable: phospholipids are estimated to account for 5.6 to 24.6% of the total lipids, free fatty acids 4.8 to 12.2%, and sterol esters 0.3 to 15.8%.

Palmitic, oleic and linoleic acid constitute 95% of free and esterified fatty acids in the oat grain. As has been determined in maize (de la Roche et al., 1971), triacylglycerol tends to carry a higher proportion of unsaturated fatty acids on its second carbon group (sn2), and contains more oleic acid and less palmitic acid than do phospholipids or glycolipids. Therefore, if the total lipid content of the grain increases, the proportion of oleic acid also increases, which means that at higher lipid content, the level of unsaturation stays high (Thro et al., 1985).

The heritability of groat oil content is high, ranging from 63% to 93% (Schipper and Frey, 1991). Lipid content responds therefore well to selection towards both high and low oil phenotypes.

2.2.3.2 Proteins

The high protein content of oat and its good amino acid balance make it nutritionally superior to other cereal grains. The proportion of threonine, methionine and particularly of lysine, three amino acids considered nutritionally limiting, tends to be higher in oat than in other cereals (Peterson and Brinegar, 1986).

Proportionally, the embryonic axis has the highest protein content (up to 44% protein), but since the embryo is such a small portion of the kernel, the biggest contributors are the bran and endosperm, adding up to about 94% of the total groat protein (Youngs, 1972).

Storage proteins are localized mainly in protein bodies, developing when proteins are deposited in cell vacuoles. These organelles have single layer membranes and the proteins are stored around a globoid of phytin, a phosphate storage substance of seeds.

During germination, hydrolytic enzymes from the aleurone degrade the substrate molecules in the endosperm, mobilizing resources for the developing embryo. There is a

peak of protease activity one or two days after imbibition, paralleled by a disappearance of endosperm proteins. The amino acid residues liberated are remobilized to synthesise proteins for the development of the embryo (Peterson and Brinegar, 1986).

Estimations of the heritability of protein content vary among studies from 0.09 to 0.90, with an average of about 0.41 (Frey, 1975). It is nevertheless believed that environment factors such as soil nitrogen content have a significant effect on this trait (Peterson and Brinegar, 1986).

2.2.3.3 Other important components of the oat grain

Starches

Starches constitute 43 to 61% of the groat weight. Their amylose content is around 23%, similar to the level observed in wheat and maize (Peterson, 1992). Pasting properties of oat flours are variable, sometimes totally inappropriate for baking, because they have high shearing susceptibility and high viscosity. The inheritance of these characteristics has not been studied.

β-glucans

β-glucans, which are unbranched polysaccharides of ß-D-glucopyranosyl, are important components of cell walls. Oat bran contains up to 7.5% of ß-glucans, and the groat 4% (Peterson, 1992). Soluble ß-glucans have potential interesting health benefits for humans, including the reduction of serum cholesterol (Ripsin et al., 1992; Karmally et al., 2005; Robitaille et al., 2005) and a decrease of glycemic response after meals (Behall et al., 2005; Tapola et al., 2005).

2.2.4 Oat genomics

2.2.4.1 Cytogenetics

Oat has a large genome, similar in size to that of wheat. One haploid genome of oat is estimated to be about 11,315 Mb (Arumuganathan and Earle, 1991), which is about 100 times larger than the *Arabidopsis thaliana* genome. The *Avena* genus is divided in

five cytogenetic groups, according to the origin and the ploidy of the genome of the different species (Table 2.1).

2.2.4.2 Genetic maps

The first map of oat was developed using a population derived from a cross between diploid oat taxa *Avena atlantica* (B. R. Baum & Fedak) and *Avena hirtula* (Lag.), using RFLP markers (O'Donoughue et al., 1992). Since then, maps have been established for several other populations, including another diploid cross and five hexaploid crosses of cultivated oat (O'Donoughue et al., 1992; Kianian et al., 1999; Wight et al., 2003; Zhu et al., 2003a; De Koeyer et al., 2004; Portyanko et al., 2005; Rayapati et al., 2006).

No consensus map for *Avena* is yet available, and the number of linkage groups in hexaploid oat maps is still larger than the number of chromosomes of the haploid genome. Although many markers have been positioned, particularly on the Kanota x Ogle map, the spacing of these markers is uneven, and clustering occurs in certain regions (Wight et al., 2003).

QTL studies have been conducted in the populations Kanota x Ogle (Kianian et al., 1999; Groh et al., 2001), Kanota x Marion (Kianian et al., 1999 and 2000, Groh et al., 2001), Terra x Marion (De Koeyer et al., 2004), Ogle x TAM O-301 (Holland et al., 1997 and 2002) and Ogle x MAM 17-5 (Zhu and Kaeppler, 2003b; Zhu et al., 2003c; Zhu et al., 2003d; Zhu et al., 2004) for several agronomic traits. Unfortunately, the absence of a consensus map and the scarcity of common markers between the different maps often limit the study of a particular QTL across the different populations.

Relatively few named genes have been mapped in oat. The literature reports the map position of genes for an ACCase (Kianian et al., 1999), a ß-glucanase (Yun et al., 1993), three avenins, two globulins (Shotwell et al., 1990; O'Donoughue et al., 1995), 22 α-amylases (Sharopova et al., 1998), three esterases, one isocitrate dehydrogenase and one peroxidase, a 6-phosphogluconate dehydrogenase, a phosphoglucomutase, a shikimate dehydrogenase (O'Donoughue et al., 1995) and 8 resistance gene analogues (Cheng and Armstrong, 2002). Many of the above were mapped using isozyme techniques, which may indicate functional differences in gene products.

Since many of the clones that were used to map RFLP loci have now been partially sequenced, we can speculate on the positions of additional gene loci. Recent oat

genomic projects have generated thousands of oat ESTs (Brautigam et al., 2005; Howard Rines, unpublished), increasing the number of oat sequences that could be positioned on genetic maps by direct or comparative mapping. However, most of the speculation on the function of these genes is based on incomplete homology with genes that are characterized in other species.

2.3 Lipid and protein biosynthesis in plants

2.3.1 Lipid biosynthesis

Detailed reviews of plant lipid biosynthesis and its regulation are available (Ohlrogge and Browse, 1995; Ohlrogge and Jaworski, 1997; Voelker and Kinney, 2001). A general account of major steps relevant to this thesis is presented below and in Figure 2.2.

2.3.1.1 Fatty acid biosynthesis

The fatty acid biosynthesis pathway is a primary metabolic pathway, and inhibitors of this pathway are lethal to the cell. The expression of the genes coding for the proteins and enzymes involved in this pathway must be closely controlled. This regulation might involve phosphorylation (Savage and Ohlrogge, 1999) and control of transcription (Elborough et al., 1994; Fawcett et al., 1994, Slabas et al., 2002).

Fatty acids are synthesised in the plastid from carbons deriving from plastidial acetyl-CoA. It is still unknown what the source of this acetyl-CoA is, but it could be generated by numerous potential pathways, including as a by-product of glycolysis or the action of Rubisco (Andrews and Kane, 1991).

The first committed step of fatty acid synthesis is catalyzed by the enzyme acetyl-CoA carboxylase, producing malonyl-CoA by carboxylation of acetyl-CoA. The malonyl group is then transferred from the Coenzyme-A group to a cofactor, the acyl carrier protein (ACP) and enters the "condensation cycle", in which the first condensation enzyme, KASIII (3-ketoacyl-ACP synthase III), adds an acetyl group from acetyl-CoA to the malonyl group, forming a C4 chain. The second condensation enzyme, KASI, further extends the carbon chain to C16 (palmitoyl-ACP), and a third enzyme, KASII, adds the two last carbons to produce C18 chains (stearoyl-ACP). The first desaturation of the chain

occurs in the plastid to form oleoyl-ACP. This reaction is catalyzed by a stearoyl-ACP desaturase. The ACP cofactor is then cleaved by two acyl-ACP thioesterases, and the free fatty acids leave the plastid by a mechanism that has not been identified yet.

2.3.1.2 Triacylglycerol biosynthesis

In the cytoplasm, fatty acids are esterified again to CoA and, once in the endoplasmic reticulum, are transferred on the glycerol backbone. The enzyme attaching a fatty acid on glycerol-3-phosphate position sn1 is a membrane-bound glycerol-3-phosphate acyltransferase (GPAT). The second fatty acid chain, on position sn2, is attached by a lysophosphatidic acid acyltransferase (LPAAT), and enzyme that has unsaturated acyl chains as preferred substrate.

Two different pathways can be involved in the last step of triacylglycerol synthesis. The first involves the enzyme diacylglycerol acyltransferase (DAGAT), transferring an acyl group from acetyl-CoA to the sn3 position of diacylglycerol to form TAG (triacylglycerol). The second pathway involves a phospholipid:diacylglycerol acyltransferase (PDAT) that transfers the sn2 acyl chain from phosphatidylcholine, a major membrane component, to the sn3 of DAG (diacylglycerol).

While esterified to phosphatidylcholine, oleoyl chains can be further desaturated to 18:2 (linoleoyl) and 18:3 (linolenoyl) by fatty acid desaturases FAD2 and FAD3.

2.3.1.3 Storage of triacylglycerol in oil bodies

Triacylglycerol is then transported to storage lipid organelles. These oil bodies are also called oleosomes or spherosomes and are surrounded by a single layer membrane. They contain mainly TAG, and 1 to 5% of highly hydrophobic proteins called oleosins. These proteins might have a role in biogenesis and stabilisation of the oil body. They might also be involved in the hydrolysis of TAG, being a potential binding site for lipases (Napier et al., 1996).

2.3.1.4 Factors regulating triacylglycerol biosynthesis in plants

Availability of precursors of fatty acids synthesis

Results obtained by Bao and Ohlrogge (1999) show that supplying elm (*Ulmus carpinifolia* Gled. and *Ulmus parvifolia* Jacq.) and cuphea *(Cuphea lanceolata* W.T.

Aiton) embryos with sucrose or glycerol did not increase the final TAG content, whereas supplying a ten-carbon acyl chain did. The availability of fatty acids might thus be limiting for TAG biosynthesis. Furthermore, Focks and Benning (1998) showed a relationship in Arabidopsis mutant *wrinkled1* between the fatty acid content of the seed and the disruption of enzymes involved in the conversion of sucrose into fatty acid precursors pyruvate and acetate. This means that the availability of fatty acids could be related to the activity of several glycolytic enzymes affecting the availability of fatty acid precursors.

Acyl-carrier proteins

Although Acyl-carrier proteins (ACP) are not enzymes, they are key cofactors in fatty acid biosynthesis. They are part of a gene family, coding for small acidic proteins of about 9 kD (Ohlrogge and Browse, 1995). The reason for the presence of different isoforms of this protein is not established, but the tissue-specificity of their pattern of expression and their different affinities for distinct enzymes of the pathway indicate that they could have a crucial role in determining the ratio of fatty acids in seed or leaf lipids. This has been confirmed by the overexpression of the ACP-1 gene in *Arabidopsis thaliana*, which results in a significant decrease in the synthesis of 16:3 fatty acids, and a corresponding increase in 18:3 acyl chains (Branen et al., 2001).

Acetyl-CoA carboxylase

Acetyl CoA carboxylase (ACCase), which is involved in the first step of fatty acid synthesis, has been extensively studied in plants as well as in yeast and mammals. Two classes of ACCases exist in plant cells (Alban et al., 1994). One is a heteromeric complex called the prokaryotic or multi-subunit ACCase. It is a 700 kD complex composed of at least three subunits: a biotin carboxylase, a biotin carboxyl carrier protein and a carboxyl transferase. The other class is a homodimeric complex of a multifunctional protein, with all three functions gathered on one large peptide of over 200 kD. This class of ACCase has a lower K_m for acetyl-CoA and is inactivated by herbicides of the diphenoxypropionic acid- and cyclohexanedione-type, to which the prokaryotic class is not sensitive. In most plants, the prokaryotic ACCase complex is located in plastids and the eukaryotic form in

the cytosol, but not in the Gramineae, where both are of the eukaryotic form. Plastidial ACCases supply malonyl-CoA for *de novo* fatty acid biosynthesis, whereas cytosolic ACCase supplies malonyl-CoA for different pathways, including cytoplasmic fatty acid elongation and flavonoid synthesis. The discovery of a eukaryotic ACCase in plastids of *Brassica napus* could indicate that this enzyme, and not the prokaryotic form, is implicated in fatty acid biosynthesis in all plants, and not only in the Gramineae (Schulte et al., 1997).

Plastidial ACCase is activated by light and inhibited by exogenous lipids, which indicates a possible feedback regulation (Ohlrogge et al., 1993). This makes ACCase a very probable limiting step in the pathway.

ACCase genes have been cloned from different plants, including Arabidopsis, maize, wheat and oat. In wheat, the complete sequence of the gene spans 12kb and contains 29 introns. Eight transcripts have been identified in wheat, six coding for a cytosolic form and two for a plastidial form of the enzyme (Gornicki et al., 1994; Podkowinski et al., 1996).

Ketoacyl-ACP synthases (KAS)

The three KAS enzymes are involved in the acyl chain elongation (Jaworski et al., 1989, Ohlrogge and Browse, 1995) . KASI is a homodimer of a 50 kD protein, coded by the gene KASB, whereas KASII is a heterodimer of KASB and a 46 kD peptide coded by KASA (Slabaugh, 1998, Mekhedov et al., 2000). In soybean, an antisense knockout of KASA has been shown to increase the proportion of 16:0 in TAG whereas a knockout of KASB causes an increase of 16:0 but also a large decrease of the activity of the entire fatty acid pathway (Voelker and Kinney, 2001). KASIII is coded by a distinct gene that has little sequence homology with KASA and KASB (Tai and Jaworski, 1993).

Delta-9-stearoyl-ACP desaturase

Delta-9-stearoyl-ACP desaturase catalyses the first desaturation of the acyl chain in the plastid. It is therefore unlikely that this enzyme is rate-limiting in TAG synthesis, but it could be implicated in controlling the amount of stearic (18:0) or oleic (18:1) acid exported to the endoplasmic reticulum, and thus the level of saturation in TAG. As

expected, the knockout of this enzyme in cotton caused an increase in stearic acid content (Liu et al., 2000). Knockout plants also showed a significant decrease in palmitic acid (16:0), which could be another indication of the regulatory role of KASII, and the existence of a retro-control by stearic acid.

Glycerol-3-phosphate acyltransferase

Glycerol-3-Phosphate Acyltransferase (GPAT) catalyses the first reaction leading to the synthesis of TAG in the endoplasmic reticulum, and attaches an acyl chain on the sn1 position of the glycerol backbone, producing monoacylglycerol. The overexpression of a safflower glycerol-3-phosphate acyltransferase in Arabidopsis has led to increases in seed oil content of up to 21% (Jain et al., 2000). This indicates that GPAT could be a limiting step of TAG synthesis.

Lysophosphatidic Acyl Acyltransferase

Lysophosphatidic acyl acyltransferase (LPAAT) attaches an acyl chain on the sn2 position of monoacylglycerol, forming diacylglycerol. DAG is a precursor of TAG, but also a major intracellular second messenger molecule. LPAAT is represented in the cell by several isoforms that have distinct selective affinities for different fatty acids. The overexpression of a yeast LPAAT in Arabidopsis resulted in an increase of up to 48% in seed oil content (Zou et al., 1997), which suggests that LPAAT is also a limiting enzyme in the synthesis of TAG.

Diacylglycerol Acyl Transferase

Diacylglycerol acyl transferase (DAGAT) is a key enzyme in plant oil synthesis in that it is the only enzyme exclusively committed to TAG biosynthesis. DAGAT activity in seeds has been shown to increase rapidly during oil accumulation, and then decrease as lipid levels stabilize (Perry and Harwood, 1991). Under conditions of high fatty acid production (when light and carbon are not limiting), it has been shown that DAG accumulates more rapidly than TAG. This suggests that the addition of the third acyl group might be a bottleneck in TAG synthesis when DAG and fatty acid supply are not limiting.

2.3.2 Storage protein biosynthesis

2.3.2.1 The journey of nitrogen through the plant

Protein biosynthesis is a complex pathway, involving itself several biochemical pathways, as well as numerous exchanges of compounds between sources and sinks of the plants (Fig. 2.3).

Assimilation of nitrogen starts in the roots with the intake of NO_3^- or NH_4^+. It has been shown that barley and wheat plants have the capacity to process more nitrogen if it is supplied by perfusion in the culm rather than in the soil (Ma et al., 1994). This results in an increase in seed protein content, suggesting that nitrogen intake by the root system is limiting.

Nitrates can be stored in the plant or reduced by the nitrate reductase, an enzyme located in both roots and leaves. NADH or NADPH is required as electron donor for this step, and therefore light levels and photosynthesis influence the activity of this enzyme. Nitrogen is then incorporated into the four nitrogen-transport amino acids of higher plants: glutamate, glutamine, aspartate and asparagine. Glutamate is also a major nitrogen donor in most transamination reactions in amino acid biosynthesis. These amino acids constitute up to 64% of the free amino acid content of Arabidopsis leaf extract (Lam et al., 1995) and are synthesised by glutamate synthases, glutamine synthases, aspartate aminotransferases and asparagine synthases, respectively. Those amino acids are then transported to the sinks of the plant, where they are either stored as free amino acids, or used in the metabolism of other nitrogen-containing molecules like amino acids (other than the four mentioned above), proteins or nucleic acids. It has been shown in maize that increased productivity is caused by the ability of the plant to accumulate nitrate in the leaves during vegetative growth, and remobilize it efficiently from senescing leaves during grain filling (Lam et al., 1995; Hirel et al., 2001).

In seeds, proteins can be accumulated to high levels in protein bodies stored in different cell compartments (Crofts et al., 2005), and will constitute an important source of carbon and nitrogen for germination.

2.3.2.2 Enzymes involved in nitrogen assimilation in plants

Glutamine Synthase

Glutamine synthases (GS) catalyze the synthesis of glutamine from NH_4^+ and glutamate. Two isozymes of this enzyme exist in plants. GS_2 is the plastidial isozyme, and is active mainly in leaves, incorporating NH_4^+ originating from primary assimilation and photorespiration. GS_2 is also involved in the assimilation of ammonium in roots, although it shares this role with the cytosolic isozyme, GS_1 (Lam et al., 1995; Hirel et al., 2001). This second isozyme does not seem to be expressed in mesophyll cells, but is implicated in the reassimilation of NH_4^+ liberated by protein hydrolysis during germination and leaf senescence. One gene for GS_2 and three genes for GS_1 have been identified in Arabidopsis (Peterman and Goodman, 1991). In tobacco (*Nicotiana tabacum* L.), transcripts for the GS_1 gene Gln-1 are undetectable in mature leaves, but their expression increases with the start of senescence (Brugiere et al., 1999).

Glutamate Synthase

Glutamate synthases (GOGAT, L-glutamine:2-oxoglutarate aminotransferase) catalyze the synthesis of glutamate from oxaloacetate in a glutamine-dependent reaction. GOGAT enzymes are thus implicated in two potentially limiting steps of protein synthesis: nitrogen assimilation and amino acid synthesis (Lam et al., 1995).

Glutamate synthases are plastidial enzymes, and exist in two different forms: ferredoxin dependent (Fd-GOGAT) and an NADH-dependent form (NADH-GOGAT). Different isozymes exist for both enzymes. Fd-GOGAT is mainly implicated in recapturing NH_4^+ from photorespiration, in conjunction with GS2, whereas NADH-GOGAT acts in the primary assimilation of N in roots (Lam et al., 1995). Transformed tobacco plants with a constitutively expressed NADH-glutamate synthase showed increased dry weight, and total nitrogen content in green tissues (Chichkova et al., 2001) but the effect on seed protein was not measured.

2.4 Monitoring gene expression

New techniques for detecting the expression of specific genes or for following the expression of many genes simultaneously are constantly being developed. A few mainstream methods with their advantages and shortcomings are presented here.

2.4.1 Detecting the expression of specific genes

2.4.1.1 Northern blot

Northern analysis (Alwine et al., 1977) remains a standard method for detection and quantitation of mRNA, and consists in the hybridization of a labelled nucleotidic probe (DNA or RNA) with RNA immobilized on a membrane after separation through gel electrophoresis.

Northern analysis allows for a straightforward determination of RNA size, amount and integrity, as well as a direct comparison of message abundance between samples. It is also the preferred method for detecting alternative splicing of transcripts. A strong advantage of this method for evaluating levels of gene expression is that it is a direct detection and quantifying method and does not require steps of PCR amplification.

The major limitation of Northern analysis is the lack of sensitivity of the method. The large amount of RNA required makes this method unsuitable in cases where the starting material is limited or the level of expression of the genes investigated is very low.

Another limitation of Northern blotting is the difficulty associated with multiple gene analysis: it is usually necessary to strip the initial probe from the membrane before hybridizing with another probe. This process can be time consuming and labour-intensive, and therefore only suitable for investigating a very limited number of genes per experiment.

Finally, this method rests on the availability of suitable probes corresponding to the genes targeted by the experiment.

2.4.1.2 Reverse transcription-polymerase chain reaction (RT-PCR)

RT-PCR is one of the most sensitive technique for mRNA detection and quantitation currently available (Carding et al., 1992). In this process, RNA strands are

first reverse-transcribed into cDNA, and the resulting DNA amplified by PCR using primers specific to the genes of interest.

Compared to Northern blot analysis, RT-PCR can quantify mRNA levels from much smaller samples, even from a single cell. It is fairly straightforward to perform, is relatively inexpensive, and can be performed on many samples simultaneously. It can sometimes even yield some useful results when the original RNA sample is of mediocre quality.

Nevertheless, there are some major limitations to this strategy. First, this method requires some knowledge of the sequence of the genes whose expression is monitored, in order to obtain gene-specific primers to be used in the PCR amplification step. A second major limitation of RT-PCR is that, as this method involves a PCR amplification, the relative quantitation of transcripts can be biased by the sequence-specific efficiency of PCR reactions. This becomes a problem particularly when samples from different genotypes are compared, and the annealing efficiency of the primers to the templates might vary because of sequence differences in the target region.

2.4.1.3 Differential display

The principle of differential display is the specific reverse-transcription and amplification of a subset of mRNAs from total RNA, due to differences in nucleotide sequence at the 3' end of modified oligo-dT primers (Liang and Pardee, 1992). The products of amplification are then separated on poly-acrylamide gels, and the profiles yielded by different samples compared.

As it relies on PCR amplification, this method can be performed with very little starting material, and on a fairly large number of samples simultaneously. It doesn't require any *a priori* information about the genes that might be differentially expressed between the sequences. Also, it allows the recovery of sequence information from differentially expressed genes through the isolation and sequencing of polymorphic bands.

As this method is also based on PCR amplification, some of the limitations of differential display are the same as those mentioned for RT-PCR, particularly relating to quantitation and the impossibility to discriminate between differential expression and differences in efficiency of PCR amplification. Another drawback of this method is the

difficulty of isolating bands corresponding to potential genes of interest from acrylamide gels, and the separation of PCR products of similar size.

2.4.1.4 Subtractive suppressive hybridization (SSH)

Subtractive suppressive hybridization is based on a technique called suppression PCR (Siebert et al., 1995), prohibiting the amplification of molecules that do not contain annealing site for a gene-specific primer. SSH combines normalization and subtraction of cDNA or genomic DNA (Diatchenko et al., 1996 and 1999). The normalization step equalizes the abundance of cDNAs within the target population (also called "tester"), and the subtraction step excludes the sequences common between the target and the reference (also called "driver") populations. After one round of subtractive hybridization, the abundance of different cDNAs is the subtracted library is normalized, and an enrichment of 1000- to 5000-fold of rare DNAs can be achieved (Diatchenko et al., 1996).

A major strength of the SSH technique is the enrichment in rare cDNAs, allowing the isolation of sequences for low-copy mRNA, such as transcription factors and other gene-expression regulatory elements. It also allows the isolation of differentially expressed genes without any prior information on gene expression in the samples studied, which is a major advantage when dealing with an organism for which little gene sequence information is available.

A technical advantage of SSH, particularly when comparing this method to differential display, is the ease with which the pool of differentially expressed genes obtained can be subsequently cloned and sequenced, without having to purify each DNA fragment from an acrylamide gel. The enriched and normalized pool of cDNA can also be used as hybridization probe to screen gene libraries, making the applications of this technique even more versatile.

Although very powerful, this technique also has limitations. First, it requires several micrograms of mRNA as starting material, which might be limiting in certain circumstances. Steps of PCR amplification can help circumvent this limit, but might introduce some bias due to different rates of amplification of some transcripts. Another limitation is that the procedure requires the digestion of transcripts with a four-base recognition site restriction enzyme. This means that the size of the cDNA fragments obtained through this procedure will be around 600 bp on average. This might be a

disadvantage when full-length cDNA are desired, and means that the genes corresponding to the clones isolated through this procedure might be difficult to characterize without further experimental steps. Finally, this procedure involves numerous steps, and is quite labour-intensive.

2.4.2 Methods for following the expression of many genes simultaneously

2.4.2.1 Serial analysis of gene expression (SAGE)

Serial analysis of gene expression produces a snapshot of the genes expressed in the sample investigated (Velculescu et al., 1995 and 1997). A short sequence tag (10-14bp) of each mRNA expressed in the sample is obtained from a region unique to each transcript, usually the 3' end. Those tags should be as short as possible and contain sufficient information to uniquely identify each transcript by comparing its sequence to known sequences in databases. The sequence tags are then linked together to form long serial molecules that can be cloned and sequenced. This provides a gene expression profile of the sample analysed, with the number of times a particular tag is observed providing an estimation of the expression level of the corresponding gene.

No a priori information on gene expression in the studied tissue is required for performing SAGE, which allows the discovery of uncharacterized genes. Nevertheless, it does require the availability of extensive sequence information from the studied organism, in order to be able to identify transcripts by a short unique oligonucleotide sequence. It has up to now mainly been used in cancer research (Hermeking, 2003), but starts to be used more widely in other organisms as available sequence information increases (Poroyko et al., 2005; Robinson et al., 2004; Chakravarthy et al., 2003).

2.4.2.2 Arrays

DNA arrays have become a common way to study large-scale expression profiles (Southern, 1996; Case-Green et al., 1998). Using arrays, expression levels can be measured for hundreds or thousands of genes simultaneously, depending on the number of DNA targets spotted on the arrays. These targets can be known or uncharacterized clones, or short oligonucleotides. Hybridizations are performed with complex probes,

most commonly mRNA populations obtained from cell lines or tissues of interest (reviewed in Freeman et al., 2000; Granjeaud et al., 1999).

DNA arrays are usually divided in macroarrays and microarrays, depending on the number and density of target spots on the array. Macroarrays feature tens to several hundred DNA targets, usually spotted on nylon membranes at a density of 10 to 100 per square centimetre (Granjeaud et al., 1999). Screening is usually done through hybridization with radioactive probes (^{32}P or ^{33}P) and results are acquired with imaging systems and software. Microarrays can be printed at a density of up to several thousand spots per square centimetre, and are more commonly spotted on glass and screened with two contrasting fluorescent probes. They can also be nylon-based, and screened with radioactive or probes.

Microarrays used to be cost-prohibitive for certain investigations, and macroarrays were much cheaper alternatives. The cost of printing microarrays, and the availability of the required equipment have made microarrays much more accessible recently. The choice between the use of macroarrays and microarrays for studying gene expression will now mainly be decided by the number of genes available for arraying. If only a few hundred to a few thousands of clones are available for the studied organism (such as in oat), macroarrays are the logical option.

The main factors influencing the quality of the results obtained with DNA arrays can be identified as the following: the quality of the arrays, the quality of the starting material for generating the complex probes, the presence of appropriate controls on the array, as well as an appropriate experimental design including biological replicates. Due to the vast amount of data generated by array experiments, careful statistical analysis of the results is also required.

Although DNA arrays have been widely used, recent insights into their limitations have led to caution in the interpretation of the data generated (Breitling, 2006; Fathallah-Shaykh, 2005; Xu, 2005; Rhodius and LaRossa, 2003). Better data analysis tools and strategies might still have to be developed to reach the full potential of DNA arrays.

Table 2.1 Ploidy levels, genome complements, and representative species in the genus *Avena*.

Ploidy	Genome	Species
Diploid	CC	*A. clauda*
Diploid	AA	*A. strigosa, A. wiestii, A. atlantica, A. hirtula, A. nuda*
Near autotetraploid	AABB	*A.abyssinica, A. vaviloviana*
Allotetraploid	AACC	*A. marocana*
Allohexaploid	AACCDD	*A. fatua, A. sativa, A. sterilis*

Figure 2.1 Schematic representation of the complex interactions between genotype, phenotype and environment. Circular arrows represent retro-control and double-headed arrows represent feedback interactions.

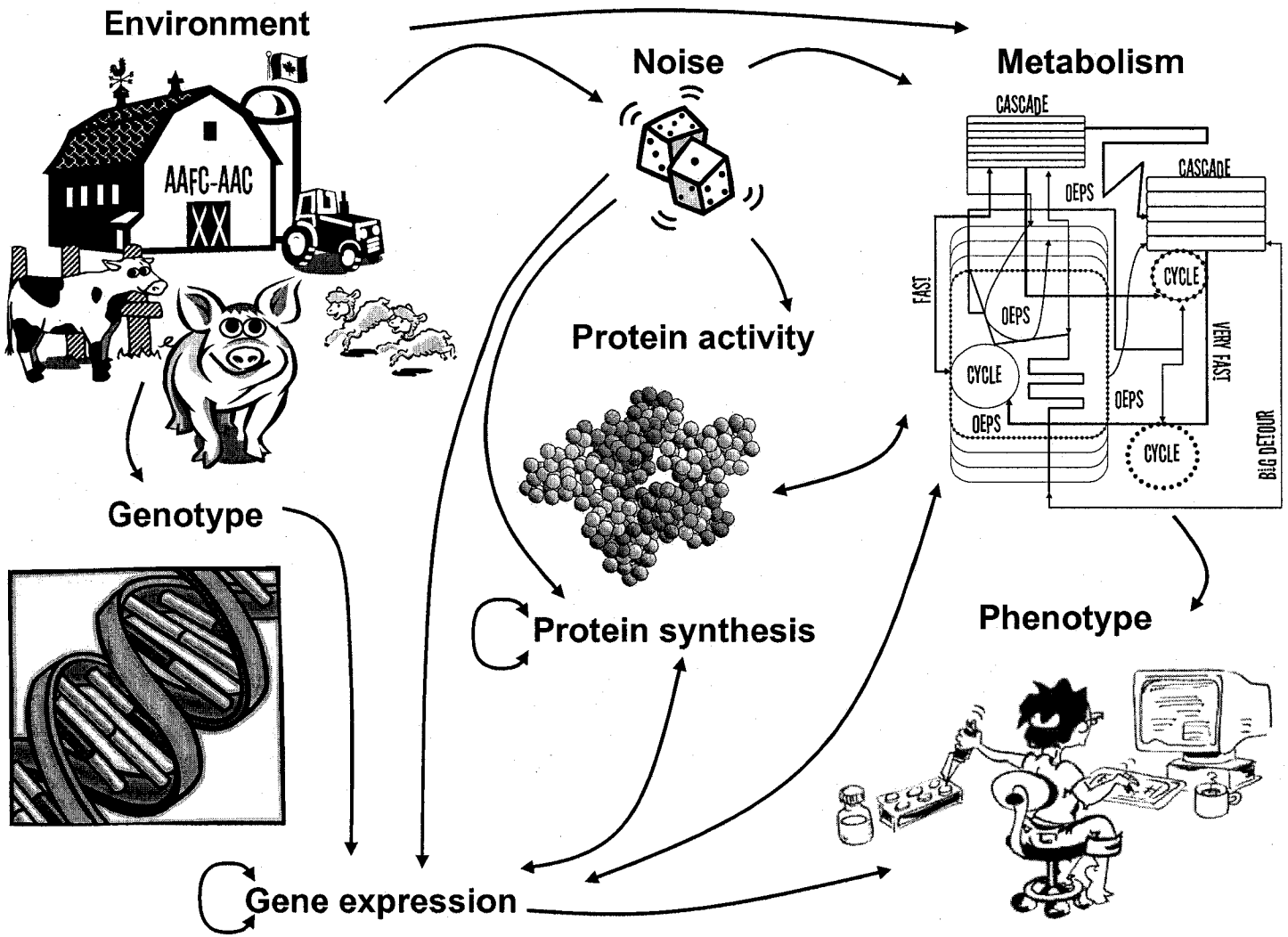

Environment
AAFC-AAC
Noise
Metabolism
CASCADE
CASCADE
OEPS
FAST
CYCLE
OEPS
VERY FAST
CYCLE
OEPS
OEPS
CYCLE
BIG DETOUR
Protein activity
Genotype
Protein synthesis
Phenotype
Gene expression

Figure 2.2 A simplified view of triacylglycerol biosynthesis in plants. The proteins in bold are among the ones that have been documented to affect lipid content in plants. ER stands for endoplastic reticulum, ACP for acyl-carrier proteins, CoA for coenzyme A, ACCase for acetyl-CoA carboxylase, G3P for glycerol-3-phosphate, GPAT for glycerol-3-phosphate acyltransferase, G2P-FA for glycerol-2-phosphate with a fatty acid attached to position 3, LPAAT for lysophosphatidyl acyltransferase, DAG for diacylglycerol, DAGAT for diacylglycerol acyltransferase and TAG for triacylglycerol.

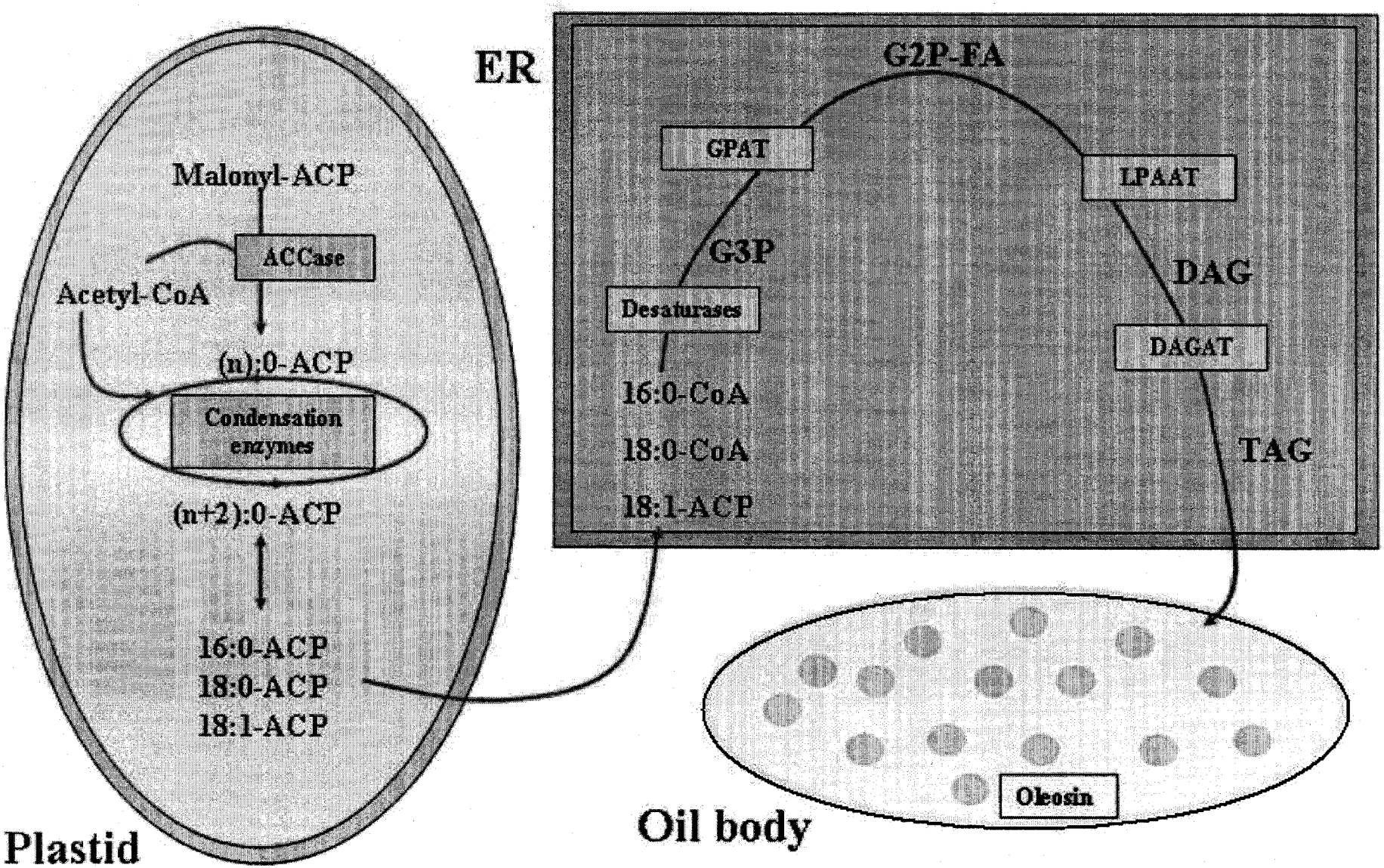
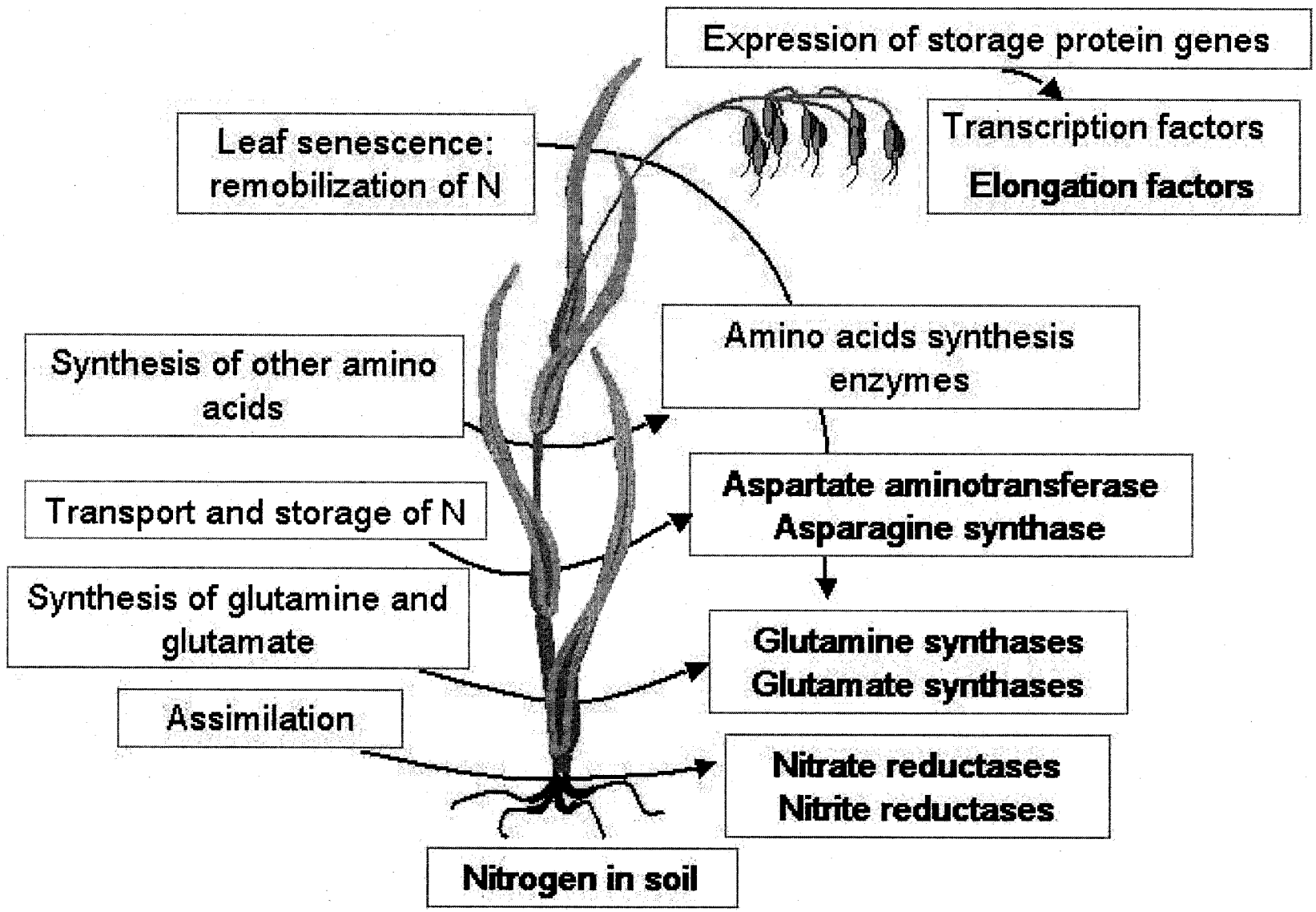

ER
G2P-FA
GPAT
LPAAT
G3P
DAG
Desaturases
DAGAT
16:0-CoA
18:0-CoA
18:1-ACP
TAG
Malonyl-ACP
ACCase
Acetyl-CoA
(n):0-ACP
Condensation enzymes
(n+2):0-ACP
16:0-ACP
18:0-ACP
18:1-ACP
Oleosin
Plastid
Oil body

Figure 2.3 A simplified view of nitrogen metabolism in plants. The proteins in bold are among the ones that have been documented to affect storage protein content in plants.


Expression of storage protein genes
Transcription factors
Elongation factors
Leaf senescence: remobilization of N
Amino acids synthesis enzymes
Synthesis of other amino acids
Aspartate aminotransferase
Asparagine synthase
Transport and storage of N
Synthesis of glutamine and glutamate
Glutamine synthases
Glutamate synthases
Assimilation
Nitrate reductases
Nitrite reductases
Nitrogen in soil

Preface to Chapter 3

In order to better understand the genetic variability underlying protein and lipid content, both of which are important quality characteristics of the oat kernel, we wanted to obtain sequences from genes involved in the metabolism of lipids and protein and to evaluate the level of sequence variability existing among cultivars of oat. When I started this project, many proteins involved in these pathways had been well characterized in plants but few oat sequences were available in public databases. I selected 21 genes corresponding to proteins that are involved in potential limiting steps of the pathways, then collected and aligned plant sequences for these genes. I identified regions in the alignments that were conserved across species. Starting with the hypothesis that these regions would also be very similar in oat, I designed PCR primers to target these regions.

Chapter 3 describes the analysis of sequences derived from PCR products obtained when these primers were used to amplify DNA from 10 oat cultivars with contrasting levels of kernel lipid and protein content. Sequences will be submitted to GenBank. The manuscript will be submitted to a refereed journal, and will be coauthored by myself, Dr. Stephen Molnar, Dr. Nicholas Tinker and Dr. Diane Mather. I selected the target genes, analyzed the template sequences, designed the primers and tested them in the laboratory, cloned the PCR products, carried the phylogenetic analysis and prepared the manuscript. Dr Stephen Molnar contributed research facilities, constructive suggestions and corrected the manuscript. Dr. Nicholas Tinker contributed constructive suggestions and corrected the manuscript. Dr. Diane Mather supervised the research, provided equipment and funding and corrected the manuscript.

Chapter 3
Nucleotide diversity in genomic DNA of oat genes involved in lipid and protein biosynthesis

3.1 Summary

In order to study sequence variability in genes involved in lipid and protein biosynthesis in oat (*Avena sativa L.),* we identified 21 genes corresponding to potential bottlenecks in triglyceride and protein biosynthesis in plants. Plant sequences for these genes were obtained from public sequence databases, and heterologous alignments were built for each candidate gene. Primers anchored in gene regions conserved among plant species were designed and used to amplify oat sequences. We obtained sequences for eight of the 21 candidate genes from each of 10 oat cultivars showing variations in lipid and/or protein content. Phylogenetic analysis showed little sequence variation among these sequences, but SNPs were nevertheless identified that could differentiate between some of the cultivars included in the study. Sequences for most genes clustered in three sequence families, with very high sequence conservation within each family. These families might correspond to homeologous versions of the genes, originating from the three homeologous oat genomes. Several sequence polymorphisms among these families could be used for the development of genome-specific markers.

3.2 Introduction

Hexaploid oat (*Avena sativa L.*) is grown in temperate regions and is used as animal feed, as human food and as an additive for cosmetic products. It has a large genome, consisting of three basic genomes (A, C and D), each containing seven pairs of chromosomes.

Among the important quality characteristics of oat are its lipid and protein content. The high protein content and good amino acid balance make it nutritionally superior to other cereal grains. The proportion of threonine, methionine and particularly of lysine, three amino acids considered nutritionally limiting, tends to be higher than in other cereals (Peterson and Brinegar, 1986). With a range of 2.0 to 16.2% of the total groat weight, oat has the highest lipid content among cereals (Schipper and Frey, 1991). Oat

lipids, like other cereal lipids, are nutritionally valuable because of their high content of unsaturated fatty acids, and particularly their abundance in linoleic acid (18:2). As the heritability of groat oil content is high, ranging from 63% to 93% (Schipper et al., 1991), lipid content responds well to selection towards either high or low oil content phenotypes and both have been breeding objectives. Breeding for a higher lipid content would make oat a high energy feed grain with good nutritional qualities, but low fat content is still considered desirable in oat grains intended for human consumption.

Metabolic pathways for lipid and protein biosynthesis in plants have been intensely studied (Larkins et al., 1982; Ohlrogge and Browse 1995; Voelker and Kinney 2001; Hills, 2004), and several steps in these pathways have been identified as potentially limiting. Many sequences are available in public databases for genes corresponding to the enzymes and proteins involved in those steps, but few of them are from oat.

In the present study, PCR primers derived from heterologous alignments of plant sequences were used to clone partial sequences from oat for selected genes implicated in lipid and protein biosynthesis. The sequences obtained, originating from 10 oat cultivars, were used to assess the degree of sequence variability in oat.

3.3 Material and methods

3.3.1 Genetic material

Oat cultivars Kanota, Ogle, Terra, Marion, Dal, Exeter, Francis, Rigodon, Hinoat and Newman were selected because the recombinant inbred line (RIL) populations derived from the crosses Kanota x Ogle, Kanota x Marion, Terra x Marion and Dal x Exeter segregate for lipid content (Kianian et al., 1999; Groh et al., 2001; De Koeyer et al., 2004; N. Tinker, personal communication), and the RIL population derived from Hinoat x Newman segregates for protein content (N. Tinker, personal communication). Two of these cultivars are high-lipid cultivars (Dal, Rigodon) and two are low-lipid cultivars (Exeter, Francis). Hinoat is a high-protein cultivar and the other five cultivars have intermediate contents of both lipid and protein (Kanota, Ogle, Terra, Marion and Newman). DNA from these ten cultivars was extracted as described by Wight et al. (2003).

3.3.2 Choice of candidate genes and acquisition of template sequences

Metabolic pathways involved in triacylglycerol biosynthesis and nitrogen metabolism in plants were examined and 21 genes encoding enzymes and proteins involved in potentially limiting steps of the pathways were selected. Priority was given to rate-limiting steps in the pathways and to genes that have been documented to influence oil and protein synthesis (Table 3.1). Across all 21 genes, a total of 361 sequences originating from 61 plant species were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.htmL). To be considered for analysis, sets of template plant sequences for each candidate gene had to include one or more complete cDNA sequences, genomic sequences to position exons and introns and grass sequences had to be represented. When possible, entries from different genotypes within species were included in order to identify regions of higher intra-specific variability.

3.3.3 Alignment of heterologous template sequences

Sequences were annotated and trimmed, then aligned using CLUSTAL W (Thompson et al., 1994) in BioEdit Sequence Alignment Editor (Hall, 1999). The alignments of cDNA and genomic DNA sequences were fitted manually. A consensus template sequence was derived from each alignment, and adjusted manually so as to bias the consensus towards the cereal sequences when some were included in the alignment.

3.3.4 Primer design and PCR

For each consensus template sequence, three to eight PCR primers anchored in conserved regions neighbouring introns were designed using Primer3 (Rozen and Skaletsky, 2000) with the following parameters: the product size range was set between 500 and 3000 bp; the minimal primer size at 18 nucleotides, the optimal size at 20 nucleotides, the maximum size at 24 nucleotides; the minimal primer melting temperature at 55°C, the optimal melting temperature at 60°C, the maximum melting temperature was set at 63°C; the optimal GC% at 50%; the maximum number of undetermined nucleotides at two.

This resulted in the design of 88 primer pairs, involving a total of 93 primers. PCR reactions were conducted using a Mastercycler Gradient thermocycler (Eppendorf,

Hamburg, Germany), with a reaction volume of 25 μL containing 1.5 mM of $MgCl_2$, 200 μM of dNTP and 2.5 units of Taq polymerase (Invitrogen, Carlsbad, CA). The optimal annealing temperature for each primer pair was determined from a gradient of temperatures ranging from 53°C to 65°C. Cycling conditions started with initial denaturation at 94°C for 3 min, followed by 40 cycles of 30 s at 94°C, 45 s at Tm, 2 min at 72°C. A final extension reaction was performed at 72°C for 10 min. PCR products were separated on 1-2% agarose gels.

3.3.5 DNA purification, cloning and sequencing

For at least one primer pair per gene, PCR products of the size expected based on the alignments of template sequences were extracted from agarose gels using the QIAquick Gel Extraction Kit (Qiagen Inc., Chatworth, CA). The purified DNA was cloned in the pDrive vector using PCR Cloning Kits (Qiagen Inc., Chatworth, CA). For each ligation, ten random clones were picked and sequenced.

Sequences were vector- and quality-trimmed and imported into a customized relational database (Couroux and Tinker, unpublished). Contigs were assembled using SeqMan (DNASTAR Inc., Madison, WI) based on an overlap-criterion of 30 bases and 90% similarity for introduction into a contig. Further details about these assembly parameters are discussed in Hattori et al. (2005).

The similarity of cloned sequences to known gene sequences from GenBank was determined using BLAST (Altschul et al., 1990) from NCBI (http://www.ncbi.nlm.nih.gov/). Subprogram blastn (2.0.14) was used to identify nucleotide similarity, and blastx (2.2.6) was used to identify translated peptide similarity. Gene homology was tentatively declared when an alignment was identified at a global database expectation value of 0.3 and results were verified manually to exclude artefacts.

3.3.6 Phylogenetic analysis of derived oat sequences

Each contig alignment of oat sequences corresponding to a target gene (as determined from BLAST results) was trimmed so that each sequence in the alignment spanned the same region of the target gene (Fig. 3.1). Identical sequences originating from the same cultivar were considered redundant and only one of them represented in the alignment. Corresponding genomic DNA sequences from rice (*Oryza sativa L.)*

including both cultivar groups *japonica* and *indica* (when available) were added to the alignment. Oat and rice sequences were aligned using CLUSTAL W (Thompson et al., 1994). Phylogenetic analysis was performed using tools from the Phylogeny Inference Package, PHYLIP 3.6 (Felsenstein, 2004). For each alignment, sequences were bootstrap re-sampled using SEQBOOT to create 100 sequence sets. Phylogenetic relationships were estimated on those sequence sets by the parsimony method (Kluge and Farris, 1969) using DNAPARS, with a rice sequence as a root and a random input order of sequences. A consensus tree was then created using CONSENSE.

A group of sequences was considered to be a sequence family when a branch occurring more than 50% of the time in the bootstrap trees supported the clade, and it included a sequence from a cultivar represented in more than one clade.

Distance matrices were generated for oat sequences of each alignment, as well as on alignments of joined exon and intron sequences, using CLUSTALDIST (Thompson et al., 1994) using the Kimura model of nucleotide distribution (Kimura, 1980), as well as the F84 model (Hasegawa et al., 1985). Gaps were included in the analysis, and no correction was made for multiple substitutions, as the sequences are evolutionarily close.

3.4 Results

3.4.1 PCR amplification of oat sequences with primers derived from heterologous sequence alignments

For each of the 21 candidate genes, at least one primer pair gave a detectable PCR product of the size expected according to the heterologous alignment. No detectable amplification size polymorphisms were observed among the 10 cultivars for any of the 88 primer pairs tested. For eight of the 21 genes selected, cloned sequences corresponding to the targeted gene were recovered (Table 3.1). For most target genes, some unrelated sequences of a similar size were amplified and cloned together with the target sequences. These sequences had no BLAST similarity to known genes, and were not considered further in this study. For the remaining 13 genes, cloned sequences did not correspond to target sequences (Table 3.1).

For the eight genes for which we obtained sequences that matched the target gene, the trimmed sequence alignments ranged in length from 238 to 1005 nucleotides. The

trimmed alignments for sequences matching elongation factor-1α and glutamate synthase fell within a single exon in the corresponding rice gene (Fig. 3.1). The other trimmed alignments each spanned at least one exon-intron boundary.

3.4.2 Phylogenetic analysis of oat sequences

Genetic distances were calculated from the sets of oat sequences matching the eight target genes with the Kimura model of nucleotide distribution and the F84 model. Both methods allows for a difference between transition and transversion rates, but the Kimura model assumes equal base frenquencies, whereas the F84 model allows for different frequencies of each of the four nucleotides. Both models gave very similar results, and only the values obtained with the Kimura model are presented here.

The average nucleotide distances in the regions studied ranged from 0.7% for glutamate synthase, to 10.2% for elongation factor-1α (Table 3.2). Coefficients of variation were high, ranging from 58% to 157%, indicating that the sequences within an alignment are not all similarly distant from each other.

For most genes, the average distance among all oat sequences is substantially larger than the average distance inside sequence families (Table 3.2). Exceptions are glutamate synthase, acetyl-CoA carboxylase and β-ketoacyl-ACP synthase III, which show little overall sequence variation.

For all genes for which both exonic and intronic sequences were available, the average distances among sequences are distinctly higher in introns than in exons (Table 3.2): 7.87% on average in intronic sequences and 2.85% in exonic sequences. For acyl-carrier protein, the average distance among intronic sequences is more than seven times higher than the average distance among exonic sequences.

For glutamine synthase, there are only two distinct families of sequences, whereas for acetyl-CoA carboxylase, there are at least five families of sequences. For the other six genes, three distinct sequence families were identified on the phylograms (Fig. 3.3 and Appendix C). These families are supported by manual grouping of sequences according to common sequence motifs (Fig. 3.2 and Appendix B). In some cases, single nucleotide polymorphisms between cultivars within a family can be identified, but in general, an

examination of sequences inside families shows very little sequence polymorphism among cultivars.

3.5 Discussion

We selected 21 genes coding for products involved in key steps of lipid and protein biosynthesis, and designed primers anchored in conserved regions of template plant sequences for these genes. Using those primers, we obtained multiple clones and partial sequences for each of eight genes from up to 10 oat cultivars.

For most primer pairs tested, multiple PCR products were generated. Many of these PCR products, when cloned and sequenced, proved to be unrelated to the sequences that the primers had been designed from. Similar amplification of unrelated sequences has already been reported in oat (Nicholas Tinker, personal communication). When excluding the case of gene families and homeologous sequences, the probability of finding several priming sites in a genome for pairs of 20-mers at a few hundred nucleotide distance is very low. It has been shown in rice that the use of oligonucleotidic "words" was not entirely random throughout the genome, and that some "words" seem to be used at a higher frequency than predicted by random occurrence (Liu et al, 2006). We can speculate that, by selecting primers in genes regions conserved across species, we might have selected oligomeric sequences corresponding to some of these high frequency "words" in the oat genome.

The sequence variability in oat observed in this study is similar to that which has been observed between rice cultivar groups *japonica* and indica. The frequency of single nucleotide polymorphisms and insertions/deletions (indels) between these two cultivar groups was estimated to be 3.22% in exons and 7.35% in introns (Yu et al., 2005) as compared to 2.85% and 7.87% (respectively) in oat.

Phylogenetic analysis allowed the identification of families of sequences closely related to each other. Three distinct families of sequences were identified for all genes, except for glutamate synthase and acetyl-CoA carboxylase. According to Mekhedov et al. (2000), there is only one copy of β-ketoacyl-ACP synthase III gene in the rice genome, which seems confirmed by a search of the Rice Genome Annotation of The Institute for Genomic Research (TIGR) (Table 3.1). We can therefore speculate that the three families

of sequences for this gene correspond to the three homeologous copies from oat basic genomes A, C and D. The other seven genes are multicopy genes in rice, with acetyl-CoA carboxylase, aspartate aminotransferase and glutamate synthase present in two copies, β-ketoacyl-ACP synthase I in up to three copies, elongation factor-1α in four copies, glutamine synthase in seven copies and acyl-carrier protein in eight copies (evaluated by searching the Rice Genome Annotation of The Institute for Genomic Research (TIGR)).

Since our PCR primers were designed to target any copy of these genes, we expected to recover up to three times these numbers of copies in oat. As oat sequences were present in only one of the template alignments (for acetyl-CoA carboxylase), the PCR primers we designed might not have recovered all copies of these genes in oat. It is possible that either homologous or homeologous copies of oat genes have diverged significantly beyond recognition by consensus primers, or that they have been lost completely, so that a PCR reaction using those primers on oat genomic DNA would not amplify any product.

It is also possible that the distance between priming sites in the oat version of some of the genes is very different from that in the template plant sequences. All PCR amplifications with the designed primer pairs gave multiple products, and we recovered only PCR products that had a size close to that predicted by priming sites on the template consensus sequence. Therefore, if the size of the oat amplicon differed greatly from the predicted size, it is possible that the right oat sequences were not recovered. Among the oat sequences corresponding to target genes, the size of exons seems very conserved but the size of introns varies slightly between oat and rice, being sometimes longer in the former or in the latter, and it is possible that for some genes, this difference in length is more pronounced.

Without a much larger sample of sequence variants, or detailed characterization of individual genes, we can only speculate that the recovered families of sequences may represent a mixture of variability within and among gene homologs of oat. Nevertheless, we cannot rule out the possibility that some versions of those genes result from recent duplications, and therefore cluster in the same sequence family. In future studies, it might be possible to differentiate between homeologous genes of the A or C genome and members of gene families by examining sequences derived from diploid species of *Avena*

sativa. It would also be interesting to verify whether the division in sequence families subsides when sequences from more oat cultivars are added to the study.

With sufficient sequence variability among cultivars, it would have been possible to develop gene-targeted molecular markers. The primers tested here did not yield any reproducible length polymorphisms in the amplified fragments, but the fragment sequences did exhibit some single nucleotide polymorphisms (SNPs). One example can be seen in the alignment for glutamate synthase (Fig. 3.2). In sequence family 2, at nucleotide position 419, Kanota and Rigodon sequences feature an A, whereas all of the other cultivars feature a G. The design and use of a marker such as this depends on having an assay that is specific enough to distinguish these SNP alleles independently from the alleles in closely related sequence families, which will presumably segregate at different loci. Having a better representation of each cultivar in all family groups of each gene would help identify useful inter-cultivar polymorphisms. The segregation of these alleles could then be studied in segregating populations. However, the general implications of these results are that DNA sequence variability within these studied genes is low, and it would be difficult to develop markers targeted to these specific regions.

Polymorphisms identified in this study may be useful in ways beyond segregation and linkage analysis. For example, primers designed to overlap the deletions at positions 342-350 and 387-389 in the alignment of aspartate aminotransferase sequences (Fig. 3.2) would easily distinguish between families 1, 2 and 3. This would be particularly interesting if some of the sequence families prove to be genome-specific. As in most polyploids, the question arises as to how each of these three genomes contributes to the functioning of oat as an organism. It would be desirable to determine whether homeologous versions of a gene exhibit differential expression during oat development. Using subgenome-specific primers to perform reverse transcriptase PCR (RT-PCR) on cDNA samples would allow the tracking of the homeologous transcripts in different tissues or developmental stages of the plant. Unfortunately, most of the indels and long sequence polymorphisms occur in introns, and would therefore be of little use in expression studies. There are some useful SNPs in exons though, such as the C/T polymorphism at position 327, which distinguishes families 1 and 2 of glutamate synthase

sequences (Fig. 3.2). Extending the study to other coding regions of these genes might highlight more useful subgenome-specific polymorphisms.

Table 3.1 Enzymes and proteins from pathways involved in lipid and protein biosynthesis, numbers of representative sequences used in alignments, and correspondence to the target genes of PCR products derived from oat cultivars

Biosynthetic pathway	Gene product	Estimated number of genes in rice [a]	Number of plant sequences used in alignment			Primer pairs tested	PCR product sequence corresponds to target gene
			Total	Genomic DNA	Cereals		
Fatty-acid biosynthesis	Acetyl-CoA carboxylase	2	48	41	44	7	Yes
	Acyl-carrier proteins	8	12	2	12	2	Yes
	3-ketoacyl reductase	2	19	12	3	4	No
	β-ketoacyl-ACP synthase I	3	15	1	4	4	Yes
	β-ketoacyl-ACP synthase III	1	9	1	5	4	Yes
	Acyl-ACP thioesterase	4	20	1	4	3	No
Triacylglycerol biosynthesis	δ-9-Desaturase	9	8	2	0	4	No
	Glycerol-3-phosphate acyltransferase	1	19	3	7	3	No
	Lysophosphatidyl acyltransferase	2	9	1	2	3	No
	Diacylglycerol acyltransferase	3	13	1	1	3	No
	Phospholipid:diacylglycerol acyltransferase	1	17	1	1	5	No

Table 3.1 (continued)

Biosynthetic pathway	Gene product	Estimated number of genes in rice [a]	Number of plant sequences used in alignment			Primer pairs tested	PCR product sequence corresponds to target gene
			Total	Genomic DNA	Cereals		
Triacylglycerol storage	Oleosins	5	38	2	38	4	No
Nitrogen assimilation	Nitrite reductase	3	17	1	6	4	No
	Nitrate reductase	1	9	2	9	3	No
	Glutamine synthase	7	21	1	13	4	Yes
	Glutamate synthase	2	8	3	5	8	Yes
	Asparagine synthase	2	15	1	4	4	No
Nitrogen metabolism	Aspartate aminotransferase	2	20	1	6	3	Yes
	Glutamate dehydrogenase	2	18	3	8	8	No
Amino acid biosynthesis	Aspartate kinase	4	10	3	10	4	No
Protein elongation	Elongation factor 1-α	4	19	1	18	4	Yes

[a] Estimated by searching the Rice Genome Annotation of The Institute for Genomic Research (TIGR)

Table 3.2 Average genetic distances among sequences and distances among exonic and intronic sequences, based on phylogenetic analyses of oat sequences corresponding to eight gene products involved in lipid and protein biosynthesis

Gene product	Number of non-redundant sequences in the alignment (number of cultivars represented)	Mean distance between sequences of the same family [a]	Mean distance between oat sequences [a] [Coefficient of variation [b]]		
			Whole sequence	Exons	Introns
Acyl-carrier protein	14 (6)	0.6	3.9 [92]	2.1 [123]	16.4 [74]
Aspartate aminotransferase	54 (10)	0.8	4.7 [157]	1.8 [87]	7.1 [175]
Elongation factor 1-α	9 (4)	2.4	10.2 [58]	10.2 [58]	not available
Glutamate synthase	37 (9)	0.6	0.7 [70]	0.7 [70]	not available
Glutamine synthase	27 (10)	1.7	6.1 [87]	2.5 [72]	13.5 [95]
Acetyl-CoA carboxylase	66 (10)	1.1	2.2 [111]	1.7 [131]	2.8 [101]
β-ketoacyl-ACP synthase I	34 (8)	0.9	3.2 [67]	1.6 [66]	4.8 [75]
β-ketoacyl-ACP synthase III	12 (5)	1.2	2.4 [118]	2.2 [105]	2.7 [129]
Average	31.6 (7.8)	1.2	4.18	2.85	7.87

[a] Number of differences as percentage of total nucleotides obtained with the Kimura nucleotide substitution model
[b] Percentage of mean

Figure 3.1 Position of the gene regions used for phylogenetic analysis of oat sequences on the corresponding rice genes. Arrows span the gene regions represented in the analysis. Boxes are exons, numbered from 5' to 3'. ACP stands for acyl carrier protein; AAT for aspartate aminotransferase; ACCase for acetyl-CoA carboxylase; EF-1α for protein elongation factor-1α; GAS for glutamate synthase; GIS for glutamine synthase; KAS for Beta-ketoacyl-ACP synthase. The numbers in brackets are the approximate total length of the rice gene.

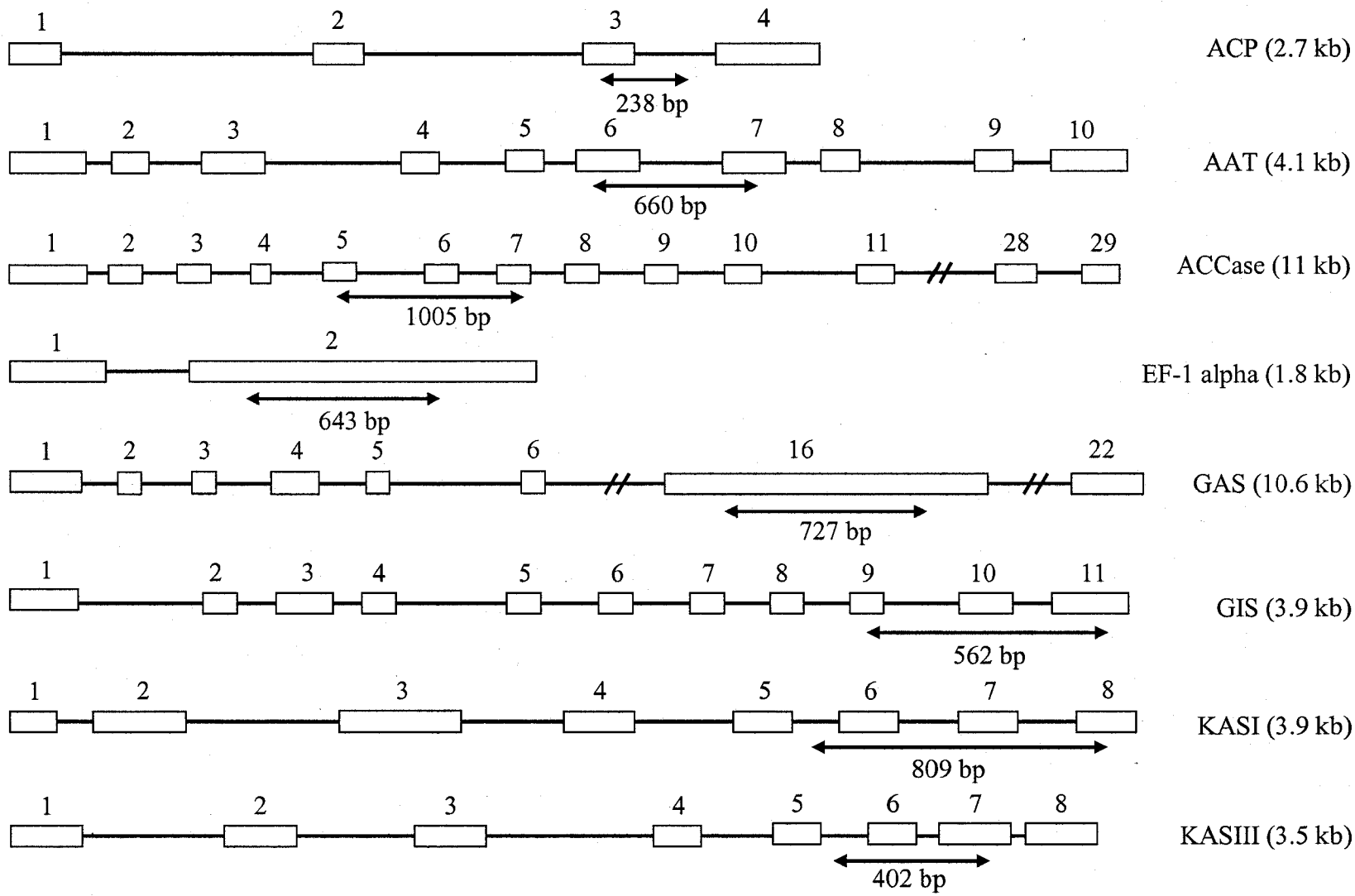
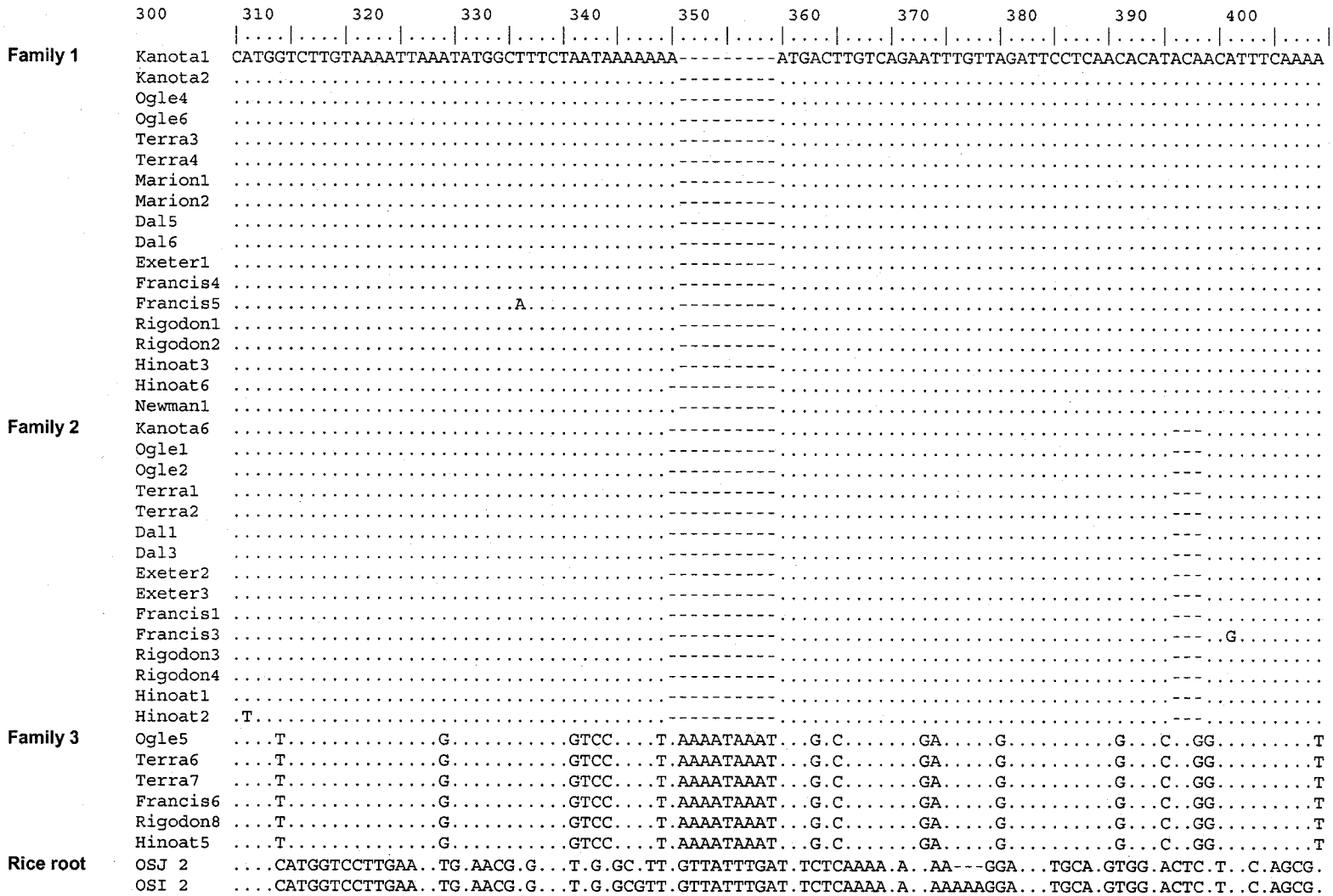
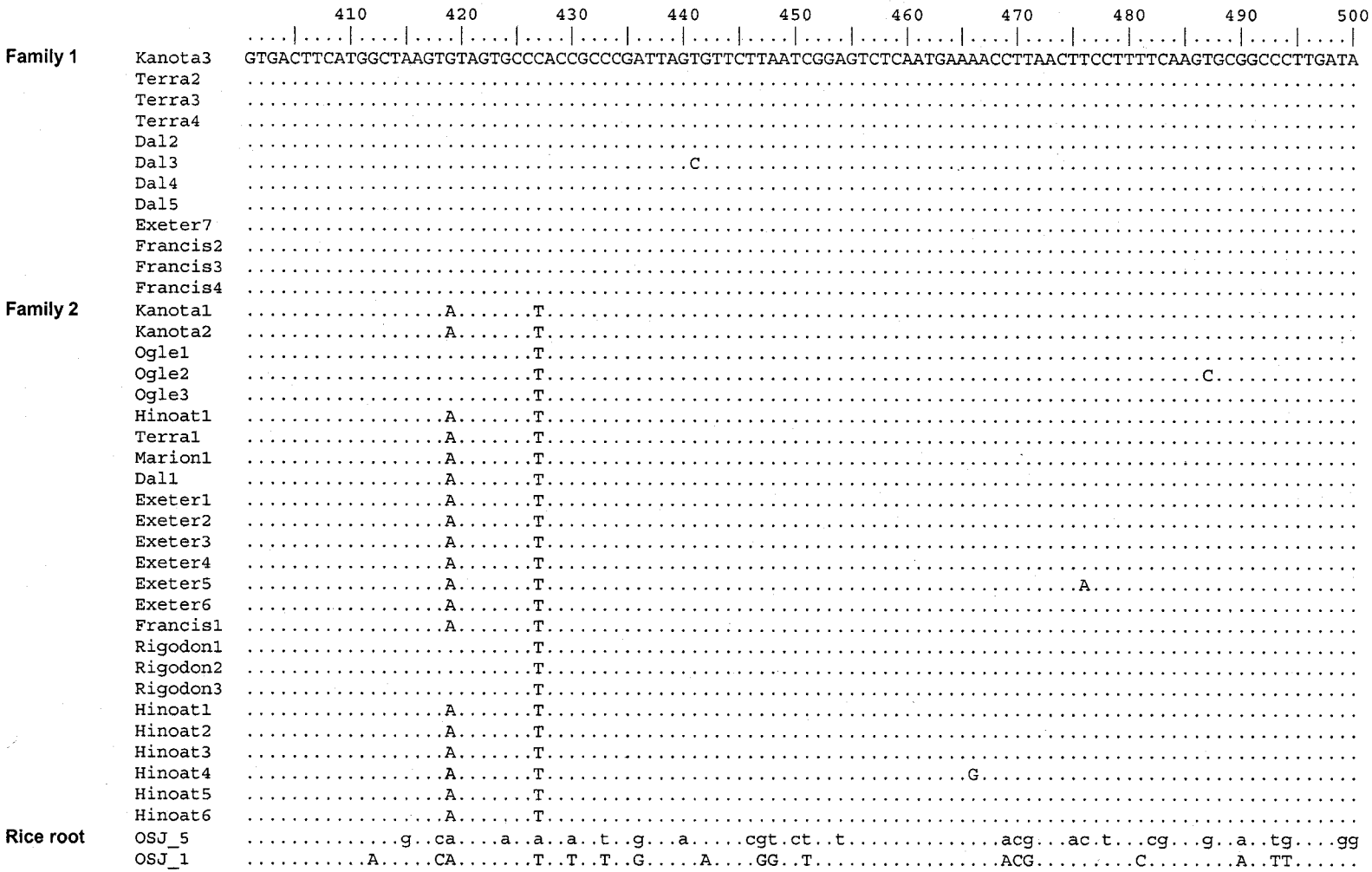

1
2
3
4
ACP (2.7 kb)
238 bp
1
2
3
4
5
6
7
8
9
10
AAT (4.1 kb)
660 bp
1
2
3
4
5
6
7
8
9
10
11
28
29
ACCase (11 kb)
1005 bp
1
2
EF-1 alpha (1.8 kb)
643 bp
1
2
3
4
5
6
16
22
GAS (10.6 kb)
727 bp
1
2
3
4
5
6
7
8
9
10
11
GIS (3.9 kb)
562 bp
1
2
3
4
5
6
7
8
KASI (3.9 kb)
809 bp
1
2
3
4
5
6
7
8
KASIII (3.5 kb)
402 bp

Figure 3.2 Partial alignments showing examples of nucleotide polymorphisms in oat sequences from aspartate aminotransferase and glutamate synthase, two of the eight genes coding for products involved in lipid and protein biosynthesis. Identity to a standard sequence is shown by a dot and a missing nucleotide by a hyphen. Each sequence in these alignments is designated by the name of the cultivar from which the sequence was obtained, followed by a number to designate a particular sequence obtained for that cultivar. Family numbers on the left represent sequence families highlighted by the phylogenetic analysis of these sequences. Only a limited portion of the total alignment is represented for each gene. Only 39 of the 54 oat sequences included in the phylogenetic analysis of aspartate aminotransferase are represented.

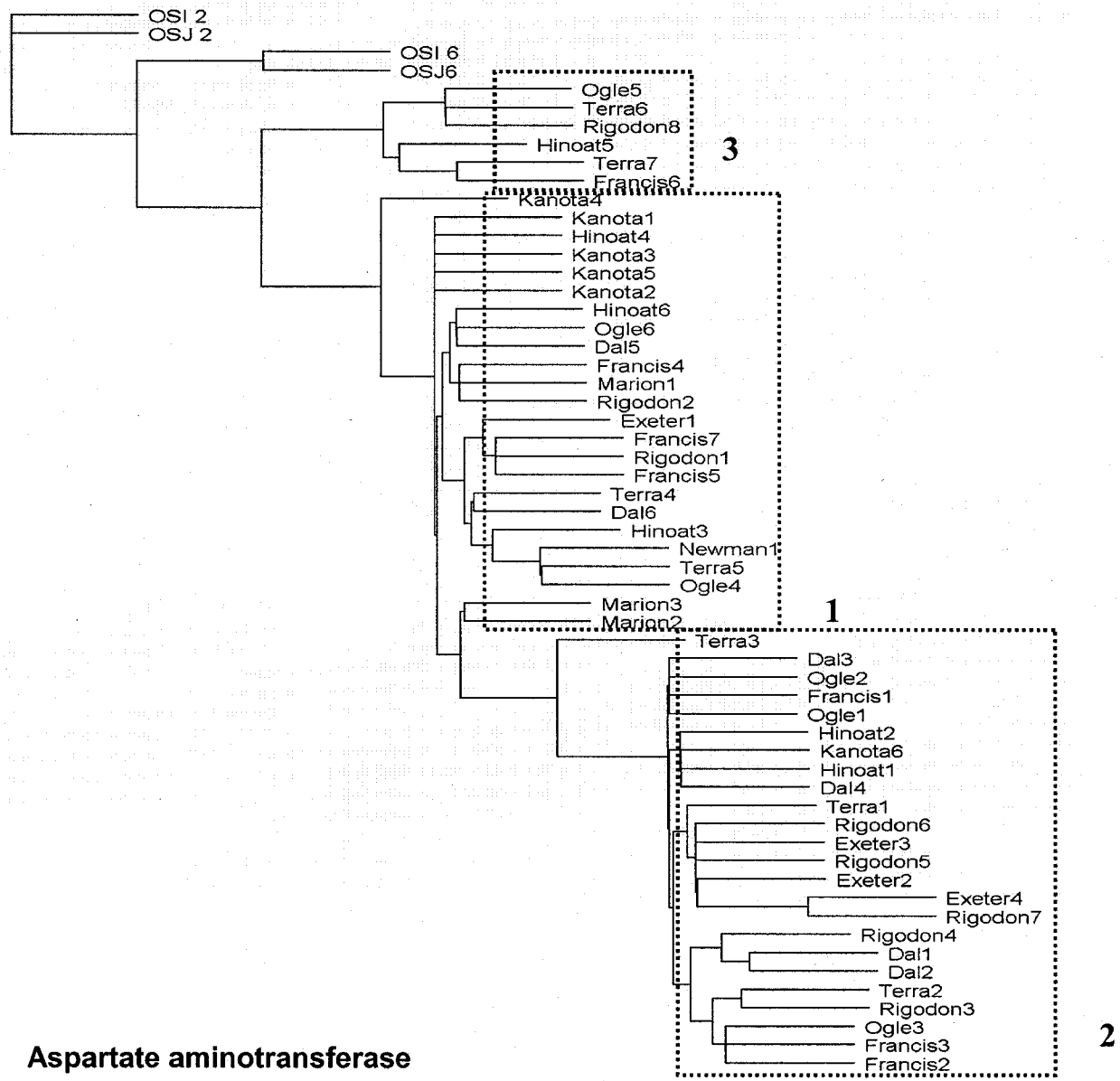


Aspartate aminotransferase (part of intron 6)

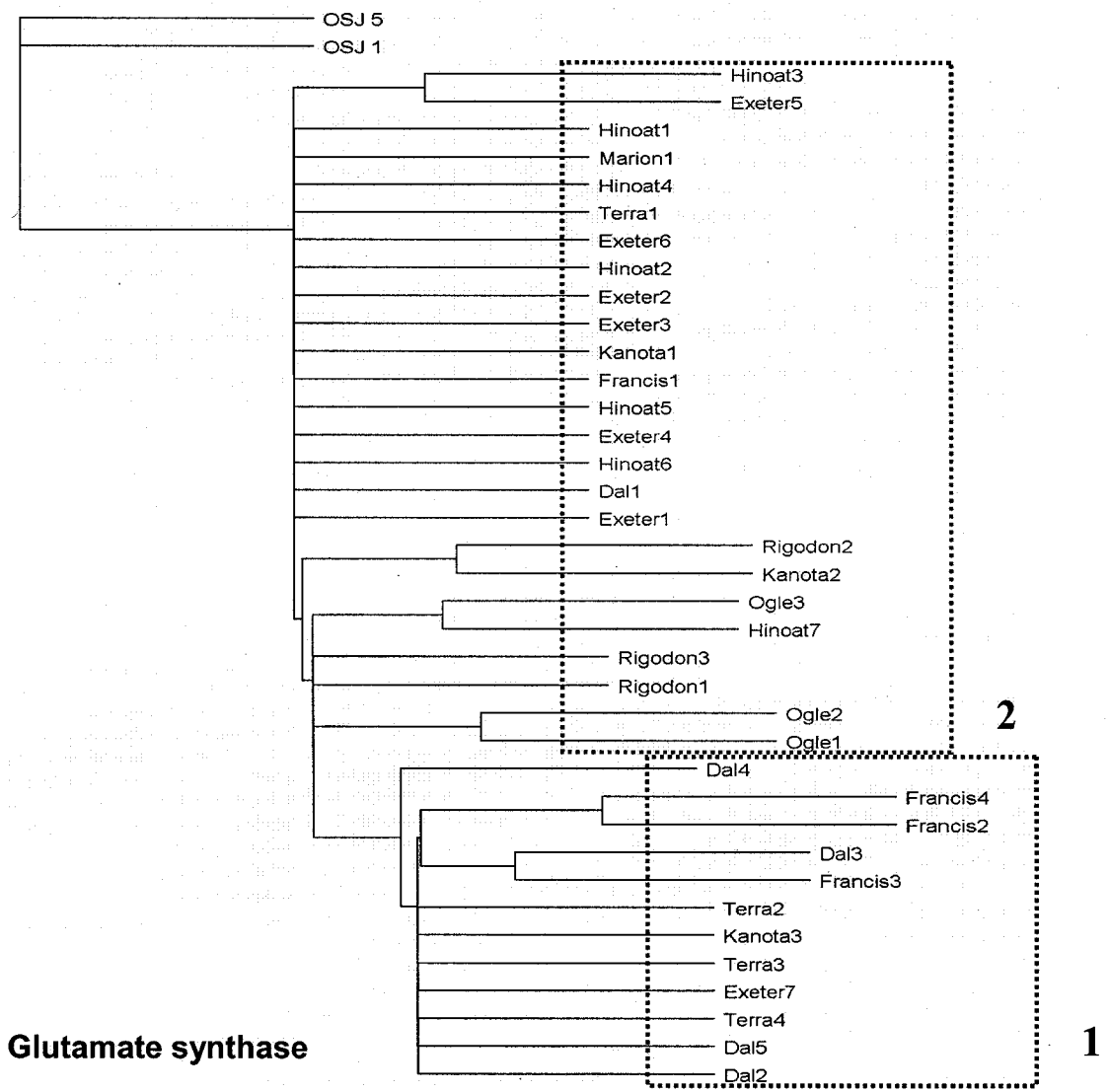


Glutamate synthase (part of exon 16)

Figure 3.3 Phylograms of oat sequences for aspartate aminotransferase and glutamate synthase, two of the eight target genes coding for products involved in lipid and protein biosynthesis. Each sequence in these alignments is designated by the name of the cultivar from which the sequence was obtained, followed by a number to designate a particular sequence obtained for that cultivar. Boxes represent sequence families. OSI and OSJ sequences are *Oryza sativa* sequences, respectively from cultivar groups *indica* and *japonica*, used as outgroups.



Aspartate aminotransferase



Glutamate synthase

Preface to Chapter 4

In Chapter 3, we examined genetic variability at the level of sequence polymorphism among cultivars. Variability between genotypes can also be manifested by variations in gene expression. Most differential gene expression studies deal with differences between treated and untreated material. Few have dealt with variability due to differences in genotypes. The work presented in Chapter 4 describes a survey of genes differentially expressed in kernels of cultivars Kanota and Ogle at the same developmental stage, eight days after the "yellow anther" stage (stage GRO:0007102 according to the Plant Ontology Consortium). Clones for those genes were obtained from reciprocal subtractive suppressive hybridization libraries of Kanota and Ogle.

Sequences will be submitted to GenBank. The manuscript will be submitted to a refereed journal, coauthored by myself, Dr. Nicholas Tinker, Dr. Stephen Molnar and Dr. Diane Mather. I designed the experiment, grew the plants, collected the material, performed the subtractive suppressive hybridization and built the libraries, analysed the sequences and prepared the manuscript. Dr. Nicholas Tinker contributed to the data analysis, provided constructive suggestions and corrected the manuscript. Dr. Stephen Molnar contributed research facilities, detailed suggestions and corrected the manuscript. Dr. Diane Mather supervised the research, provided equipment and funding and corrected the manuscript.

Chapter 4
Genotype-specific gene expression in young oat kernels

4.1 Summary

In order to explore genetic differences between two oat cultivars and to identify new candidate genes for oat kernel quality traits, we surveyed differences in gene expression in young kernels of Kanota and Ogle. Two subtractive suppressive hybridisation (SSH) libraries were obtained from the reciprocal subtractions of young kernel cDNA from cultivars Kanota and Ogle. Good quality sequences from clones of those libraries were assembled in a total of 195 contig sequences exclusive to one or the other library. Most sequences had homology to unidentified sequences or did not have homology to any known sequence When possible, contigs and singletons were grouped in categories based on the biological function of their primary BLAST hit.

Among the identified sequences, differences were observed in the numbers of sequences from each cultivar that belonged to specific functional categories. Several of the sequences expressed specifically in Ogle are related to genes involved in amino acid and storage protein metabolism, whereas several of the sequences specific to Kanota belong to genes implicated in regulation of gene expression or are related to transposable elements.

4.2 Introduction

Chemical composition traits of the oat (*Avena sativa* L.) kernel include lipid, protein and β-glucan content, which are important quality characteristics for the food and feed uses of oat. Kernel lipid and storage proteins start to accumulate in the endosperm and embryo a few days after flowering. By eight days post-anthesis, storage proteins contribute about half of the total endosperm proteins (Peterson and Brinegar 1986). The few days after anthesis are the stage of kernel development where gene expression activity might be expected to be at its highest in the young kernel. At this stage, differences between oat cultivars in timing and dosage of the expression of key genes might have an effect on the chemical composition of the mature kernel.

The Kanota x Ogle recombinant inbred line population was developed from a cross between parents Ogle and Kanota. Ogle is a spring cultivar developed in Illinois, whereas Kanota is a winter oat grown mainly as forage in the southern United States. The Kanota x Ogle map was the first complete hexaploid oat map to be developed and is currently the most complete (O'Donoughue et al., 1995; Wight et al., 2003). It has been used to detect quantitative trait loci (QTLs) for agronomic traits (Siripoonwiwat et al., 1996; Holland et al., 1997), kernel characteristics and chemical composition (Kianian et al., 1999; Kianian et al., 2000; Groh et al., 2001).

To gain better understanding of the path between genotype and phenotype, evaluating transcriptional differences as a source of phenotypic variation between genotypes of a species is a promising strategy. Subtractive suppressive hybridization (SSH) has been used by Botha et al. (2005) to study mechanisms of resistance of wheat (*Triticum aestivum* L.) to Russian wheat aphid (*Diuraphis noxia* Kurdj.), by comparing resistant or susceptible near-isogenic lines. Another study, by Yao et al. (2005) has examined differences in gene expression between wheat/spelt (*Triticum aestivum* subsp. spelta (L.) Thell.) hybrids and their parental inbreds, in order to examine the role of differential gene expression in heterosis. In order to explore genetic differences between two oat cultivars and to identify new potential candidate genes for oat kernel quality traits, we surveyed differences in gene expression in young kernels of Kanota and Ogle.

4.3 Materials and methods

4.3.1 Plant material

The oat cultivars Kanota and Ogle were grown in a growth cabinet (16 h at 20°C, 8 h at 16°C, 16 h of light) in 13 cm pots, with three plants per pot. Each panicle of each plant was harvested eight days after its uppermost floret reached anthesis ("yellow anther" stage, or stage GRO:0007102 according to the Plant Ontology Consortium, http://dev.plantontology.org/docs/growth/growth.html), frozen in liquid nitrogen and kept at -70°C till processing. Each individual floret (caryopsis, lemma and palea) was separated from its rachilla. All florets originating from the same pot were pooled and ground in liquid nitrogen with a mortar and pestle.

4.3.2 RNA extraction, mRNA purification and construction of SSH

RNA was extracted from 0.5 g of frozen ground tissue using the RNAwiz™ isolation reagent (Ambion Inc., Austin, TX), and stored at -70 °C in formaldehyde until further use. Messenger RNA was purified from 500μg of total RNA of Kanota and of Ogle, using Dynabeads® Oligo(dT)25 columns (Dynal Biotech, Oslo, Norway) following manufacturer's instructions. Messenger RNA from each parent was then used alternatively as the tester or the driver to build two reciprocal subtractive suppressive hybridization (SSH) libraries (Diatchenko et al., 1996), using a PCR-Select™ cDNA Subtraction Kit (BD Biosciences, San Jose, CA) following the manufacturer's instructions (http://www.clontech.com/images/pt/PT1117-1.pdf). The cDNA pools were cloned in the pDrive vector using the PCR Cloning Kit (Qiagen Inc., Chatworth, CA), and electroporation in SURE® Competent Cells (Stratagene, La Jolla, CA). The cDNA library derived using Kanota as the tester and Ogle as the driver was named K_minus_O, and the reciprocal library O_minus_K.

4.3.3 Sequencing and sequence analysis

By single-pass sequencing, 833 and 919 sequences were obtained from K_minus_O and O_minus_K respectively, reaching redundancy rates of 71% and 46% based on contig assemblies described below. Sequences were trimmed of vector contamination and low-quality segments and imported into a customized relational database (Tinker and Couroux, unpublished). Contigs were assembled using SeqMan (DNASTAR Inc., Madison, WI) based on overlap-criteria of 30 bases and 90% similarity for introduction into a contig. Further details and considerations of these assembly parameters were discussed by Hattori et al. (2005).

The similarity of singleton and contig sequences to known gene sequences was determined using BLAST (Altschul et al., 1990) from NCBI (http://www.ncbi.nlm.nih.gov/). Subprogram blastn (2.0.14) was used to identify nucleotide similarity, and blastx (2.2.6) was used to identify translated peptide similarity. Gene homology was tentatively declared when an alignment was identified at a global database expectation value of 0.3 and results were verified manually to exclude artefacts. Genes were classified into functional gene categories based on the description of their

primary BLAST hit and according to the Gene Ontology Consortium (2000; http://www.geneontology.org/), and grouped in the following eight groups: storage proteins, repetitive sequences, cellular metabolism (GO:0044237), photosynthesis and photorespiration (GO:0015979 and GO:0009853), rRNA processing (GO:0006364), cytoskeleton organization and biogenesis (GO:0007010), gene expression and signal transduction (GO:0040029 and GO:0007165) and cell cycle and DNA repair (GO:0007049 and GO:0006281).

For contigs that contained sequences from both libraries, CLUSTALDIST (Thompson et al., 1994) was used to estimate distances between sequences from the two cultivars. Gaps were included in the analysis, and no correction was made for multiple substitutions.

4.4 Results and discussion

Good quality sequences of clones from K_minus_O and O_minus_K were assembled in a total of 202 contigs (Table 4.1). Only seven of these contigs contained sequences from both K_minus_O and O_minus_K clones, showing that the overlap between the two libraries was limited, as expected from reciprocal subtractive libraries.

Alignments of those seven contigs show some sequence polymorphism between sequences from Kanota and Ogle. Eighty-two single nucleotide polymorphisms and 10 indels can be identified from those sequences (Table 4.2). Most indels were only one nucleotide long, but an insertion of 26 nucleotides is present in the Ogle version of the sequence corresponding to contig ecas485 (a contig with no significant BLAST hit).

While 76% and 66% of contig consensus sequences from K_minus_O and O_minus_K (respectively) had homology with a known nucleotide or protein sequence, only 31% and 38% of singletons had a BLAST hit. On average, sequences with no BLAST hits were shorter than the sequences that did have a hit (247 nucleotides against 426). A higher percentage (34% vs. 25%) also corresponded to highly variable 3' UTR regions of messenger RNA, as determined by the presence in the sequence of a polyA tail (10 nucleotides or longer).

Contigs and singletons with BLAST hits were grouped in categories based on the biological function of their primary BLAST hit. Differences appear in the gene categories

expressed differentially in each cultivar at this stage of development (Table 4.3). Repetitive sequences, and particularly retrotransposon-related genes, are more frequent in K_minus_O than in O_minus_K. Genes involved in cell metabolism, such as amino acid biosynthesis and glucose metabolism are more frequent in O_minus_K than in K_minus_O. Genes for storage proteins including 11S and 12S globulin are present in O_minus_K, and absent from K_minus_O.

The prominence of transcripts related to repetitive sequences in Kanota compared to Ogle intrigued us at first. Although retrotransposons and other repetitive sequences contribute to a large proportion of plant genomes, they were believed to be transcriptionally silent in plant genomes and activated by stresses or radical changes in the environment (Grandbastien et al., 1997). Nevertheless, a survey of publicly available plant ESTs has showed that retrotransposon genes were present at a frequency of one in 1,000 sequences, and up to 1.75 in 1,000 in grass EST collections (Vicient et al., 2001b). The same team reported active transcription of retrotransposon genes in Triticeae, barley and oat (Vicient et al., 2001a; Vicient et al., 2001b).

The fact that those genes have been isolated from SSH libraries means that they are either expressed in one cultivar and not the other, or expressed in different amounts in young kernels of the two cultivars. It is possible that between four and eight days after anthesis, cells in young kernels of both cultivars are involved in different processes: perhaps with Ogle going through amino acid and storage protein metabolism, but Kanota implicated in regulation of gene expression and in expression of many transposable element-related genes. It is tempting to formulate the hypothesis that those differences in gene expression could result in different biochemical composition of the mature kernel, but another possibility is that after a fixed time following anthesis, Kanota and Ogle are not at the same developmental stage, and the expression of genes involved in gene expression and signal transduction could simply be a stage preceding the expression of storage proteins and carbohydrate metabolism enzymes.

It is nevertheless likely that genes expressed differentially between Kanota and Ogle have specific functions in conferring distinct phenotypes to both cultivars. It is important to notice that 34% of sequences isolated here have homologies to uncharacterized plant sequences, and that another 48% have no homology to known

sequences, which suggests that our approach was effective for identifying transcripts that might encode rare or unknown proteins. As more genes are characterized in other organisms, it will be possible to identify a larger number of the genes differentially expressed between Kanota and Ogle, and gain a better understanding of what mechanisms are involved in each cultivar at this developmental stage.

Table 4.1 Summary of unique oat cDNA sequences and contigs obtained from each of two reciprocal subtractive suppressive hybridization libraries: K_minus_O and O_minus_K

	Subtractive suppression hybridization library	
	K_minus_O	O_minus_K
Total number of contig sequences assembled [a]	109	86
Number (and percentage) of contig sequences with at least one BLAST hit	82 (75%)	57 (66%)
Mean length of contig sequences that had at least one BLAST hit (in nucleotides)	570	465
Total number of singleton sequences	75	144
Number (and percentage) of singleton sequences with at least one BLAST hit	23 (31%)	55 (38.%)
Mean number of nucleotides in singleton sequences that had at least one BLAST hit	426	427
Mean length in singleton sequences that had no BLAST hits (in nucleotides)	255	239

[a] Contigs containing sequences from both libraries were not included.

Table 4.2 Sequence variations and primary BLAST hit of the seven contigs containing sequences from both reciprocal oat subtractive suppressive hybridization libraries K_minus_O and O_minus_K

Contig name	Number of sequence polymorphisms		Distance between sequences [a]	BLAST hit description
	SNPs	Indels		
ecas115	8	2	4.6	*Arabidopsis thaliana* clone 3522 5S ribosomal RNA gene
ecas241	6	0	3.0	*Avena strigosa* Ty1-copia retrotransposon TAS1:GAG, AP and IN
ecas415	1	0	0.4	*Oryza sativa* (*japonica* cultivar-group) cDNA clone:001-104-H01
ecas426	30	0	5.3	*Hordeum vulgare* eIF4E gene, complete sequence.
ecas434	10	1	4.4	Unknown protein [*Oryza sativa* (*japonica* cultivar-group)].
ecas485	19	6	6.0	No significant hit
ecas632	8	1	6.7	No significant hit

[a] Number of differences as percentage of total nucleotides

Table 4.3 Distribution of contigs or singletons from reciprocal oat subtractive suppressive hybridization libraries K_minus_O and O_minus_K in functional gene categories based on best BLAST hits

Functional category [a]	K_minus_O [b]	O_minus_K [b]
rRNA processing	4 (2.1%)	9 (3.9%)
Cytoskeleton organization and biogenesis	1 (0.5%)	0 (0%)
Storage proteins	0 (0%)	6 (2.6%)
Cellular metabolism	3 (1.6%)	17 (7.3%)
Gene expression and signal transduction	10 (5.3%)	11 (4.7%)
Repetitive sequences	9 (4.8%)	2 (0.9%)
Photosynthesis and photorespiration	0 (0%)	5 (2.1%)
Cell cycle and DNA repair	2 (1.1%)	3 (1.3%)
Uncharacterized sequences	79 (42.0%)	61 (26.2%)
No BLAST hit	80 (42.6%)	119 (51.1%)

[a] Based on annotations of matching DNA or protein sequences detected using BLAST (blastn and blastx)
[b] The number of sequences fitting each category is followed by the percentage of total contigs and singletons (in parentheses).

Preface to Chapter 5

After analysing differences in gene expression among Kanota and Ogle, we wanted to survey the segregation of the level of expression of some of those differentially expressed genes. We decided to use macroarrays to perform this survey, which required some optimization of the protocol and the procedures for data analysis. Chapter 5 presents our assessment of the reliability of macroarray analyses. First, we tested the reproducibility of the macroarray printing process, then we tried to find a data transformation that would reduce the variation related to separate hybridizations, such as labelling intensity and exposure time as much as possible.

The manuscript will be submitted as a technique note to a refereed journal, coauthored by myself, Dr. Nicholas Tinker, Dr. Stephen Molnar and Dr. Diane Mather. I designed the experiment, performed the laboratory work, the data analysis and prepared the manuscript. Dr. Nicholas Tinker contributed suggestions and corrected the manuscript. Dr Stephen Molnar contributed research facilities, suggestions and corrected the manuscript. Dr. Diane Mather supervised the research, provided equipment and funding and corrected the manuscript.

Chapter 5
Reliability of DNA-macroarray printing and reduction of the effect of exposure time on spot intensity readings

5.1 Summary

In array experiments, the quality of the array and the methods used to analyse the hybridization data are crucial to obtain meaningful results. In preparation to perform an experiment involving DNA macroarrays, we tested the reliability of the macroarray printing process. We found that the amount of DNA spotted on macroarrays using a 96-pin replicator proved very consistent inside arrays and between replicate arrays. We also tested methods of data transformation applicable to data originating from radioactively labelled arrays. Subtracting the local background from the signal was found to increase the variation between replicate readings, whereas dividing the signal by the intensity of the median array background was found to reduce variation due to exposure time.

5.2 Introduction

DNA arrays are a powerful tool to study the transcriptional regulation of hundreds to thousands of genes simultaneously, and their use has allowed the investigation of gene expression at the level of a cell as well as the level of an organism. A DNA array can be defined as a collection of spots on a matrix, each spot containing a specific DNA fragment. The evaluation of expression levels with the array technique is based on hybridization of nucleic acids, where sequence complementarity leads to the pairing between two single-stranded nucleic acid molecules, one of which is immobilized on the array and the other labelled so as to be detectable by the experimenter.

One of the challenges of array experiments is the difficulty of comparing results originating from different hybridization events. Several experimental variables make this challenging, including differences between arrays due to spotting (or "printing") inconsistencies, differences due to the quality of the cDNA, efficiency of the labelling reaction and reading of the signal. These sources of variation are not (or unlikely to be) related to the genetic variation targeted by the array experiment, and it is therefore

important to remove as much of this variability as possible from the data to improve the accuracy of measurement and to avoid systematic bias.

When conducting an experiment involving microarrays, it is possible to compare a control signal to an experimental signal for each spot (Dudoit and Fridlyand, 2002), but this is not the case with macroarrays for which only one labelling signal (usually radioactivity) is used. Two common standardization methods used with macroarray data are (1) expressing the radioactive signal from each spot as a percentage of the total array signal (Petersohn et al., 2001; Weber and Jung, 2002) and (2) expressing spot signal as the fold difference to one or more normalization control spots, usually housekeeping genes such as actin or ubiquitin (Ji et al., 2003; Zhou et al., 2004). Other normalization methods mentioned in the literature but less commonly used for macroarrays are the global (or centring) normalization, consisting of a subtraction of the median or mean array signal from each spot signal, and the Z score transformation (Cheadle et al., 2003).

Each of these methods is based on questionable assumptions. Expression of the signal as a percentage of the total array signal, centring normalizations and Z score transformation all assume that most genes on the chip are not differentially expressed and that most of the observed variation originates from an "array effect". Therefore, either the total signal on an array (percentage of the total), or the array mean or median (centring normalization), or the array mean and standard deviation (Z score transformation) are considered constant across the cDNA samples tested. These assumptions might not be valid in all experiments, particularly if a majority of clones spotted on the arrays have been pre-selected because they might be differentially expressed between the samples tested. Similarly, the common assumption that "housekeeping" genes such as actin make good expression level standards is questionable when considering samples from genetically distinct individuals.

In this study, we examined the reliability of the macroarray printing process and tested a method of data transformation for data originating from radioactively labelled arrays, consisting of dividing the signal by the intensity of the median array background.

5.3 Material and methods

A 96-pin replicator (VP Scientific Inc., San Diego, CA) was used to print two replicate arrays (Array 1 and Array 2) on BiodyneB nylon membrane (Pall, East Hills, NY). The 864-spot arrays consisted of 96 spots of an undiluted solution of phage *Lambda* DNA labelled by random priming with α-^{32}P, 192 replicate spots of each of three dilutions (1/2, 1/4 and 1/8) of the same DNA and 192 empty spots (Fig. 5.1).

Both arrays 1 and 2 were exposed to a phosphor screen (Molecular Dynamics, Sunnyvale CA) for 1 h, 2 h, 3 h, 4 h, 7 h, 16 h, 20 h and 68 h, and the screens were scanned with a Storm™ 840 PhosphorImager™ (Molecular Dynamics, Sunnyvale CA) at a resolution of 200 μm.

The scanned image files were analyzed using GenePix Pro 5.1 software (Axon, Union City CA), with the colour selection set to "Rainbow" for visualization of intensity levels in artificial colours. The array grid was positioned manually on the image. The area for which the signal was measured for each spot was adjusted manually.

The variables considered for analysis were the spot intensity (I = the average of the median signal intensities for all replicate spots on an array), as well as the spot intensity minus the local background (GenePix® Pro 5.0 User's Guide & Tutorial, 2003) (I-B = the average of the median signal intensities minus the median surrounding background for all replicate spots on an array). Correlation coefficients and coefficients of variations were calculated for these variables.

A transformation called I/E was applied by dividing each value of I by the median intensity of all empty (blank) spots on the array.

5.4 Results

5.4.1 Spotting consistency

Scans of the arrays after different exposure times show a good consistency of the amount of DNA spotted in replicates within arrays (Fig. 5.2). For both I and I-B values, there was good reproducibility of spot intensity between replicate arrays (Fig. 5.3), indicating a good reproducibility of printing between arrays. The correlation coefficients between data from Array 1 and Array 2 are high: $r = 0.96$ for I and $r = 0.95$ for I-B.

When coefficients of variation between replicate positions of the negative control spots are calculated, the coefficients are much higher (195.0% and above) for I-B (Table 5.1) than for I (between 14.8 and 21.1%; Table 5.1). For spots containing labelled DNA, coefficients of variation ranged from 14.0% to 21.2% for I-B (Table 5.1) and from 9.1 to 16.5% for I (Table 5.1). This indicates that subtraction of the background (as evaluated by GenePix) increases the variation between replicate spots. Direct intensity readings (I) were therefore preferred over I-B.

5.4.2 Effect of exposure time

The curves in Fig. 5.4 represent the mean spot intensity for all replicate spots of both arrays plotted against the concentration of the labelled DNA at different exposure times. For the original data (I), the curves corresponding to different array exposure times do not coincide. To obtain an evaluation of the intensity that is as independent as possible from the length of the exposure, it would be desirable to have a transformation that would cause the eight curves to coincide. After dividing I values by the median background value (I/E), all eight curves were approximately coincident except for some divergence at the higher values for the two curves generated from the longest exposure times.

5.5 Discussion

We observed a good reproducibility of direct readings of spot intensity among replicated spots within and across arrays. The variability among replicate readings increased when local background intensity was subtracted from spot intensity, probably because of the influence of neighbouring spots on the measurement of the local background.

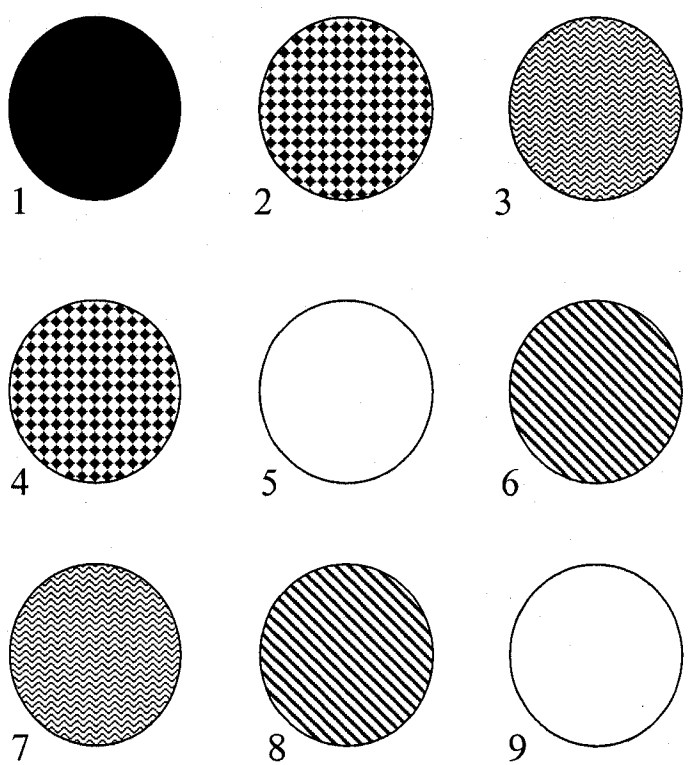
Furthermore, we found that dividing spot intensity by the median background intensity seemed to reduce variation due to exposure time, particularly for low and moderate radioactive signal intensities. For high signal intensities, bleeding of the radioactive signal into the areas designed for background estimation and/or saturation of the screens seem to decrease the precision of the intensity as well as the estimation of the array background.

This transformation method could also prove useful for the analysis of data from experiments other than macroarrays also relying on radioactive hybridization, such as Northern or Southern blots, especially when results from different hybridization events have to be compared and no other appropriate comparison standard is available.

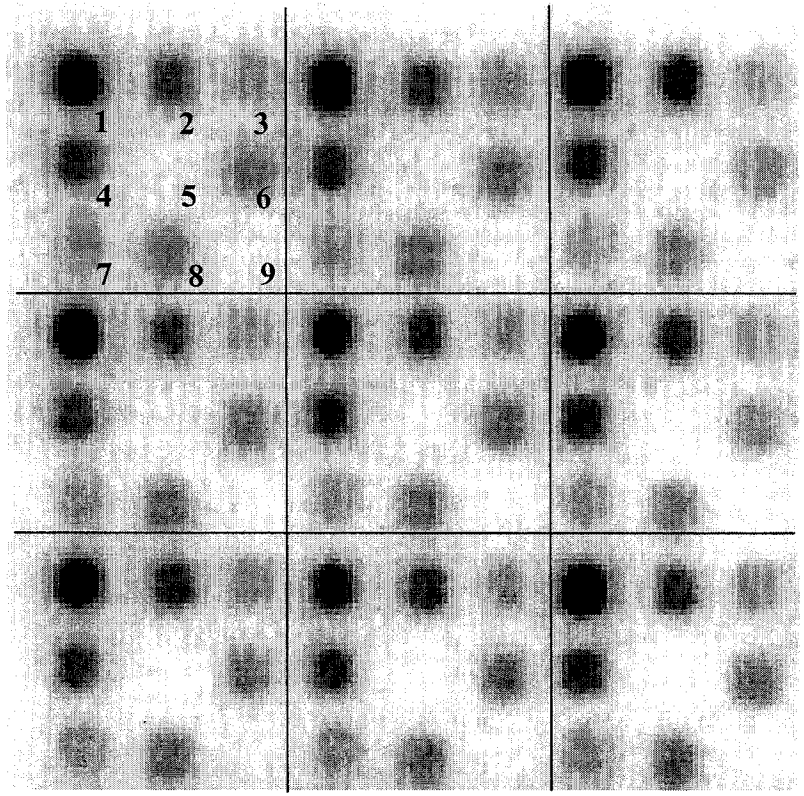
Table 5.1 Coefficients of variation of spot intensity (I) and spot intensity *minus* the local background (I-B) between replicate spots at eight exposure times for arrays printed with a solution of labelled DNA

	Dilution of the DNA solution	Exposure time (h)								
		1	2	3	4	7	16	20	21	68
Coefficient of variation for I-B	0	196.0	195.5	196.5	196.0	195.0	196.7	196.7	196.4	197.4
	1/8	18.7	20.8	16.7	19.9	21.2	16.7	14.7	15.4	18.7
	1/4	17.1	19.0	15.3	17.6	18.9	15.8	14.5	17.1	17.7
	1/2	17.7	19.2	17.3	18.5	17.7	18.7	16.5	20.1	19.0
	1	17.0	18.9	17.2	16.8	17.4	14.4	14.0	16.0	14.7
Coefficient of variation for I	0	21.0	19.6	18.1	19.9	17.9	20.0	17.4	14.8	21.1
	1/8	11.3	13.2	9.9	11.4	12.5	10.5	9.1	11.1	11.4
	1/4	11.7	12.7	10.7	11.8	12.8	10.7	10.1	13.1	12.4
	1/2	14.6	15.3	14.5	15.2	14.5	15.4	13.8	16.5	15.6
	1	15.4	17.0	15.6	15.1	15.6	12.8	12.6	13.9	13.1

Figure 5.1 A. Diagram showing the layout used for each of 96 three-by-three blocks of printing positions making up an 864-spot array. Within each such three-by-three block, seven of the nine positions were used for spots of a ^{32}P-labeled DNA solution, with that solution undiluted in position 1, diluted to 1/2 in positions 2 and 4, diluted to 1/4 in positions 6 and 8 and diluted to 1/8 in positions 3 and 7. The remaining positions (5 and 9) in each block were left empty as negative controls. **B**. Phosphorimaging scan of an 81-spot section (nine three-by-three blocks) of an array after 4 h of exposure.



A



B

Figure 5.2 Phosphorimaging scans of arrays printed with dilutions of a labelled DNA solution after 1h to 68h of exposure to a phosphor screen.

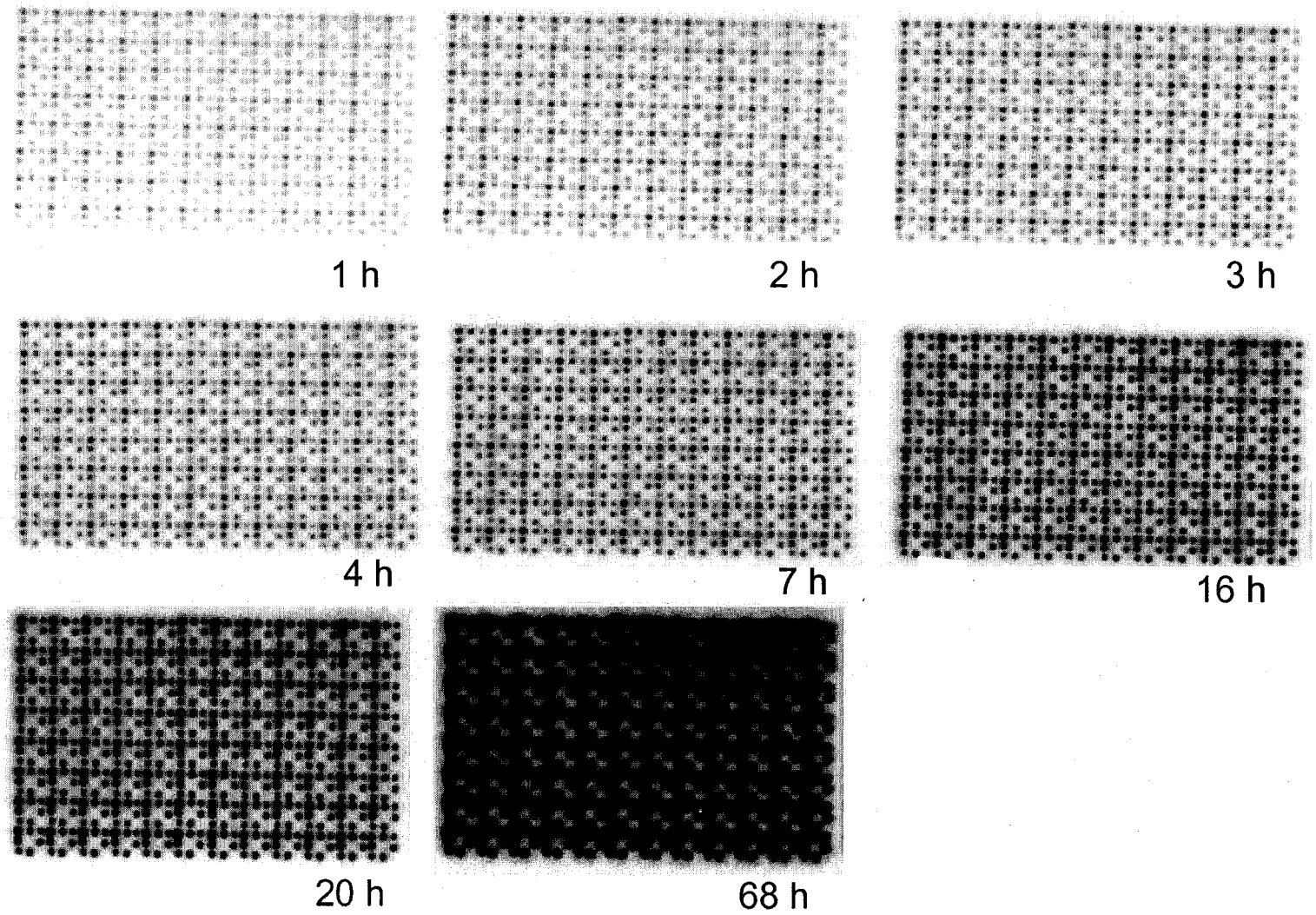

1 h
2 h
3 h
4 h
7 h
16 h
20 h
68 h

Figure 5.3 Plot of replicate values of I and I-B from two sets of replicate arrays (array 1 and array 2) printed with a solution of labelled DNA, after 16h of exposure of the array to a phosphorimaging screen. The x-axis represents the intensity of spots on array 1, and the y-axis the intensity of the corresponding spot on array 2. The two clusters of spots of low intensity (100 and under) correspond to the empty control spots on the array.

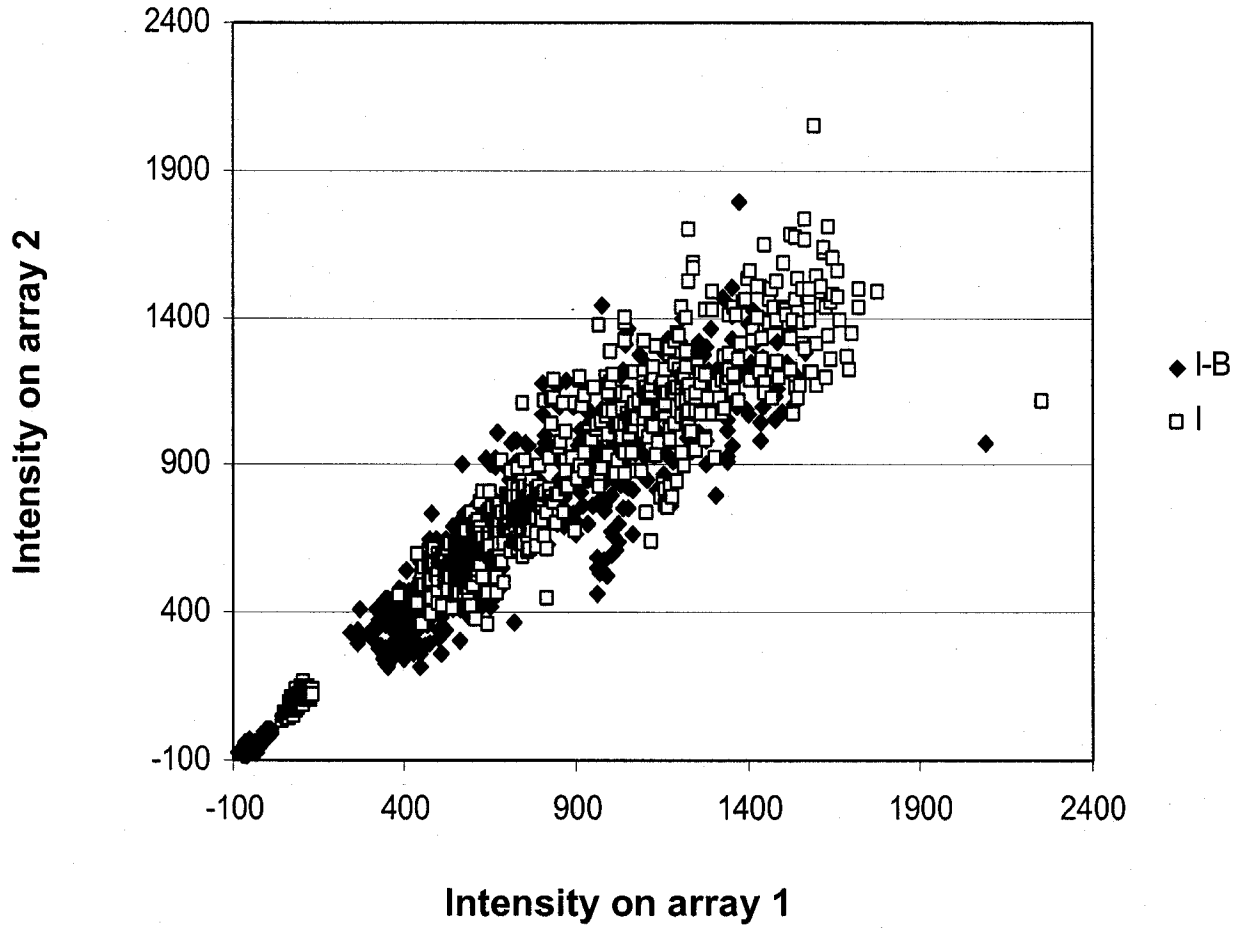
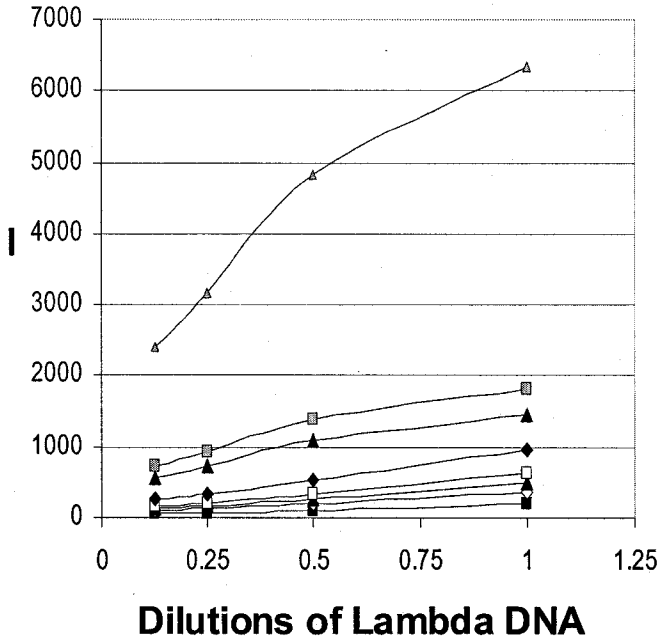
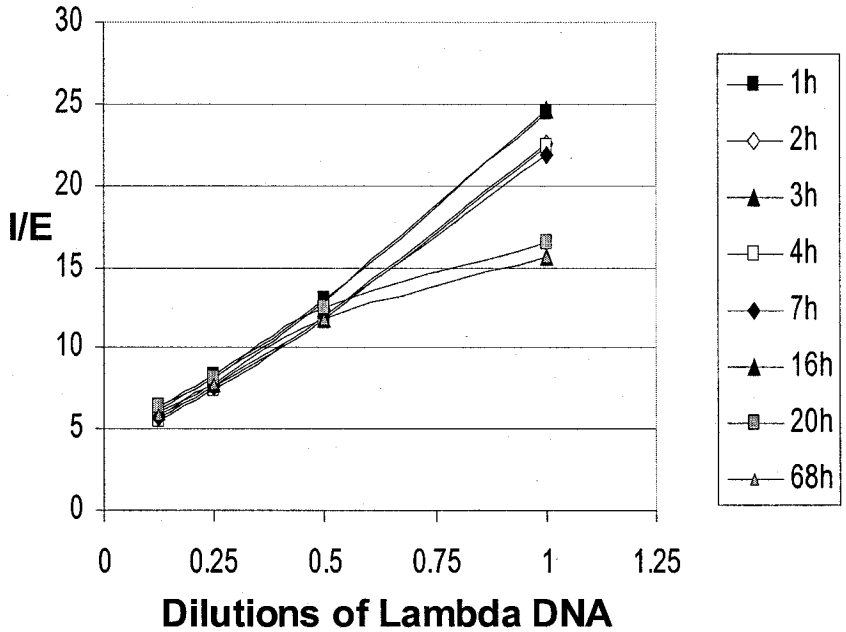

2400
1900
1400
900
400
-100
-100
400
900
1400
1900
2400
Intensity on array 2
Intensity on array 1
I-B
I

Figure 5.4 Intensity of signal at eight exposure times for arrays printed with a solution of labelled DNA. The x-axis represents the concentration of the DNA solution (in dilution factors of the labelled reaction). The y-axes represent the average intensity (I) and average intensity divided by the background (I/E) of spots (average of all replicate spots from both sets of arrays)

.


7000
6000
5000
4000
3000
2000
1000
0
I
0
0.25
0.5
0.75
1
1.25
Dilutions of Lambda DNA


30
25
20
15
10
5
0
I/E
0
0.25
0.5
0.75
1
1.25
Dilutions of Lambda DNA
1h
2h
3h
4h
7h
16h
20h
68h

Preface to Chapter 6

This chapter describes the analysis of gene expression levels in the Kanota x Ogle RIL population using macroarrays, and the detection of expression QTLs (eQTLs). A large part of the results presented in this chapter covers the statistical analysis of the experimental data. This is to emphasize the challenges we met in comparing large datasets deriving from different experimental events, even though some problems were partially solved by the data transformation presented in Chapter 5. The clones spotted on the macroarray came mainly from libraries K_minus_O and O_minus_K described in Chapter 4, but some of the PCR-derived clones described in Chapter 3 were also included.

The manuscript will be submitted to a refereed journal, coauthored by myself, Dr. Nicholas Tinker, Dr. Stephen Molnar and Dr. Diane Mather. I designed the experiment, grew the plants, collected the material, designed the macroarrays, carried out the hybridizations and the image analysis, analysed the data and prepared the manuscript. Dr. Nicholas Tinker contributed to the data analysis, contributed constructive suggestions and corrected the manuscript. Dr. Stephen Molnar contributed research facilities, suggestions and corrected the manuscript. Dr. Diane Mather supervised the research, provided equipment and funding, extensive help with data analysis, contributed detailed and constructive suggestions and corrected the manuscript. Phenotypic data from the Kanota x Ogle population was provided by Dr. Howard Rines.

Chapter 6
Expression level variations in a population of oat recombinant inbred lines

6.1 Summary

In this study, we considered gene expression levels as quantitative traits potentially influenced by genetic factors segregating in the Kanota x Ogle oat mapping population. We designed a macroarray featuring 288 oat clones, most of which had been isolated from a previous experiment designed to capture transcripts that were differentially expressed between young kernels of Kanota and Ogle. Two replicate sets of these arrays were hybridized with cDNA from RILs from the Kanota x Ogle population. Four methods of standardizing signal intensity (I) were tested: a Z-score transformation of I, I as a percentage of the total array signal, I relative to that of an actin gene, and I divided by the median array background. Division by the median array background gave the best concordance between results of replicate arrays. With this transformation, 63.9% of clones showed a significant variation among cDNA samples, and 33 significant eQTL peaks were detected at $P < 0.05$. Most of these eQTLs cluster to one locus on KO linkage group 29_43, constituting an apparent "hot-spot" for the regulation of gene expression.

6.2 Introduction

Many phenotypic traits in plants and animals exhibit quantitative variation. Underlying this variation, there may be genetic and environmental factors interacting via complex networks to influence the transcriptome, proteome and metabolome. Despite significant progress in knowledge of the physiology, biochemistry and molecular biology of plants and animals, large gaps remain in our understanding of the path between genotype and phenotype.

Several studies have examined the genetic control of quantitative variation in the transcriptome and proteome. This type of analysis was pioneered by Damerval et al. (1994), who mapped loci affecting quantitative variation of 72 proteins in an F_2 population of maize (*Zea mays* L.). Seventy protein quantity loci (PQLs) were mapped for these proteins, and up to 12 chromosomal regions seemed to affect the amounts of

individual proteins, indicating that the regulatory systems involved might be complex. Similar work has been conducted since, based on transcript abundance of a few candidate genes at a time using Northern hybridizations (Consoli et al., 2002), or using microarrays to investigate thousands of genes simultaneously in fruit fly (*Drosophila melanogaster*), yeast (*Saccharomyces cerevisiae*), mouse (*Mus musculus*), maize, Arabidopsis (*Arabidopsis thaliana* L.), human (*Homo sapiens*) and rat (*Rattus norvegicus*) (Wayne and McIntyre 2002; Schadt et al., 2003; Yvert et al., 2003; Bystrykh et al., 2005; Chesler et al., 2005; DeCook et al., 2005; Hubner et al., 2005). Results show that mapping of loci affecting transcript abundance can be a successful strategy for dissecting complex traits and identifying candidate genes. By screening microarrays of mouse and maize genes, (Schadt et al., 2003) identified significant expression QTLs (eQTLs) for 16% of the mouse genes, and 34% of maize leaf tissue genes tested.

Few of the eQTLs that have been identified co-locate with the map positions of the genes whose expression levels were considered. This, together with the fact that few causative mutations have been identified in candidate genes for human obesity (Chagnon et al., 2003), led Pomp et al. (2004) to propose the hypothesis that most genes regulating the inheritance of complex traits act in *trans* on the primary physiological pathways involved. Mapping of eQTLs or PQLs might allow us to estimate the chromosomal positions and effects of these *trans*-factors.

Molecular markers have been used to construct genetic maps in many crop species, on which genes and QTLs have been positioned (Kearsey and Farquhar, 1998). Plant improvement relies on the accumulation of desired alleles for economically important traits. As many of those traits exhibit quantitative variation, molecular markers are also useful to trace desirable alleles in segregating populations.

In oat (*Avena sativa* L.), molecular marker maps have been established for several populations (O'Donoughue et al., 1992; Kianian et al., 1999; Zhu and Kaeppler, 2003a; Wight et al., 2003; De Koeyer et al., 2004; Portyanko et al., 2005; Rayapati et al., 2006), but the Kanota x Ogle map is currently the most complete (O'Donoughue et al., 1995; Wight et al., 2003). QTL studies have been carried out in the populations Kanota x Ogle (Kianian et al., 1999; Groh et al., 2001), Kanota x Marion (Kianian et al., 1999 and 2000, Groh et al., 2001), Terra x Marion (De Koeyer et al., 2004), Ogle x TAM O-301 (Holland

et al., 2002) and Ogle x MAM 17-5 (Zhu and Kaeppler 2003b; Zhu et al., 2003c; Zhu et al., 2003d; Zhu et al., 2004). Published QTLs for each population include loci affecting yield, plant height, maturity, kernel morphology and disease resistance, as well as grain quality traits such as lipid, protein and β-glucan contents (Kianian et al., 1999 and 2000, Groh et al., 2001, Wight et al., 2003, Zhu and Kaeppler, 2003b, Zhu et al., 2003c, 2003d and 2004, De Koeyer et al., 2004).

Unfortunately, few genes have yet been positioned on oat maps, making it difficult to assess whether candidate genes co-locate with QTLs for the corresponding traits. In the research reported in Chapter 4 of this thesis, clones were obtained for oat genes that were differentially expressed between Kanota and Ogle early in kernel development. Here, macroarrays featuring those oat clones were used, along with labelled cDNA from Kanota x Ogle recombinant inbred lines (RILs), to attempt to detect and map eQTLs.

6.3 Materials and methods

6.3.1 Plant material and RNA extraction

The oat cultivars Kanota and Ogle, and a random sample of 72 $F_{9:10}$ RILs derived from the Kanota x Ogle population (O'Donoughue et al., 1995, Wight et al., 2003), were grown in a growth cabinet (16 h at 20°C, 8 h at 16°C, 16 h of light) in 13 cm pots, with three plants per pot. One pot of each RIL and two replicate pots of each parent were placed in a completely randomized arrangement within the cabinet. Each panicle of each plant was harvested eight days after its uppermost floret reached anthesis ("yellow anther" stage, or stage GRO:0007102 according to the Plant Ontology Consortium (http://dev.plantontology.org/docs/growth/growth.html)), frozen in liquid nitrogen and kept at -70°C till processing. Each individual floret (caryopsis, lemma and palea) was separated from its rachilla. All florets originating from the same pot were pooled and ground in liquid nitrogen with a mortar and pestle.

RNA was extracted from 0.5 g of frozen ground tissue using the RNAwiz™ isolation reagent (Ambion Inc., Austin, TX), and stored at -70°C in formaldehyde until further use.

6.3.2 Array design and printing

Two hundred and seven clones were selected from two SSH libraries (K_minus_O and O_minus_K) to represent all of the non-redundant BLAST hits obtained (see Chapter 4 of this thesis). Eighty-one clones derived from PCR amplification of oat genomic DNA with primers targeting candidate genes encoding proteins involved in kernel lipid and protein biosynthesis (see Chapter 3) were added. A wheat (*Triticum aestivum* L.) actin clone (GI:22303396) provided by Dr. Therese Ouellet (Agriculture and Agri-Food Canada, Ottawa, Canada) was used as an internal control. A fragment of a gene coding for human nebulin, a large actin-binding protein not showing nucleotide homology to any known plant gene, was used as an external control (spike).

Two sets of five 96-well plates (sets 1 and 2) were used for PCR amplifications and each of these sets was later used to print a set of 76 864-spot macroarrays. Within each set of five plates, one plate was used for the wheat actin clone, with that clone assigned to all 96 wells of that plate. Similarly, one plate in each set was used for the human nebulin clone. The 288 oat clones acting as probes were randomly assigned to the 288 individual wells of the remaining three plates of each set. PCR amplifications were conducted using M13 primers and 4 μL of heat-lysed bacterial culture of the appropriate clone in each well. PCR products were denatured by adding one volume of 0.4 N NaOH to the reaction after amplification. Five microliters of a concentrated solution of xylene cyanol were added to each well as a visual aid for printing.

A 96-pin replicator (VP Scientific Inc., San Diego, CA) was used to print arrays of 864 positions onto 10 cm x 11 cm Biodyne B membranes (Pall, East Hills, NY). The 864-position array consisted of 96 three-by-three blocks of nine spots, corresponding to nine possible printing positions of the replicator (Fig. 6.1). Of the nine positions, one was left blank, one was used to print the 96 replicates of the actin clone, one was used to print the 96 replicates of the nebulin clone, and six were used to print the oat clones in two replicate spots. Thus, there were two sets of 76 membranes (sets 1 and 2, corresponding to the two sets of 96-well plates) and each membrane contained 96 replicates of the internal (actin) control, 96 replicates of the external (nebulin) control, 96 replicates of the negative (blank) control and two replicates of each of 288 oat clones. Each nine-position

block within each array contained one internal control, one external control, one negative control and two replicates of each of three oat clones.

Printed arrays were scanned prior to hybridization and printing irregularities (incomplete or partial spotting, presence of impurities or scratches on the membrane), made visible by the presence of xylene cyanol, were recorded.

6.3.3 cDNA synthesis, labelling and hybridization

For each RIL and for each replicate of Kanota and Ogle, first-strand cDNA was synthesized from 2 μg of total RNA using the Omniscript RT Kit (Qiagen Inc., Chatworth, CA). Second-strand synthesis was performed as described by Sambrook and Russell (2001). The quality of the cDNA was tested by PCR using primers located in the 3' end of an actin gene. The cDNA samples were labelled by random-priming using both α-dATP^{32} and α-dCTP^{32} (Sambrook and Russell, 2001). Hybridization was performed at 65 °C as described by Wight et al., (2003) with the following changes: use of a hybridization oven and hybridization bottles, with 10 mL of hybridization buffer and addition of a rinse with a solution of 2x SSC buffer (0.3 M NaCl, 0.03 M sodium acetate) and 2% (w/v) SDS and addition of a 30 min wash with a solution of 0.5x SSC and 0.5% (w/v) SDS. Each of the cDNA samples was incubated overnight with one array from each set. Hybridized and washed arrays were then exposed to phosphor screens (Molecular Dynamics, Sunnyvale CA) until a sharp image was obtained, with as little saturation as possible for the stronger spots (2 h to 48 h according to the intensity of the labelling). The screens were scanned with a Storm™ 840 PhosphorImager™ (Molecular Dynamics, Sunnyvale CA) at a resolution of 200 µm. Samples were prepared and arrays hybridized in random order.

6.3.4 Image analysis

The scanned image files were analyzed using GenePix Pro 5.1 software (Axon, Union City CA), with the colour selection set to "Rainbow" for visualization of intensity levels in artificial colours. The array grid was positioned manually on the image. The area for which the signal was measured for each spot was adjusted manually. Spots presenting printing anomalies or high surrounding background were flagged and excluded from

further analysis. Pixel intensity data were then exported and manipulated in spreadsheets. The GenePix software allows for adjustment of intensity values for local background intensity but this adjustment was not used here as it has been found to increase the coefficient of variation among replicate spots, probably due to bleeding from neighbouring spots (see Chapter 5 of this thesis).

6.3.5 Data analysis

The variable considered for analysis was median signal intensity (I) of all pixels belonging to a given spot. Values of I were averaged over replicate spots on the array. Statistical analyses were applied to I and to four types of transformed data: I divided by the total array signal including background (I/T); the deviation of I from the array mean, divided by the standard deviation of the array (I_S); I divided by the average intensity of actin spots on the array (I/A); I divided by the background intensity of the array (I/E) where E was derived from the average of all readings originating from blank spots.

Analysis of variance, t-tests and calculation of correlation coefficients were done using PROC GLM, PROC TTEST and PROC CORR of SAS version 8.2 (SAS Institute Inc., Cary NC).

6.3.6 Genetic map

A linkage map was constructed using G-Mendel Win32 version 0.8b (Holloway and Knapp, 1993), based on genotypic data for 286 molecular marker loci from the Kanota x Ogle linkage map published by Wight et al. (2003) and 23 additional loci that had each been scored on the RILs used in this study. In selecting marker loci for inclusion on the linkage map, consideration was given to their estimated genomic positions, the number of lines on which they had been scored (favouring markers that had been scored on at least 48 of the 72 RILs for which expression data were collected) and absence of segregation distortion (favouring markers for which the genetic ratio was near 1:1). The map was constructed as described in Wight et al. (2003). In cases where marker order was ambiguous, the order was kept consistent with the map published by Wight et al. (2003).

6.3.7 QTL mapping

Analysis was conducted on data from array set 1, array set 2 and the average values of the two sets. Simple interval mapping was conducted using NQTL software (Tinker and Mather, 1995) with a walking speed of 5 cM and with significance thresholds set by permutation to provide expected genome-wide probabilities of type I error of 0.10, 0.05 and 0.01. One thousand data permutations were performed to set the significance thresholds. Phenotypic input files were prepared using a custom-made Perl script and fed to NQTL in batch mode.

Phenotypic data from the Kanota x Ogle population as described in Kianian et al. (1999 and 2000) and Groh et al. (2001) were provided by Dr. Howard Rines (USDA-ARS, University of Minnesota, St. Paul, MN) for the following traits: mean kernel width as determined by image analysis (Dmin), flow injection analysis (FIA) of β-glucan content, near infrared (NIR) spectroscopy prediction of oil by acid-hydrolysis, groat percent of the kernel, kernel plumpness as assessed by image analysis (100*Fshape), number of days from planting to 50% panicle emergence, lodging severity and rating for presence of tertiary kernels.

6.4 Results

6.4.1 Correlation coefficients between replicate arrays

The overall correlation between corresponding spots on replicate arrays (r) was computed across all 288 oat clones and 76 cDNA samples (21,888 values, minus some missing spots). This overall correlation coefficient was positive but low for untransformed I values ($r = 0.25$, $P < 0.05$) and somewhat higher for each of the four sets of transformed values ($r = 0.33$ for I/T, $r = 0.35$ for I_S, $r = 0.42$ for I/A and $r = 0.49$ for I/E).

When the correlation coefficients between corresponding spots on the two sets of arrays were calculated separately for each of the 76 cDNA samples, the correlation coefficients of intensity values between corresponding spots on the two sets of arrays (r_{lines}) ranged from 0.04 to 0.64 ($P < 0.05$ for all but one coefficient), with a median correlation coefficient of 0.35 (Table 6.1, Fig. 6.2).

When the correlations between corresponding spots on the two sets of arrays were calculated separately for each of the 288 oat clones (r_{clones}), correlation coefficients ranged from -0.21 to 0.87 for I, -0.70 to 0.76 for I/T, -0.73 to 0.96 for I_S, -0.79 to 0.96 for I/A and -0.47 to 0.88 for I/E. All four transformation methods again increased the correlation between the two sets of data. The I/E transformation gave the strongest positive correlations (Table 6.2, Fig. 6.3), with 223 of the 288 clones exhibiting significant ($P < 0.05$) positive correlation between the two sets of arrays, compared to only 39 for the non-transformed values. Clones showing negative or non-significant positive correlations for I/E all had low intensity values (mean I/E across the two array sets of 5.3 or less) (Fig. 6.4).

6.4.2 Effects of cDNA sample, array set and residual effect

With the cDNA sample and the array set considered as sources of variation, only 10.5% of clones showed a significant variation among cDNA samples in a two-way ANOVA on untransformed data. All four transformations increased the number of clones showing a significant effect of the cDNA sample. The highest percentage of significant clones (63.9%) was obtained for I/E transformed data, which also gave the highest median value of F (Table 6.3). The I/E transformation method was therefore selected over the other three methods for further analysis.

6.4.3 QTL scans

Thirty-three significant eQTL peaks were detected based on analysis of averaged I/E transformed data, at $P < 0.05$, and 11 of these peaks were significant even at $P < 0.01$. Sixteen of these peaks were also detected in analyses of set 1 or set 2 data only, with two of these detected in both set 1 and set 2. Among those sixteen peaks, 10 peaks had a higher ratio of test statistic by threshold of significance in the averaged data than in either of set 1 and set 2. Clones for which a peak significant at $P < 0.05$ in averaged I/E data was detected are listed in Table 6.5.

Twenty-seven of the 33 significant peaks detected co-located to KO linkage group 29_43; three co-located to unlinked marker ac06.625, two to unlinked marker UMN110,

and one to linkage group 6 (for marker and linkage group information see Wight et al., 2003).

The eQTLs significant at $P < 0.05$ explained between 7% and 27% of the phenotypic variation (Table 6.6), with a median of 19%. One clone (combo6J05_JF557_023 at position G7_Position2) showed two significant peaks, one on linkage group 6 and one at unlinked marker UMN110. For all but one peak, Kanota contributed the positive allele (Table 6.6). The only peak for which Ogle contributed the positive allele was obtained for clone combo6J05_JF557_023 (array position G7_Position2), and mapped to linkage group 6.

Among the 33 clones showing a significant TS peaks at $P < 0.05$ in the averaged data, three also showed a significant expression level difference ($P < 0.05$) between parents Kanota and Ogle (Table 6.6).

When correlations between I/E values for clones showing significant eQTLs at $P < 0.05$ and phenotypic data for the 72 RILs were tested, several weak but significant correlations were found. Traits showing significant correlations are mean kernel width (D min), length (D max) and plumpness (100*Fshape) as determined by image analysis, β-glucan content, percentage of groat, test weight, heading date, lodging, and tertiary kernels rating (Table 6.7).

No significant epistatic effect of linkage group 29_43 on other loci of the genome was detected for clones showing a prominent peak on that linkage group.

6.5 Discussion

The goal of this study was to try to identify quantitative trait loci for gene expression level in the Kanota x Ogle population. For this purpose, we designed two replicate sets of macroarrays featuring 288 oat clones, most of which were selected because they were differentially expressed in young kernels of Kanota and Ogle. These arrays were hybridized with cDNA from 72 progeny RILs from the Kanota x Ogle population. We detected 33 eQTLs at $P < 0.05$ which is more than twice what could be expected by random chance.

Number of progeny lines

As in most QTL studies the question arose as to how many progeny lines should be included, since the procedure, from collecting material to cDNA synthesis and Southern hybridization, is labour-intensive and costly. Originally, our experiment included 112 lines of the extended Kanota x Ogle RI population (described in Wight et al., 2003) and four biological replicates of each parent. As the experiment proceeded, samples had to be dropped due to poor RNA or cDNA quality, or weak ^{32}P-labelling efficiency, leaving only 72 RILs for eQTL detection. This population size is smaller than the 137 lines that have been used for phenotypic QTL studies in Kanota x Ogle (Kianian et al., 1999 and 2000, Groh et al., 2001), but similar to the 76 F_{2-3} maize population used for eQTL detection in Schadt et al. (2003), and twice as large as the rodent populations used for eQTL detection in Chesler et al., Bystrykh et al. and Hübner et al. (2005) and the population of 30 *A. thaliana* RILs used in DeCook et al. (2005). Nevertheless, the small size of the population used in the current study means that the power of QTL detection was low. The effect of eQTLs on the variation of expression level of the corresponding gene, up to 27%, is also probably overestimated (Melchinger et al., 1998).

Data standardization

Another challenge this work raised was the choice of a standardization method. It was necessary to select a method to standardize the data that would remove as much systematic variation among arrays as possible, in order to allow the comparison of data originating from different hybridization events. As only one labelling signal (usually radioactivity) is used in macroarrays, it is not possible to compare a control signal to an experimental signal for each spot, as it is done for microarrays (Dudoit and Fridlyand, 2002). We reviewed several standardization methods that are commonly used with macroarrays and microarray data (see part 5.2 of this thesis). Rather than simply adopting one of these, we tested three of them: the Z score transformation (I_S), signal intensity expressed as percent of the total array signal (I/T), and an actin gene as an expression control (I/A). A fourth data transformation consisting of dividing each spot signal by the median background of blank control spots (I/E) was also tested. This transformation does not account for the effect of cDNA sample quality or labelling strength, but we found in a

previous optimization experiment that it was an effective way to remove the effect of differences in array exposure time to phosphorimaging cassettes (see Chapter 5 of this thesis).

This last transformation (I/E) was ultimately selected for further analysis of the data. It gave the best concordance between replicate array results, the highest correlation coefficients and the most clones showing a positive and significant correlation among data from the replicate arrays. It is also in the I/E data that the highest percentage of array clones showing a significant difference among cDNA samples was obtained.

Sensitivity of the procedure to capture expression level variations

The fact that 63.9% of clones showed a significant effect of the cDNA sample (Table 6.3) indicates that this protocol is sensitive enough to detect some non-random variation in gene expression levels. Nevertheless, few clones showed a significant difference of expression level between the two parents (Table 6.6 and data not shown). This number would have been expected to be high, as about 72% of spots on the array (excluding control spots) originate from subtractive libraries built to select genes differentially expressed between Kanota and Ogle. This low number could be caused by the technical difficulty of detecting small changes in expression level of many different genes with a protocol based on hybridization, as discussed by Xu (2005). A further factor to consider is that hybridization analysis is most sensitive and accurate for the detection of transcripts with high abundance, whereas the SSH libraries may have successfully isolated many transcripts with low relative abundance. Another possible cause of this low level of significant differences among the parents is the fact that the t-tests were based on only two replicates for each parent, and were therefore not very powerful. The QTL analysis has a higher power of detection, as it exploits replications of allelic classes among progeny lines (with an expectation of 36 replicates of each parental type among the 72 lines).

Detection of eQTLs

In this study, gene expression levels have been considered as quantitative traits potentially influenced by genetic factors segregating in the Kanota x Ogle population.

Previous work in this population has shown that other quantitative traits are affected by genetic factors (Kianian et al., 1999 and 2000, Groh et al., 2001), and it is reasonable to assume that some of these traits may be influenced by the regulation of transcription of certain genes. Prior to conducting this experiment, we formulated the hypothesis that eQTLs might be found that co-located with other QTLs previously detected in this population. These loci might coincide with the actual locations of the structural genes whose expression was quantified, or might be located elsewhere in the genome, acting in *trans* on the level of expression of the studied genes. Knowing the genes whose transcription is affected by loci where QTLs and eQTLs co-locate could lead to hypotheses about genetic mechanisms of QTL effects. However, the eQTLs detected in this study are largely independent from the locations of previously detected QTLs.

Potential reasons that previously detected QTLs are not related to the changes in transcriptional activity measured in this experiment are that these changes have little influence on the traits that have been measured, or that experiments have not been powerful enough to detect coinciding events. The first of these explanations seems plausible: the growth stage at which transcription levels have been investigated may not have a fundamental influence on other quantitative traits (mainly traits of economic importance, estimated on mature kernels and plants) that have been studied in this population. The issue of experimental power is also important. The parents of this population could have many hundreds of genetic differences that are capable of influencing transcription, but only some of these would be detected.

Nevertheless, significant correlations have been found between the level of expression of several genes and phenotypic data for nine traits measured in the Kanota x Ogle RIL population. Only one of these nine traits showed a significant QTL (heading date, on linkage group 7_10_28) when screened over the 72 RIL lines included in this study. QTLs for most of these traits have previously been reported based on analysis of a larger set of lines from this population (Kianian et al., 1999 and 2000, Groh et al., 2001), and might simply not be detectable in this subset of 72 lines.

A positive correlation between the expression level of a gene and a phenotype could indicate that the gene is implicated in regulating the trait, but also that they might be influenced by a common genetic factor. It is interesting to notice that the expression of

an acetyl-CoA carboxylase seems positively correlated to kernel plumpness and negatively to kernel length (at $P < 0.01$). An acetyl-CoA carboxylase gene maps to linkage group 11-41, and is linked to a major oil QTL in Kanota x Marion (Kianian et al., 1999), but also to an overlapping QTL for kernel area (Groh et al., 2001). Those are two indications that acetyl-CoA carboxylase might be implicated in the determination of kernel shape in oat.

Co-location of eQTLs on linkage group 29_43

One eQTL location on linkage group 29_43 was detected more consistently, and for more genes than any other. Schadt et al. (2003) reported a similar phenomenon in mouse, where eQTLs clustered to seven "eQTL hot-spots". DeCook et al. (2005) similarly reported five eQTL hot-spots in a study involving Arabidopsis RILs from a Columbia x Landsberg erecta cross, two of the hot spots coinciding with QTLs influencing shoot regeneration suggesting that some of the heritable gene expression changes observed were related to differences in shoot regeneration efficiency between ecotypes.

Six cDNA clones have been previously mapped to linkage group 29_43 (Wight et al., 2003). Two of them (UMN360 and UMN856) have similarities to puroindoline (a grain softness protein in wheat), and one to a glucosamine-fructose 6-R aminotransferase, an enzyme involved in amino-sugar synthesis and glutamate metabolism. The other three clones don't show any homology to identified sequences. Among the 27 clones for which an eQTL was detected on group 29_43 one had sequence similarities to an alanine aminotransferase, an enzyme involved in amino acid biosynthesis (see Table 6.6). Four genes connected to protein biosynthesis seem thus linked to group 29_43, but no QTL related to the protein content of kernels has been detected on that linkage group.

As 16 of the 33 clones for which an eQTL was detected had no significant similarity to characterized sequences, it is difficult to establish if other patterns of relationship exist among the genes affected by this segregating region. It nevertheless seems possible that this region contains one or more transcriptional regulators that affect many genes. No major developmental QTL having been found in this region, it is also possible that these differences are related to a genetic event that affects the timing of

transcriptional regulation. For example, many recent studies have shown that plant genes are regulated in response to circadian rhythms (Blasing et al., 2005; Kim et al., 2005; Turner et al., 2005), and it is possible that a single event segregating in Kanota x Ogle could affect this timing.

Table 6.1 Correlation coefficients calculated to compare data from two replicate sets of macroarrays hybridized with cDNA samples from 72 RILs from the Kanota x Ogle population.

	Description	Data sets compared	Number of correlation coefficients	Comments
r	General correlations between array sets	Spot intensity from all clones and for all cDNA samples	5	One correlation coefficient is calculated in untransformed data and each of the 4 transformation methods.
r_{lines}	Correlation among lines	Spot intensity from all clones for each cDNA sample	76	One correlation coefficient is calculated for each cDNA sample
r_{clones}	Correlations among clones	Spot intensity for each clone across the 76 cDNA samples	5 x 288	One correlation coefficient is calculated for each clone, in untransformed data and each of the 4 transformation methods.

Table 6.2 Summary of 288 Pearson correlation coefficients (r_{clones}, for 288 oat clones) calculated across paired observations from two replicate sets of arrays probed with 76 different cDNA samples

Transformed variable	Simple linear (Pearson) correlation coefficients			Percentage of clones for which the correlation was positive and significant ($P < 0.01$)
	Minimum	Median	Maximum	
I	-0.21	0.05	0.87	7.6
I/T	-0.70	0.16	0.76	15.3
I_S	-0.73	0.27	0.96	50.7
I/A	-0.79	0.29	0.96	53.1
I/E	-0.47	0.43	0.88	69.4

Table 6.3 Effects of cDNA sample and of array set on spot intensity evaluated by two-way ANOVA for each of five variables (spot intensity and four transformations of spot intensity)

Variable	F values			Percentage of clones with significant cDNA sample effect ($P < 0.05$)
	Minimum	Median	Maximum	
I	0.71	1.07	3.56	10.7
I/T	0.64	1.23	6.35	16.5
I_S	0.48	1.53	7.77	53.3
I/A	0.26	1.11	8.44	28.9
I/E	0.58	3.05	47.96	63.9

Table 6.4 Number of test statistic peaks above significance threshold at $P < 0.1$, $P < 0.05$ and $P < 0.01$, obtained for I/E data from two sets of replicate arrays and for the average between the two sets

		Number of significant peaks		
		$P < 0.1$	$P < 0.05$	$P < 0.01$
Data sets (I/E)	Set 1	30	13	1
	Set 2	47	26	8
	Average	50	33	11
Common peaks between data sets	Set 1 and Set 2	5	1	0
	Set 1 and Average	14	7	1
	Set 2 and Average	22	10	5

Table 6.5 Array position, clone name and BLAST hits for clones showing significant peaks at $P < 0.05$ in averaged I/E data. QTLs for expression level were identified across 72 RIL lines of the population Kanota x Ogle for 288 oat clones spotted on macroarrays

Clone position on array	Clone Name [a]	Description
A11_Position3	hippo2A04_JF219_016	*Avena sativa* LTR-retrotransposon OARE-1 gag-pol pseudogene for polyprotein
A12_Position2	combo2F19_JF444_076	Similar to copia-type pol polyprotein [*Oryza sativa* (*japonica* cultivar-group)]
A12_Position3	hippo2A21_JF222_095	Isoprenoid biosynthesis-like protein [*Oryza sativa* (*japonica* cultivar-group)]
A5_Position3	hippo1M02_JF229_003	Wheat cold-stressed seedling cDNA library *Triticum aestivum* cDNA clone
A7_Position2	combo2D11_JF444_046	*Secale cereale* DNA for dispersed repeat sequence (R173-2)
B10_Position3	hippo2H14_JF221_057	No BLAST hit
B3_Position1	combo1B11_JF427_048	Putative 60S RIBOSOMAL PROTEIN L36 [*Oryza sativa* (*japonica* cultivar-group)] gi|21104629
B3_Position3	hippo2C03_JF222_014	*Hordeum vulgare* sp. vulgare cultivar Morex BAC clone 773k14, complete sequence
B6_Position1	combo1B18_JF428_079	Putative pM5 collagenase [*Oryza sativa* (*japonica* cultivar-group)] gi|14495219
B8_Position1	combo1C04_JF426_014	Epsilon1-COP [*Oryza sativa*]
B9_Position3	hippo2F20_JF221_076	No BLAST hit
C11_Position2	combo3H13_JF447_057	Acetyl-coenzyme A carboxylase [*Alopecurus myosuroides*]
C11_Position3	hippo2P18_JF221_065	*Avena sativa* isolate Pc68LrkC5 sequence containing retrotransposon and repetitive DNA linked to receptor kinase gene

Table 6.5 (continued)

Clone position on array	Clone Name [a]	Description
C5_Position3	hippo2L18_JF221_069	Unknown protein [*Oryza sativa* (*japonica* cultivar-group)] gi|18873850
C8_Position3	hippo2N19_JF220_068	No BLAST hit
C9_Position1	combo1F01_JF427_011	*Triticum aestivum* FGAS: Library 5 GATE 7 *Triticum aestivum* cDNA
C9_Position3	hippo2O09_JF222_033	No BLAST hit
D10_Position3	ogle5C03_JF423_095	*Avena strigosa* DNA for dispersed repeat region, clone As22
D11_Position1	combo1G21_JF395_089	Stromal cell-derived factor 2-like protein [*Oryza sativa* (*japonica* cultivar-group)] gi|37806069
D12_Position1	combo1H06_JF428_025	Alanine aminotransferase [*Deschampsia antarctica*]
D3_Position3	ogle2E10_JF396_072	F7F22.17 [*Arabidopsis thaliana*]
D4_Position2	combo3K08_JF446_022	NADH-dependent Glutamate Synthase [*Oryza sativa*]
D9_Position3	ogle4_5H02_JF422_002	*Avena sativa* receptor-like kinase extracellular domain rlk2a13 pseudogene, complete sequence
F5_Position1	combo1J11_JF427_040	Hypothetical protein- *Arabidopsis thaliana* gi|8439892 with similarity to DNA repair protein RAD51 homolog 4 (TRAD) from *Homo sapiens* gi|6174940.
F9_Position1	combo1K07_JF395_022	No BLAST hit
G7_Position2	combo6J05_JF557_023	Similar to Transposon MAGGY gag and pol gene homologues [*Oryza sativa* (*japonica* cultivar-group)]

Table 6.5 (continued)

Clone position on array	Clone Name [a]	Description
G7_Position2	combo6J05_JF557_023	Similar to Transposon MAGGYgag and pol gene homologues [*Oryza sativa* (*japonica* cultivar-group)]
G8_Position1	combo1N03_JF427_004	No BLAST hit
H11_Position1	combo2A09_JF429_047	Expressed protein [*Arabidopsis thaliana*]
H12_Position1	combo2A19_JF429_080	No BLAST hit
H3_Position1	combo1O07_JF395_018	*A.vaviloviana* ty1-copia like retrotransposon DNA
H4_Position1	combo1O10_JF426_033	No BLAST hit
H4_Position2	combo7C20_JF553_054	O*ryza sativa* (*japonica* cultivar-group)

[a] Note about library names: libraries K_minus_O and O_minus_K were originally named Hippopotame and Pamplemousse, which is still reflected in the name of the clones. All clones with a name starting with Hippo originate from cultivar Kanota, and clones with a name starting with Pamp originate from cultivar Ogle.

Table 6.6 Clones showing significant QTL effects ($P < 0.05$) based on transformed hybridization intensity (I/E) of cDNA from 72 RIL lines of the Kanota x Ogle oat population.

Clone Name	Parent contributing the positive allele	KO Linkage group	TS at peak	Estimated QTL effect (% of total phenotypic variance)	t value [d]
hippo2A04_JF219_016 [a]	Kanota	29_43	21.1	27	-2.46
combo2F19_JF444_076	Kanota	29_43	18.5	24	1.09
hippo2A21_JF222_095 [a]	Kanota	29_43	18.6	24	1.82
hippo1M02_JF229_003	Kanota	Unlinked (ac06.625)	13.1	18	-2.21
combo2D11_JF444_046	Kanota	29_43	12.4	17	-0.14
hippo2H14_JF221_057 [a]	Kanota	29_43	12.6	17	0.27
combo1B11_JF427_048	Kanota	29_43	13.5	18	-0.64
hippo2C03_JF222_014	Kanota	29_43	14.1	19	-0.76
combo1B18_JF428_079	Kanota	Unlinked (ac06.625)	4.9	7	-3.15 [c]
combo1C04_JF426_014	Kanota	29_43	14.5	19	-0.22
hippo2F20_JF221_076	Kanota	29_43	14.1	19	-0.23
combo3H13_JF447_057	Kanota	29_43	15.6	21	2.73
hippo2P18_JF221_065	Kanota	29_43	6.0	9	-0.82
hippo2L18_JF221_069	Kanota	29_43	16.3	22	-2.08
hippo2N19_JF220_068	Kanota	29_43	14.7	20	-1.44
combo1F01_JF427_011	Kanota	29_43	12.7	17	-1.06
hippo2O09_JF222_033	Kanota	29_43	14.6	20	-1.55
ogle5C03_JF423_095	Kanota	29_43	18.1	24	0.21

Table 6.6 (continued)

Clone Name	Parent contributing the positive allele	KO Linkage group	TS at peak	Estimated QTL effect (% of total phenotypic variance)	t value [d]
ogle5C03_JF423_095	Kanota	29_43	18.1	24	0.21
combo1G21_JF395_089	Kanota	29_43	15.9	21	-1.10
combo1H06_JF428_025	Kanota	29_43	14.0	19	-0.66
ogle2E10_JF396_072	Kanota	29_43	15.5	21	-1.20
combo3K08_JF446_022	Kanota	Unlinked (ac06.625)	9.8	14	1.14
ogle4_5H02_JF422_002	Kanota	29_43	14	19	-0.62
combo1J11_JF427_040	Kanota	29_43	13.5	18	-3.59 [c]
combo1K07_JF395_022	Kanota	29_43	14.1	19	-4.53 [c]
combo6J05_JF557_023	Ogle	6	9.0	13 (24) [b]	-1.61
combo6J05_JF557_023	Kanota	Unlinked (UMN110)	10.0	14 (24) [b]	-1.61
combo1N03_JF427_004	Kanota	29_43	14.4	19	-0.69
combo2A09_JF429_047	Kanota	29_43	14.2	19	-0.46
combo2A19_JF429_080	Kanota	29_43	12.2	17	-0.62
combo1O07_JF395_018 [a]	Kanota	29_43	14.5	19	-0.46
combo1O10_JF426_033 [a]	Kanota	29_43	17.4	24	-0.01
combo7C20_JF553_054	Kanota	Unlinked (UMN110)	19.3	25	-1.79

[a] Peaks appearing in set 1 and set 2

[b] Combined effect of both QTLs

[C] Significant at $P < 0.05$

[d] The t-values result from pooled t-tests for differences in expression levels between Kanota and Ogle for the 33 clones. Each parent was represented by two replicate cDNA samples.

Table 6.7 Number of environments where significant correlations ($P < 0.05$) were found between averaged I/E values for clones showing significant eQTLs and phenotypic traits in the Kanota x Ogle population (correlation coefficients for each environment are shown in parentheses)

Clone	Heading date[a]	Lodging[b]	Tertiary rating[c]	Dmin[d]	Dmax[e]	Fshape[f]	Groat[g]	Protein[h]	β-glucan[i]
Number of environments where trait was measured[j]	3	3	3	5	5	5	5	5	7
hippo2A04_JF219_016		2 (0.29; 0.29)	1 (0.26)						
hippo2A21_JF222_095							1 (0.25)		
hippo1M02_JF229_003									1 (0.38)
combo1B11_JF427_048			1 (0.26)						
hippo2C03_JF222_014		2 (0.32; 0.30)	1 (0.24)		1 (-0.32)		1 (0.30)	1 (-0.27)	2 (0.32; 0.31)
combo3H13_JF447_057						3 (0.26; 0.26; 0.35)			
hippo2L18_JF221_069							1 (0.25)		
hippo2N19_JF220_068		2 (0.35; 0.31)	1 (0.45)		1 (0.33)			3 (0.30; -0.25; -0.36)	
hippo2O09_JF222_033		2 (0.31; 0.27)	1 (0.31)				1 (0.30)	1 (-0.31)	
ogle5C03_JF423_095				1 (0.27)					

Table 6.7 (continued)

Clone	Heading date[a]	Lodging[b]	Tertiary rating[c]	Dmin[d]	Dmax[e]	Fshape[f]	Groat[g]	Protein[h]	β-glucan[i]
combo1H06_JF428_025							1 (0.27)		
ogle2E10_JF396_072	1 (0.26)			1 (0.26)					1 (0.31)
ogle4_5H02_JF422_002					1 (-0.27)				
combo1J11_JF427_040									1 (0.32)
combo1K07_JF395_022		2 (0.26; 0.25)					1 (0.26)		1 (-0.25)
combo1N03_JF427_004		3 (0.32; 0.31; 0.27)			1 (0.33)		1 (0.26)		1 (0.34)
combo2A09_JF429_047							1 (0.25)		
combo1O07_JF395_018			1 (0.42)		1 (0.27)			2 (0.28; -0.38)	
combo1O10_JF426_033			1 (0.29)					1 (-0.32)	

[a] number of days from planting to 50% panicle emergence
[b] lodging severity; [c] rating for presence of tertiary kernels
[d] mean kernel width as determined by image analysis
[e] mean kernel length as determined by image analysis
[f] kernel plumpness as determined by image analysis
[g] groat percent of the seed
[h] NIR prediction of protein content
[i] Flow injection analysis (FIA) of β-glucan
[j] Phenotypic data from the Kanota x Ogle population were communicated by Dr. Howard Rines (USDA-ARS, University of Minnesota, St. Paul, MN)

Figure 6.1 Diagram showing the layout used for each of 96 three-by-three blocks of printing positions making up an 864-spot array. Within each such three-by-three block, six of the nine positions were used for two replicate spots of each of three oat clones, with positions 1 and 8 used for one clone, positions 2 and 6 used for a second clone and positions 3 and 4 used for a third clone. Position 5 of each block was used for a human nebulin spot as an external control. Position 7 of each block was used for an actin clone spot as an internal control. Position 9 was left empty as a negative control.

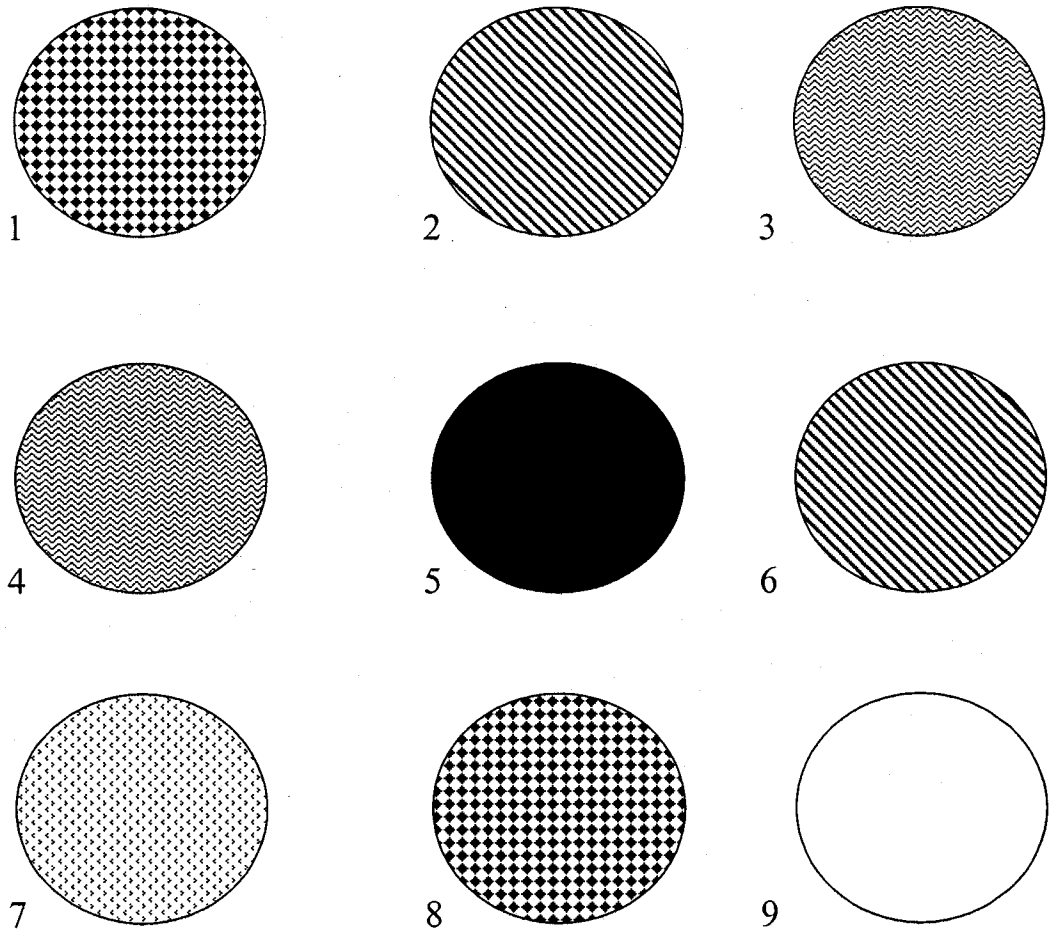
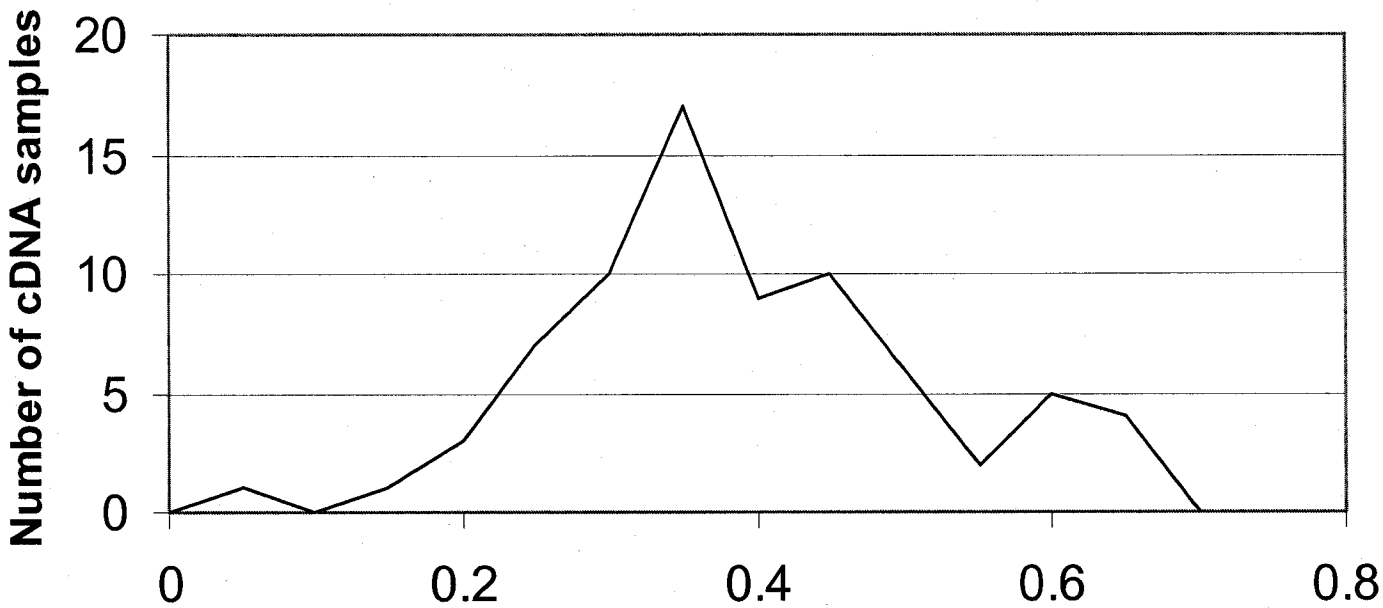
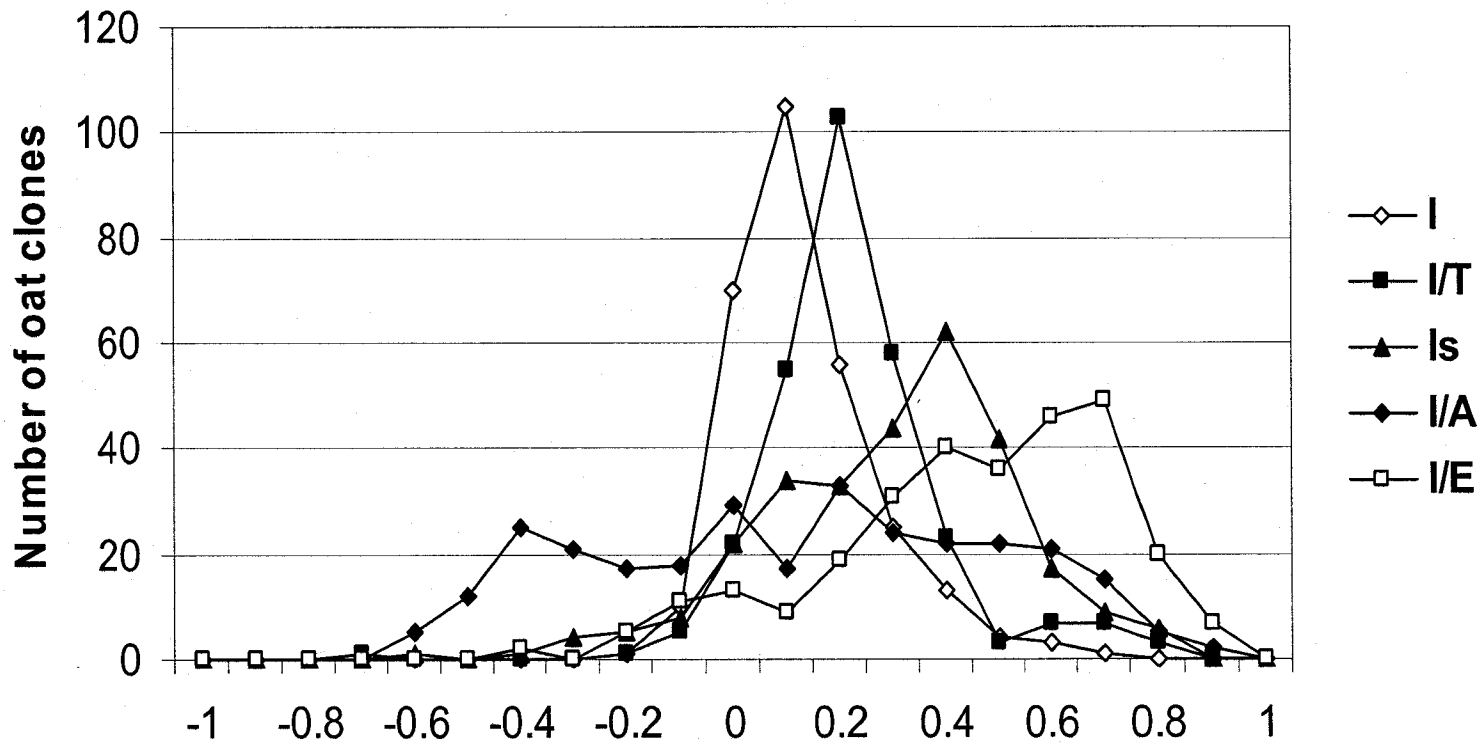

1
2
3
4
5
6
7
8
9

Figure 6.2 Distribution of simple linear correlation coefficients for associations between spot intensity data of two sets of 76 macroarrays among cDNA samples.



Simple linear correlation coefficient (r_{lines}) between sets of macroarrays

Figure 6.3 Distributions of simple linear correlation coefficients for associations between spot intensity data of two sets of 76 macroarrays. Each polygon is the distribution of the coefficients for 288 oat clones for each of five variables (spot intensity and four transformations of spot intensity).



Simple linear correlation coefficient (r_{clones}) between sets of macroarrays

Figure 6.4 Scatter plots representing the distribution of correlation coefficients (r) between spot intensity data of the two sets of macroarrays versus the average I/E of the two replicate spots for each clone on an array.

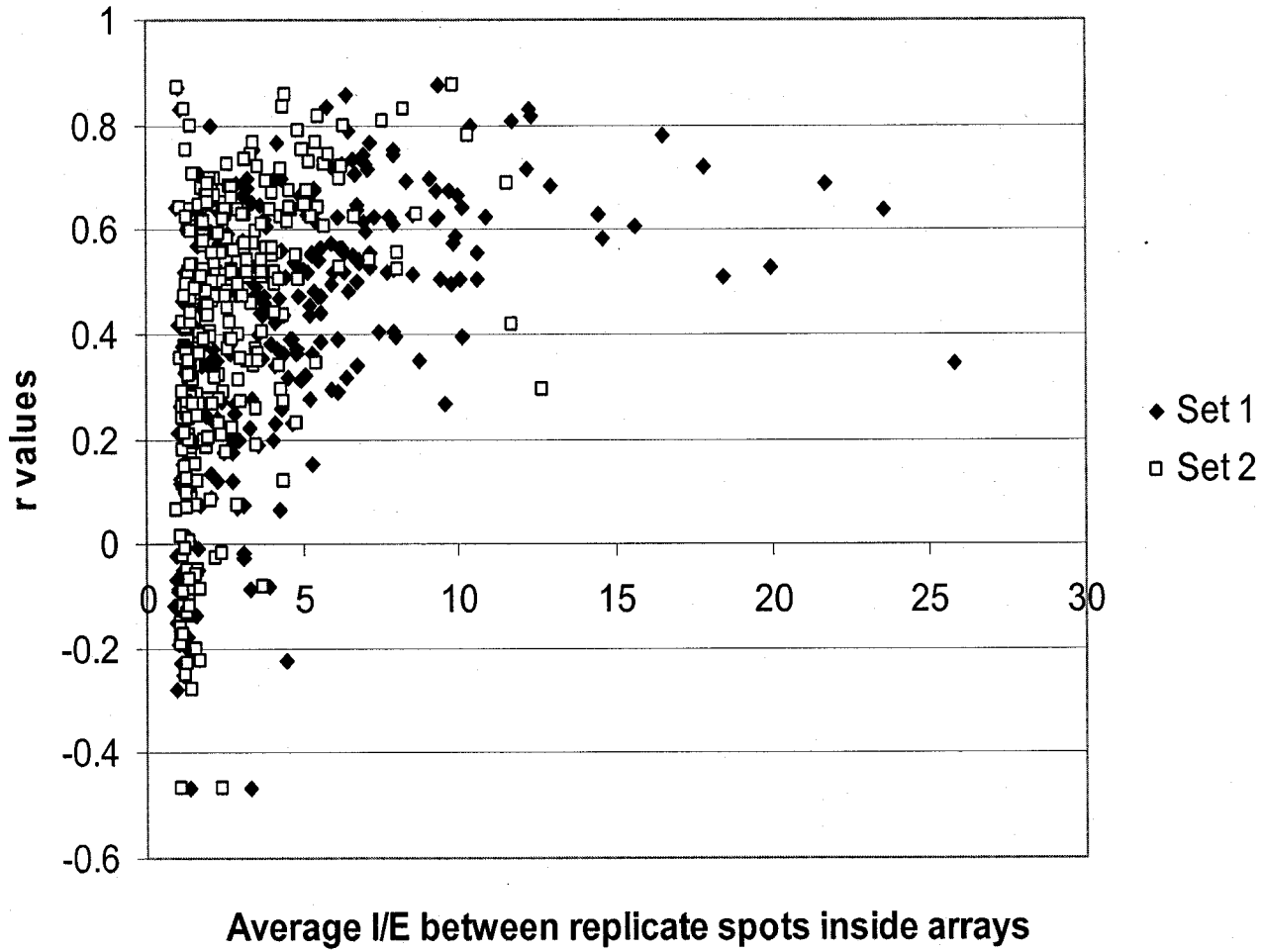
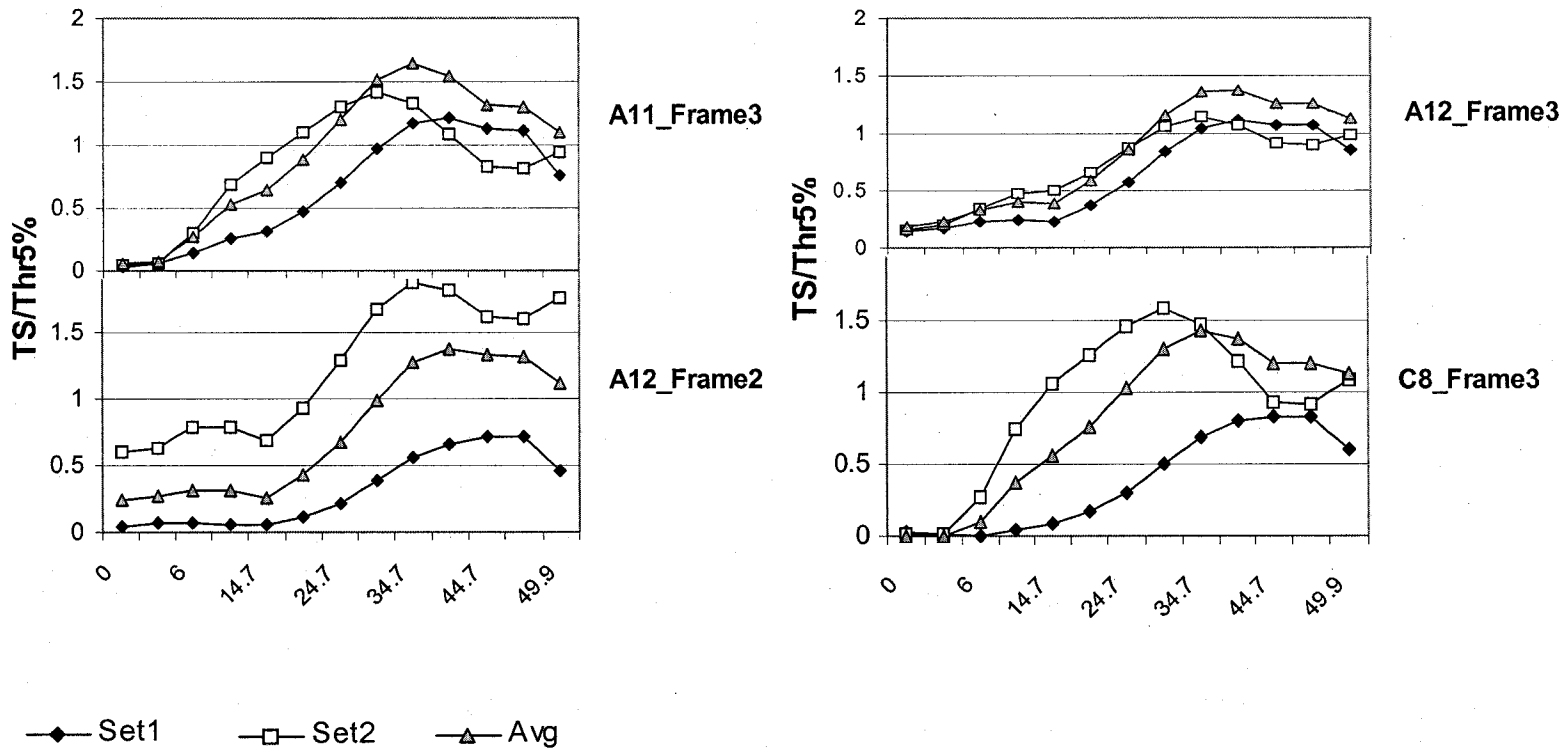
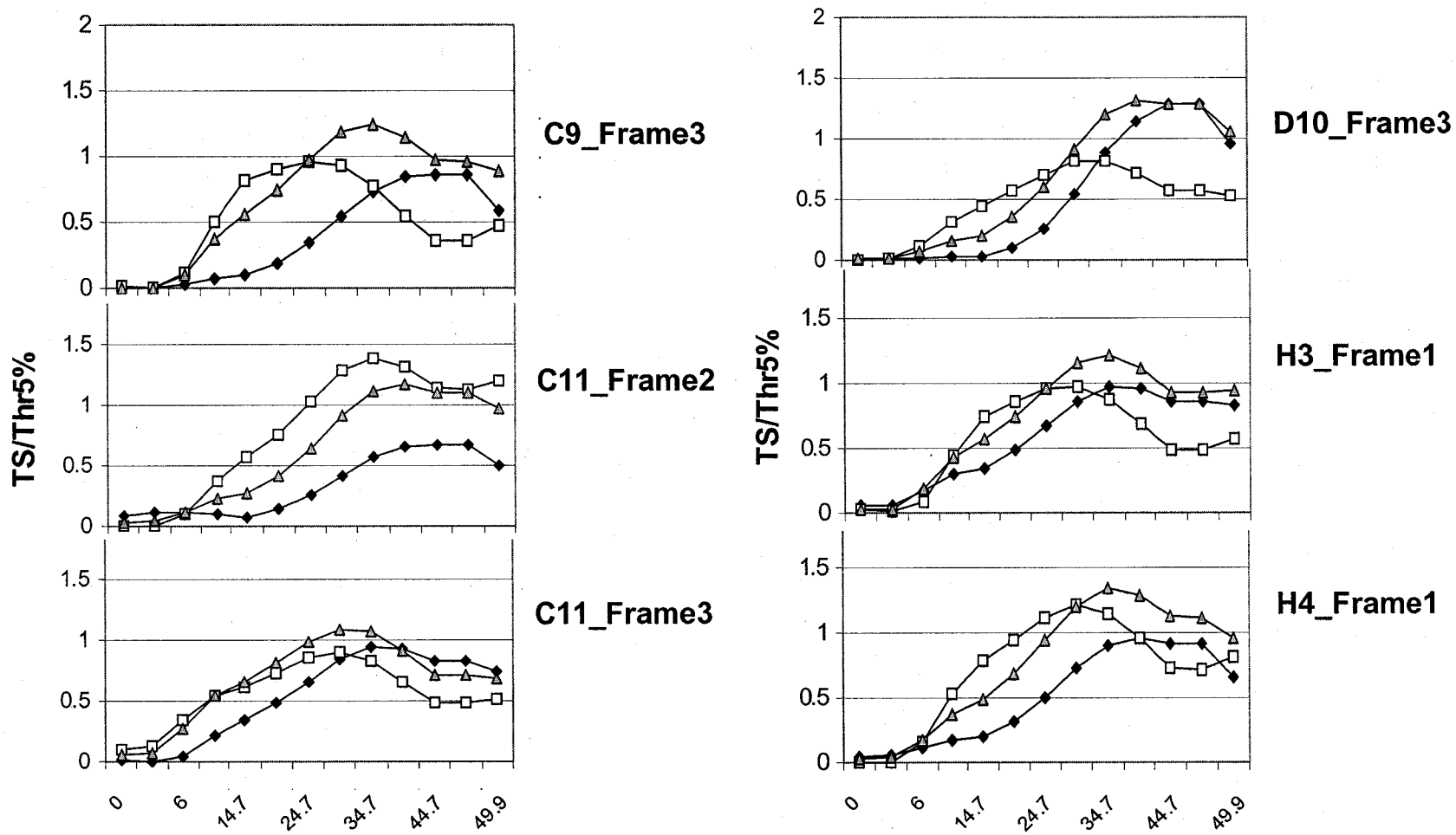

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
r values
0
5
10
15
20
25
30
Average I/E between replicate spots inside arrays
Set 1
Set 2

Figure 6.5 Results of QTL mapping of expression levels of selected genes in the Kanota/Ogle population. The clones selected showed a peak significant at $P < 0.05$ in the average I/E data, and did not map to an unlinked marker. The x-axis represents the cumulative genetic distance (in cM) on KO linkage group 29_43. The y-axis represents the test statistic divided by the 5% significance threshold obtained after 1000 data permutations.


TS/Thr5%
A11_Frame3
A12_Frame2
A12_Frame3
C8_Frame3
0
0.5
1
1.5
2
0
6
14.7
24.7
34.7
44.7
49.9
Set1
Set2
Avg


TS/Thr5%
C9_Frame3
C11_Frame2
C11_Frame3
D10_Frame3
H3_Frame1
H4_Frame1
0
0.5
1
1.5
2
0
6
14.7
24.7
34.7
44.7
49.9

Chapter 7
General Discussion

Most genetic variability has its source in sequence variation, but it can be manifested in the cell in different ways. The first possible manifestation is structural, and involves a variation in protein function or performance because of a change in the coding DNA level. Structural variability includes alterations of enzyme activity and specificity, or changes in protein structure or stability. An extreme case would be the total absence of the product of the gene. Structural variability can have repercussions on the proteome and metabolome of the cell. A second manifestation of genetic variability would be the control of gene expression. The gene products might not differ in their properties, but the level or timing of expression might be variable. This includes the control of transcription, as well as splicing mechanisms and messenger RNA stability. Recent hypotheses consider variations in the regulation of gene expression as the main impact of genetic variability on phenotype (see review by Pomp et al., 2004).

Both types of mechanisms are investigated in the work presented here. The first approach, investigating sequence variability at the structural gene level, consisted of evaluating sequence variability among cultivars in genes coding for enzymes involved in lipid and protein biosynthesis. The second approach was to investigate manifestations of genetic variability at the level of gene expression control. Therefore we surveyed genes that were differentially expressed between two oat cultivars at a defined development stage. Finally, we assessed the variability in levels of expression of genes across a RIL population of oat, and identified genome regions influencing those levels.

One of the outcomes of the study of sequence variability among oat cultivars is the identification of SNPs and indels distinguishing cultivars. In the partial gene sequences studied here, the average frequency of SNPs and indels was low, similarly to what has been observed between cultivars groups of rice. As we did not analyse complete gene sequences, the average number of SNPs and indels reported is only an estimate of the number of SNPs and indels in the whole gene sequence, and could be lower or higher in other regions of the gene. The enzymes and proteins encoded by the studied genes are not only involved in storage compound synthesis, but also in several other key cellular processes. As variation in the activity of these enzymes would result in major effects on

the cell or plant, it is not surprising that the coding regions of these genes are highly conserved. Intronic sequences were also conserved among cultivars, and sequence variation was mainly limited to SNPs. This would make the design of cultivar-specific markers challenging to design.

Another result of the study of sequence variability among cultivars is that, for the eight studied genes the sequences analysed tend to cluster into groups of very closely related sequences (families) originating from different cultivars, rather than according to the cultivar they were obtained from. Sequence variability inside sequence families was distinctly lower than among sequences from the same cultivar. For six out of the eight genes studied, three distinct families were observed, which may correspond to homeologous genes originating from the three subgenomes of hexaploid oat. This could be confirmed by cloning and sequencing the corresponding gene fragments from diploid or tetraploid oat. If it is verified that those sequence families correspond to homeologous genes, SNPs and indels distinguishing families would allow the design of genome-specific markers. Those markers would allow the tracking of the expression of homeologous versions of the genes across different tissues and developmental stages of oat. This would help us understand how each subgenome participates in the development and metabolism of diploid, tetraploid or hexaploid oat, and maybe help formulate new hypotheses on why polyploidy is common in plants.

Surveying sequence variability is an easy strategy to implement, as long as the target sequence is known. Gene sequences can be obtained by approaches relying on gene conservation among related organisms, such as the approach described in Chapter 3. For sequence variations affecting gene expression, it would be more difficult to proceed in a similar way. In general regulatory sequences such as promoters tend to be less conserved among organisms than structural gene sequences (Vikkula et al., 1992). Little is known still about sequences influencing gene expression in *cis,* and even less about those influencing gene expression in *trans*. The effects of genetic diversity acting on gene expression are therefore usually evaluated through the differential expression of genes in a particular tissue and at a particular developmental stage. One of the many difficulties of this type of investigations is the choice of appropriate tissues and stages. Another major difficulty is the identification of the genome regions influencing the expression of the

target genes. *Cis* regulatory regions can be isolated by methods such as reverse PCR. *Trans* regulatory regions on the other hand can be located anywhere in the genome, and there are no direct methods to detect them. One way to identify such regions is to carry out a QTL study in a population in which this regulatory factor might segregate, by considering the expression levels of the target genes as quantitative traits.

We chose to carry out our expression level study in young kernels of oat sampled eight days after the "yellow anther" stage (stage GRO:0007102 according to the Plant Ontology Consortium, http://dev.plantontology.org/docs/growth/growth.html). This stage just precedes the accumulation of storage compounds in the kernel, and we speculated it would be an appropriate stage to detect variations in key factors influencing kernel traits such as protein and lipid content. In order to increase the chance of observing a segregation of expression level in the mapping population, we selected genes that were found to be differentially expressed between the two parents of the population. Most of the sequences obtained did not correspond to any known gene sequence. For other sequences it was possible to associate a potential gene and gene function, through the annotation of their BLAST hits. Those gene functions seemed to indicate that cells of the young kernels of Kanota and Ogle were involved in different metabolic processes at this stage of development. Clones specifically expressed in Ogle included clones with homology to genes coding for storage proteins and for products involved in photosynthesis and photorespiration. No gene category was exclusive to Kanota, but it showed a distinctly higher number of differentially expressed retrotransposon-related sequences. The different patterns of gene expression were reflected in segregation of expression in the mapping population: among the 17 clones with BLAST hits for which an eQTL was detected, six had homologies to transposable elements, and two to enzymes involved in protein biosynthesis (alanine aminotransferase and NADH-dependent glutamate synthase).

A striking result of this study is the co-location of nearly 82% of the detected eQTLs to a single locus on linkage group 29_43. Similar eQTL "hotspots" have been reported in mouse and Arabidopsis (Schadt et al., 2003, DeCook et al., 2005). In the mouse and Arabidopsis studies, the eQTLs overlapped with QTLs for traits segregating in the populations, but the oat linkage group 29_43 does not feature any QTL for the traits

that have been evaluated in the Kanota x Ogle population. cDNA clones mapping to this linkage group have homologies to puroindolins and an enzyme involved in amino acid metabolism, suggesting the involvement of genes on this linkage group in protein biosynthesis.

Nevertheless, the lack of reported QTLs on this linkage group indicates that the genetic control exercised by that region does not have a detectable impact on traits evaluated in the Kanota x Ogle population. This was disappointing, as by working with young kernels, we were hoping to identify genomic regions affecting traits such as protein and lipid content through their effects on the expression of candidate genes. However, it is possible that the eQTL on linkage group 29_43 influences one or more phenotypic traits that have not yet been evaluated in the population. This again raises the challenge of identifying the right stage and tissue to search for genetic variability influencing a particular trait, without complete understanding of the biochemical processes and developmental steps involved.

This challenge is particularly strong when dealing with crop plants, for which most of the available phenotypic data is for economically important traits affecting agricultural productivity and end-use quality. Those traits often do not correspond to distinct physiological properties or anatomical characteristics of the organism. A typical example is yield: it is the sum of many different properties of the plant, including resistance to known or unknown pathogens, efficiency of different metabolic processes, amount of resources invested in the organ of agronomic interest, and many more factors. Each of these "sub-traits" might segregate independently from each other, but the only phenotypic data available bulks them all in one complex trait. Evaluating the genetic variability at the origin of such composite traits is extremely difficult, mainly because of the multitude of weak effects involved. Decomposing some traits of economic importance in oat and other well-studied crop plants into genetically simpler component-traits would allow us to draw more conclusions from work such as that presented in this thesis, and help understand the genetic basis of important phenotypes.

Chapter 8
Conclusions

Sequence variability among cultivars of oat

Sequences from eight oat genes coding for products involved in lipid and protein biosynthesis exhibit low frequencies of SNP and indel polymorphisms among oat cultivars (2.85% and 7.87%, respectively).

Clustering of sequences into gene families

Across oat cultivars, sequences from eight genes coding for products involved in lipid and protein biosynthesis cluster in families of very closely related sequences, often in three distinct families. Within-cultivar sequence variability among these families is greater than among-cultivar sequence variability within sequence families. Sequence variability among homeologous oat sequences is greater than among homologous sequences.

Potential for designing gene-targeted markers

Single-nucleotide and indel polymorphisms among the component genomes of oat are frequent enough to support the design of DNA markers capable of distinguishing among gene copies on different genomes. The design of cultivar-specific markers would be more difficult, as only a few among-cultivar SNPs were discovered.

Differential gene expression among oat cultivars and lines

Oat cultivars and lines can exhibit differential gene expression in developing kernels. In this research, identified gene sequences expressed in developing kernels of Ogle oat exhibited homology mainly to genes involved in amino acid and storage protein metabolism, whereas identified gene sequences expressed in kernels of Kanota oat exhibited homology mainly to sequences implicated in the regulation of gene expression and to genes related to transposable elements. Most of the sequences that exhibited differential expression between these two cultivars also exhibited significant variation in expression among recombinant inbred lines derived from a cross between Kanota and Ogle.

Transcription of retrotransposons

Retrotransposons are actively transcribed in young kernels of oat.

Detection of eQTLs in oat

A region on linkage group 29_43 of the Kanota x Ogle oat linkage map affects the expression level of numerous genes. This eQTL region is a potential "hot spot" for the regulation of gene expression in developing kernels of oat. Several individual DNA markers that have not been assigned to any linkage group in oat are associated with expression levels of individual sequences. None of these eQTLs coincided with previously detected QTLs in oat.

Association of acetyl-CoA carboxylase expression with kernel shape

The expression of an acetyl-CoA carboxylase gene exhibits a positive association with kernel plumpness and a negative association with kernel length.

Reduction of the effect of exposure time on ^{32}P-labelled arrays

Effects of exposure time on the intensity of signals obtained from ^{32}P-labelled arrays can be reduced by dividing the signal intensity of each individual spots on an array by the median background signal intensity of that array.

Chapter 9
Contributions to knowledge

When this project was started, public databases contained only few sequences from oat (889 nucleotide entries for *Avena*, and only 720 for *Avena sativa* in GenBank as of 15/11/01). This work contributed 2184 new oat sequences that will be added to public databases.

Among those sequences, 1208 are genomic sequences, obtained by PCR amplification of genomic DNA from 10 different cultivars of oat with primers targeting conserved regions of genes involved in lipid and protein biosynthesis pathways. They constitute a valuable resource that can be mined for sequence polymorphisms, and used to design allele specific markers for the study of segregation or transcript expression.

This is the first report of a multi-gene study of sequence variability among cultivars of oat. Sequence analysis showed similar levels of sequence variability among cultivars as has been reported between rice cultivar groups *indica* and japonica. Phylogenetic analysis showed that sequences clustered in families of closely related sequences, and that sequence variability among gene families was higher than sequence variability among cultivars. This is an important result that explains the difficulty of designing cultivar-specific markers in oat. We speculate that these families correspond to homeologous versions of the genes originating from the three subgenomes of hexaploid oat. If this is confirmed, the sequence variations among sequence families will be useful to design markers specific to homeologous versions of these genes.

This is the first time differential expression between two oat cultivars has been studied. It generated 976 sequences corresponding to cDNA sequences from cultivars Kanota and Ogle. Clones were isolated by SSH so that the selected genes were differentially expressed between young kernels of the two cultivars. Many of the corresponding genes did not have significant homology to known genes or proteins, and may represent novel genes that can be characterized more fully in future studies.

In this work, we report the detection of eQTLs in the Kanota x Ogle mapping population. This makes oat one of the few plant species, with maize and Arabidopsis, in which quantitative gene expression has been mapped. We identified 33 significant eQTLs

($P < 0.05$). We report the clustering of several eQTLs to a single genomic region; similar clustering has been observed in mouse and Arabidopsis.

This work presents the first comparison of data transformation methods for standardizing data originating from different ^{32}-P labelled hybridization events. We report a novel data transformation method consisting of dividing signal intensity by the median background of the array. This method resulted in the highest degree of correlation between replicate arrays. We show that this method also reduces the effect of different exposure time on the measure of radioactive signal intensity.

Chapter 10
Suggestions for future research

In this work, I have taken several approaches to examine genetic variation in oat. Each of these could be carried further in future research. The work could also be extended to examine the impact of the identified genetic variability on the phenotype. Some suggested extensions of this work are listed below.

1. The panel of lines and cultivars for which sequence variability was evaluated could be extended. This would allow confirmation of the grouping of sequences in families, as well as the number of these families. Including diploid and tetraploid oat in the panel would allow testing of the hypothesis that the identified sequence families correspond to the subgenomes of oat. Of the three families appearing for some genes in hexaploid oat, only two should appear in tetraploid oat, and only one in diploid oat.

2. A few cultivar-specific markers could be derived from the sequence information obtained in this work. Additional markers might be obtained from a wider panel of lines.

3. The eight genes included in this study could be mapped and their position compared to existing QTLs and eQTLs.

4. Markers that are specific to sequence families could be designed. If the correspondence between sequence families and homeologous genomes is confirmed, those markers could be used to map specific homeologous versions of the genes, and to associate the linkage groups on which they map with the corresponding genomes.

5. The cDNA corresponding to the eight genes studied in Chapter 3 could be amplified by RT-PCR and cloned from different genetic backgrounds of oat, different tissues and different developmental stages in order to determine which homeologous versions of these genes are transcribed and when they are transcribed.

6. In order to understand why the plant tolerates abundant expression of retrotransposon-related genes, the expression levels of such genes could be evaluated across different genetic backgrounds, different tissues and different developmental stages of oat to determine where and when they are expressed, and to understand why some genotypes tolerate abundant expression of these transcripts.

7. Genes for which significant eQTLs were detected could be mapped to compare their map location with the eQTL location.

8. The eQTL results obtained through macroarray hybridization could be confirmed by Northern hybridization.

9. Homologous regions corresponding to the eQTL hot-spot on linkage group 29-43 in other crosses of oat or other cereals could be examined by comparative mapping, and genes present in this area considered as potential candidate genes influencing gene expression.

Bibliography

Alban C, Baldet P and Douce R (1994) Localization and characterization of two structurally different forms of acetyl-CoA carboxylase in young pea leaves, of which one is sensitive to aryloxyphenoxypropionate herbicides. Biochem. J. **300**: 557-565

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. **215**:403-410

Alwine JC, Kemp DJ, Stark GR (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc.Natl.Acad.Sci U.S.A. **74**:5350-5354

Andrews TJ and Kane HJ (1991) Pyruvate is a by-product of catalysis by ribulosebisphosphate carboxylase/oxygenase. J. Biol.Chem. **266**: 9447-9452

Arumuganathan K and Earle ED. (1991) Nuclear DNA content of some important plant species. Plant Mol.Biol.Rep. **9**: 208-219

Bao X and Ohlrogge J (1999) Supply of fatty acid is one limiting factor in the accumulation of triacylglycerol in developing embryos. Plant Physiol. **120**: 1057-1062

Baum BR. (1977) *OATS: Wild and Cultivated, a Monograph of the Genus Avena L. (Poaceae)*. Canada Department of Agriculture, Research Branch

Behal KM, Scholfield DJ, Hallfrisch J (2005) Comparison of hormone and glucose responses of overweight women to barley and oats. J. Am. Coll. Nutr. **24**: 182-188

Blasing OE, Gibon Y, Gunther M, Hohne M, Morcuende R, Osuna D, Thimm O, Usadel B, Scheible WR, and Stitt M (2005) Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in *Arabidopsis*. Plant Cell. **17**: 3257-3281

Botha AM, Lacock L, van Niekerk C, Matsioloko MT, du Preez FB, Loots S, Venter E, Kunert KJ, and Cullis CA (2005) Is photosynthetic transcriptional regulation in *Triticum aestivum* L. cv. 'TugelaDN' a contributing factor for tolerance to *Diuraphis noxia* (Homoptera: Aphididae)? Plant Cell Rep. **25**: 41-54

Branen JK, Chiou TJ, and Engeseth NJ (2001) Overexpression of acyl carrier protein-1 alters fatty acid composition of leaf tissue in *Arabidopsis*. Plant Physiol. **127**: 222-229

Brautigam M, Lindlof A, Zakhrabekova S, Gharti-Chhetri G, Olsson B, Olsson O (2005) Generation and analysis of 9792 EST sequences from cold acclimated oat, Avena sativa. BMC Plant Biol. **5**:18

Breitling R (2006) Biological microarray interpretation: the rules of engagement. Biochim Biophys Acta. **1759:** 319-327

Brugiere N, Dubois F, Limami AM, Lelandais M, Roux Y, Sangwan RS, and Hirel B (1999) Glutamine synthase in the phloem plays a major role in controlling proline production. Plant Cell. **11**: 1995-2012

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, and de Haan G (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat.Genet. **37**: 225-232

Carding SR, Lu D and Bottomly KA (1992) A polymerase chain reaction assay for the detection and quantification of cytokine gene expression in small number of cells. J.Immunol.Methods **151**: 277-287

Case-Green SC, Mir KU, Pritchard CE, and Southern EM (1998) Analysing genetic information with DNA arrays. Curr Opin Chem Biol. **2**: 404-410

Chagnon YC, Rankinen T, Snyder EE, Weisnagel SJ, Perusse L, and Bouchard C (2003) The human obesity gene map: the 2002 update. Obes.Res. **11**: 313-367

Chakravarthy S, Tuori RP, D'Ascenzo MD, Fobert PR, Despres C, and Martin GB. (2003) The tomato transcription factor Pti4 regulates defense-related gene expression via GCC box and non-GCC box cis elements. Plant Cell. **15**:3033-3050

Cheadle C, Vawter MP, Freed WJ, and Becker KG (2003) Analysis of microarray data using Z score transformation. J.Mol.Diagn. **5**: 73-81

Cheng DW and Armstrong KC (2002) Direct capture and cloning of receptor kinase and peroxidase genes from genomic DNA. Genome **45**: 977-983

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, and Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat.Genet.**37**: 233-242

Cheung VG and Spielman RS (2002) The genetics of variation in gene expression. Nat.Genet.**32 Suppl**: 522-525

Chichkova S, Arellano J, Vance CP, and Hernandez G (2001) Transgenic tobacco plants that overexpress alfalfa NADH-glutamate synthase have higher carbon and nitrogen content. J.Exp.Bot. **52**: 2079-2087

Consoli L, Lefevre A, Zivy M, de Vienne D, and Damerval C (2002) QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. Plant Mol.Biol. **48**: 575-581

Crawford DC, Akey DT, and Nickerson DA (2005) The patterns of natural variation in human genes. Annu.Rev.Genomics Hum.Genet. **6**: 287-312

Crofts AJ, Washida H, Okita TW, Satoh M, Ogawa M, Kumamaru T and Satoh H (2005) The role of mRNA and protein sorting in seed storage protein synthesis, transport, and deposition. Biochem Cell Biol. **83**: 728-37

Damerval C, Maurice A, Josse JM, and de Vienne D (1994) Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics **137**: 289-301

De Koeyer DL, Tinker NA, Wight CP, Deyl J, Burrows VD, O'Donoughue LS, Lybaert A, Molnar SJ, Armstrong KC, Fedak G, Wesenberg DM, Rossnagel BG, and McElroy AR (2004) A molecular linkage map with associated QTLs from a hulless x covered spring oat population. Theor.Appl.Genet. **108**: 1285-1298

de la Roche IA, Weber EJ, and Alexander DE (1971) The selective utilization of diglyceride species into maize triglycerides. Lipids **6**: 531-540

Decook R, Lall S, Nettleton D, and Howell SH (2005) Genetic regulation of gene expression during shoot development in *Arabidopsis*. Genetics **172**: 1155-1164

Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, and Siebert PD (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc.Natl.Acad.Sci U.S.A. **11**: 6025-6030

Diatchenko L, Lukyanov S, Lau YF, and Siebert PD (1999) Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. Methods Enzymol. **303**: 349-380

Dudoit S and Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. **3:** Epub 2002 Jun 25

Elborough KM, Swinhoe R, Winz R, Kroon JTM, Farnsworth L, Fawcett T, Martinezrivas JM and Slabas AR (1994) Isolation of cDNAs from Brassica napus encoding the biotin-binding and transcarboxylase domains of acetyl-CoA carboxylase: assignment of the domain-structure in a full-length Arabidopsis thaliana genomic clone. Biochem J **301**: 599-605

El-Din El-Assal S, onso-Blanco C, Peeters AJ, Raz V, and Koornneef M (2001) A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. Nat.Genet. **29**: 435-440

Elowitz MB, Levine AJ, Siggia ED, and Swain PS (2002) Stochastic gene expression in a single cell. Science **297**: 1183-1186

Fathallah-Shaykh HM (2005) Microarrays: applications and pitfalls. Arch Neurol. **62**:1669-1672

Fawcett T, Simon WJ, Swinhoe R, Shanklin J, Nishida I, Christie WW and Slabas AR (1994) Expression of messenger-RNA and steady-state levels of protein isoforms of enoyl-ACP reductase from Brassica napus. Plant Mol Biol **26**: 155-163

Flatt T (2005) The evolutionary genetics of canalization. Q. Rev. Biol. **80**: 287-316

Focks N and Benning C (1998) wrinkled1: A novel, low-seed-oil mutant of *Arabidopsis* with a deficiency in the seed-specific regulation of carbohydrate metabolism. Plant Physiol. **118**: 91-101

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, and Tanksley SD (2000) fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. Science **289**: 85-88

Frey KJ and Hammond EG (1975) Genetics, characteristics, and utilization of oil in caryopses of oat species. J.Am.Oil Chem.Soc. **52**: 358-362

Galli-Taliadoros LA, Sedgwick JD, Wood SA, and Korner H (1995) Gene knock-out technology: a methodological overview for the interested novice. J. Immunol.Methods **181**: 1-15

Gornicki P, Podkowinski J, Scappino LA, DiMaio J, Ward E, and Haselkorn R (1994) Wheat acetyl-coenzyme A carboxylase: cDNA and protein structure. Proc.Natl.Acad.Sci U.S.A. **91**: 6860-6864

Grandbastien MA, Lucas H, Morel JB, Mhiri C, Vernhettes S, and Casacuberta JM (1997) The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. Genetica **100**: 241-252

Grattapaglia D (2004) Integrating genomics into Eucalyptus breeding. Genet. Mol. Res. **3**: 369-379

Groh S, Kianian SF, Phillips RL, Rines HW, Stuthman DD, Wesenberg DM, and Fulcher RG (2001) Analysis of factors influencing milling yield and their association to other traits by QTL analysis in two hexaploid oat populations. Theor.Appl.Genet. **103**: 9-18

Gupta PK, Rustgi S, and Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol. Biol. **57**: 461-485

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl.Acids.Symp.Ser. **41**: 95-98

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. **22**:160-74

Hattori J, Ouellet T, and Tinker NA (2005) Wheat EST sequence assembly facilitates comparison of gene contents among plant species and discovery of novel genes. Genome **48**: 197-206

Hermeking H. (2003) Serial analysis of gene expression and cancer. Curr Opin Oncol. **15**: 44-49

Hills MJ (2004) Control of storage-product synthesis in seeds. Curr. Opin. Plant Biol. **7**: 302-308

Hirel B, Bertin P, Quillere I, Bourdoncle W, Attagnant C, Dellay C, Gouy A, Cadiou S, Retailliau C, Falque M, and Gallais A (2001) Towards a better understanding of the genetic and physiological basis for nitrogen use efficiency in maize. Plant Physiol. **125**: 1258-1270

Holland JB, Moser HS, O'Donoughue LS, and Lee M (1997) QTLs and epistasis association with vernalization responses in oat. Crop Sci. **37**: 1306-1316

Holland JB, Portyanko VA, Hoffman DL, Lee M (2002) Genomic regions controlling vernalization and photoperiod responses in oat. Theor.Appl Genet **105**: 113-126

Holloway JL and Knapp SJ. (1993) G-MENDEL 3.0: software for the analysis of genetic markers and maps. Oregon State University. Corvalis, OR.

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G , Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, and Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat.Genet.**37**: 243-253

Hyperdictionnary www.hyperdictionary.com, 2006

Ingvarsson PK (2005) Molecular population genetics of herbivore-induced protease inhibitor genes in European aspen (*Populus tremula* L., Salicaceae). Mol.Biol.Evol. **22**: 1802-1812

Jain RK, Coffey M, Lai K, Kumar A, MacKenzie SL (2000) Enhancement of seed oil content by expression of glycerol-3-phosphate acyltransferase genes. Biochem.Soc.Trans **28**: 958-961

Jansen RC (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl.Genet. **85**: 252-260

Jansen RC, Van Ooijen JW, Stam P, Lister C, Dean C (1995) Genotype by environment interaction in genetic mapping of multiple quantitative trait loci. Theor.Appl.Genet. **91**: 33-37

Jaworski JG, Clough RC and Barnum SR (1989) A cerulenin insensitive short chain 3-ketoacyl-acyl carrier protein synthase in Spinacia oleracea leaves. Plant Physiol. **90**: 41-44

Ji SJ, Lu YC, Feng JX, Wei G, Li J, Shi YH, Fu Q, Liu D, Luo JC, and Zhu YX (2003) Isolation and analyses of genes preferentially expressed during early cotton fiber development by subtractive PCR and cDNA array. Nucleic Acids Res.**31**: 2534-2543

Karmally W, Montez MG, Palmas W, Martinez W, Branstetter A, Ramakrishnana R, Holleran SF, Haffner SM, Ginsberg HN (2005) Cholesterol-lowering benefits of oat-containing cereal in Hispanic americans. J. Am. Diet. Assoc. **105** 967-970

Kearsey MJ and Farquhar AG (1998) QTL analysis in plants; where are we now? Heredity **80** (2): 137-142

Kellis M, Birren BW, and Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**: 617-624

Kianian SF, Egli MA, Phillips RL, Rines HW, Somers DA, Gengenbach BG, Webster FH, Livingston SM, Groh S, O'Donoughue LS, Sorrells ME, Wesenberg DM, Stuthman DD, and Fulcher RG (1999) Association of a major groat oil content QTL and an acetyl-CoA carboxylase gene in oat. Theor.Appl.Genet. **98**: 884-894

Kianian SF, Phillips RL, Rines HW, Fulcher RG, Webster FH, and Stuthman DD (2000) Quantitative trait loci influencing Beta-glucan content in oat (*Avena sativa*, 2n=6x=42). Theor.Appl.Genet. **101**: 1039-1048

Kidwell JF (1963) Genotype x Environment interaction with isogenic lines of *Drosophila melanogaster*. Genetics **48**: 1593-1604

Kim WY, Hicks KA, and Somers DE (2005) Independent roles for EARLY FLOWERING 3 and ZEITLUPE in the control of circadian timing, hypocotyl length, and flowering time. Plant Physiol. **139**: 1557-1569

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. **16**: 111-20

Kluge AG and Farris FS (1969) Quantitative phyletics and the evolution of anurans. Syst Zool **18**: 1–32

Kojima S, Takahashi Y, Kobayashi Y, Monna L, Sasaki T, Araki T, and Yano M (2002) Hd3a, a rice ortholog of the *Arabidopsis* FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions. Plant Cell. Physiol. **43**: 1096-1105

Kraakman AT, Niks RE, Van den Berg PM, Stam P, and Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics **168**: 435-446

Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppala J, Koch M, Mitchell-Olds T, Langley CH, and Savolainen O (2004) Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. Genetics **168**: 1575-1584

Lam HM, Coschigano K, Schultz C, Melo-Oliveira R, Tjaden G, Oliveira I, Ngai N, Hsieh MH, and Coruzzi G (1995) Use of *Arabidopsis* mutants and genes to study amide amino acid biosynthesis. Plant Cell. **7**: 887-898

Lander ES and Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185-199

Larkins BA, Mason AC, and Hurkman WJ (1982) Molecular mechanisms regulating the synthesis of storage proteins in maize endosperm. Crit. Rev. Food Sci. Nutr. **16**: 199-215

Liang P and Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science **257**: 967-971

Liu J, Van EJ, Cong B, and Tanksley SD (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc.Natl.Acad.Sci U.S.A **99**: 13302-13306

Liu S, Tinker NA and Mather DE. (2006) Exact word matches in rice pseudomolecules. Genome. **49**:1047-1051

Lynch M and Walsh B (1997) Genetics and Analysis of Quantitative Traits. American Sinauer Associates, Inc., Sunderland, MA: 980 pages

Ma BL, Leibovitch S, Smith DL (1994) Plant growth regulator effects on protein content and yield of spring barley and wheat. J.Agron.& Crop Sci **172**: 9-18

McCallum CM, Comai L, Greene EA, and Henikoff S (2000) Targeting induced local lesions IN genomes (TILLING) for plant functional genomics. Plant Physiol. **123**: 439-442

McCouch SR and Doerge RW (1995) QTL mapping in rice. Trends Genet. **11**: 482-487

MedicineNet. www.medicinenet.com, 2006

Mekhedov S, de Ilarduya OM, and Ohlrogge J (2000) Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. Plant Physiol. **122**: 389-402

Melchinger AE, Utz HF, and Schon CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics **149**: 383-403

Molecular systematics and evolution. http://www.dbbm.fiocruz.br/james/GlossaryI.html, 2006

Morton NE (2005) Linkage disequilibrium maps and association mapping. J. Clin.Invest. **115**: 1425-1430

National Human Genome Research Institute. http://www.genome.gov, 2006

Ness SA (2006) Basic microarray analysis: strategies for successful experiments. Methods Mol Biol. **316**: 13-33

O'Donoughue LS, Wang Z, Röder M, Kneen B, Leggett M, Sorrells ME, Tanksley SD (1992) An RFLP-based linkage map of oats based on a cross between two diploid taxa (*Avena atlantica* × *A. hirtula*). Genome **35**: 765-771

O'Donoughue LS, Kianian SF, Rayapati J, Penner GA, Sorrells ME, Tanksley SD, Phillips RL, Rines HW, Lee M, Fedak G, Molnar SJ, Hoffman DL, Murphy P, Wu BC, Autrique E, Van Deyzne A (1995) A molecular linkage map of cultivated oat. Genome **38**: 368-380

Ohlrogge J, Jaworski J, Post-Beittenmlller D, Roughan G, Roessler P, Nakahlra K. (1993) Regulation of flux through the fatty acid biosynthesis pathway. *In* Biochemistry and molecular biology of membrane and storage lipids of plants: 102-112. American Society of Plant Physiologists, Rockville, MD. Somerville C. (ed)

Ohlrogge J and Browse J (1995) Lipid biosynthesis. Plant Cell. **7**: 957-970

Ohlrogge JB and Jaworski JG (1997) Regulation of fatty acid synthesis. Annu.Rev.Plant Physiol Plant Mol.Biol. **48**: 109-136

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, and van OA (2002) Regulation of noise in the expression of a single gene. Nat.Genet **31**: 69-73

Paran I and Zamir D (2003) Quantitative traits in plants: beyond the QTL. Trends Genet. **19**: 303-306

Perry HJ and Harwood JL (1991) Lipid metabolism during seed development in oilseed rape (*Brassica napus* L. cv. Shiralee). Biochem.Soc.Trans. **19**: 243S

Peterman TK and Goodman HM (1991) The glutamine synthetase gene family of *Arabidopsis thaliana*: light-regulation and differential expression in leaves, roots and seeds. Mol.Gen.Genet. **230**: 145-154

Petersohn A, Brigulla M, Haas S, Hoheisel JD, Volker U, and Hecker M (2001) Global analysis of the general stress response of *Bacillus subtilis*. J. Bacteriol.**183**: 5617-5631

Peterson DM and Brinegar AC (1986) Oat storage proteins. *In* Oats Chemistry and Technology: 153-203. American Association of Cereal Chemistry, St. Paul, MN. Webster FH (ed)

Peterson DM (1992) Composition and nutritional characteristics of oat grain and oat products. *In* Oat Science and Technology: 265-292. American Society of Agronomy and Crop Science Society of America, Madison, WI. Marshall HG, Sorrells ME (ed)

Plant Ontology Consortium. http://dev.plantontology.org/docs/growth/growth.html, 2006

Podkowinski J, Sroga GE, Haselkorn R, and Gornicki P (1996) Structure of a gene encoding a cytosolic acetyl-CoA carboxylase of hexaploid wheat. Proc.Natl.Acad.Sci U.S.A **93**: 1870-1874

Pomp D, Allan MF, and Wesolowski SR (2004) Quantitative genomics: exploring the genetic architecture of complex trait predisposition. J. Anim.Sci. **82** E-Suppl: E300-E312

Poroyko V, Hejlek LG, Spollen WG, Springer GK, Nguyen HT, Sharp RE, and Bohnert HJ. (2005) The maize root transcriptome by serial analysis of gene expression. Plant Physiol **138**:1700-1710

Portyanko VA, Chen G, Rines HW, Phillips RL, Leonard KJ, Ochocki GE, and Stuthman DD (2005) Quantitative trait loci for partial resistance to crown rust, *Puccinia coronata*, in cultivated oat, *Avena sativa* L. Theor.Appl.Genet. **111**: 313-324

Prioul JL, Quarrie S, Causse M, and de Vienne D (1997) Dissecting complex physiological functions into elementary components through the use of molecular quantitative genetics. J. Exp.Bot. **48**: 1151-1163

Quarrie SA, Steed A, Calestani C, Semikhodskii A, Lebreton C, Chinoy C, Steele N, Pljevljakusic D, Waterman E, Weyen J, Schondelmaier J, Habash DZ, Farmer P, Saker L, Clarkson DT, Abugalieva A, Yessimbekova M, Turuspekov Y, Abugalieva S, Tuberosa R, Sanguineti MC, Hollington PA, Aragues R, Royo A , and Dodig D (2005) A high-density genetic map of hexaploid wheat (*Triticum aestivum* L.) from the cross Chinese Spring x SQ1 and its use to compare QTLs for grain yield across a range of environments. Theor.Appl.Genet. **110**: 865-880

Raser JM and O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. Science **304**: 1811-1814

Raser JM and O'Shea EK (2005) Noise in gene expression: origins, consequences, and control. Science **309**: 2010-2013

Rayapati PJ, Gregory JW, Lee M, Wise RP (1994) A linkage map of diploid *Avena* based on RFLP loci and a locus conferring resistance to nine isolates of *Puccinia coronata* var. *avenae*. Theor.Appl.Genet. **89**: 831-837

Ripsin CM, Keenan JM, Jacobs DR, Jr., Elmer PJ, Welch RR, Van HL, Liu K, Turnbull WH, Thye FW, Kestin M (1992) Oat products and lipid lowering. A meta-analysis. JAMA **267**: 3317-3325

Robinson SJ, Cram DJ, Lewis CT, and Parkin IA (2004) Maximizing the efficacy of SAGE analysis identifies novel transcripts in Arabidopsis. Plant Physiol. **136**: 3223-3233

Robitaille J, Fontaine-Bisson B, Couture P, Tchernof A, Vohl MC (2005) Effect of an oat bran-rich supplement on the metabolic profile of overweight premenopausal women. Ann. Nutr. Metab. **49**: 141-148

Rhodius VA and LaRossa RA (2003) Uses and pitfalls of microarrays for studying transcriptional regulation. Curr Opin Microbiol. **6**: 114-119

Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132, 365-386

Saitou N and Nei M (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol.Biol.Evol. **4**: 406-425

Sambrook J and Russell D (2001) Molecular Cloning III: A Laboratory Manual. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, New York: 2,344 pages

Savage LJ, Ohlrogge JB (1999) Phosphorylation of pea chloroplast acetyl-CoA carboxylase. Plant J **18**: 521-527

Scarr-Salapatek S (1971) Race, social class, and IQ. Science **174**: 1285-1295

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, and Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature **422**: 297-302

Schipper H and Frey KJ (1991) Observed gains from three recurrent selection regimes for increased groat-oil content of oat. Crop Sci **31**: 1505-1510

Schmalhausen II (1949) Factors of Evolution: the Theory of Stabilizing Selection. University of Chicago Press, Chicago: 451 pages

Schrickel DJ (1986) Oats Production, Value and Use. *In* Oats Chemistry and Technology: 1-11. American Association of Cereal Chemistry, St. Paul, MN. Webster FH (ed)

Schulte W, Topfer R, Stracke R, Schell J, and Martini N (1997) Multi-functional acetyl-CoA carboxylase from *Brassica napus* is encoded by a multi-gene family: indication for plastidic localization of at least one isoform. Proc.Natl.Acad.Sci U.S.A **94**: 3465-3470

Selmanoff MK, Jumonville JE, Maxson SC, and Ginsburg BE (1975) Evidence for a Y chromosomal contribution to an aggressive phenotype in inbred mice. Nature **253**: 529-530

Sharopova NR, Portyanko VA, and Sozinov AA (1998) Genetics of alpha-amylases in hexaploid oat species. Biochem.Genet. **36**: 171-182

Shotwell MA, Boyer SK, Chesnut RS, and Larkins BA (1990) Analysis of seed storage protein genes of oats. J Biol.Chem. **265**: 9652-9658

Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA. (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. **25**: 1087-1088

Siripoonwiwat W, O'Donoughue LS, Wesenberg D, Hoffman DL, Barbosa-Neto JF, and Sorrells ME (1996) Chromosomal regions associated with quantitative traits in oat. J. Quant.Trait Loci **2** (available at http://www.cabi-publishing.org/jag/papers96/paper396/oatqtl3g.html)

Slabas AR, White A, O'hara P, Fawcett T (2002) Investigations into the regulation of lipid biosynthesis in Brassica napus using antisense down-regulation. Biochem Soc Trans. **30**:1056-1059

Slabaugh M, Leonard J, and Knapp S (1998) Condensing enzymes from Cuphea wrightii associated with medium chain fatty acid biosynthesis. Plant J. **13**: 611–620

Slade AJ and Knauf VC (2005) TILLING moves beyond functional genomics into crop improvement. Transgenic Res. **14**: 109-115

Southern EM (1996) High-density gridding: techniques and applications. Curr Opin Biotechnol. **7**: 85-88

Swain PS, Elowitz MB, and Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc.Natl.Acad.Sci. U.S.A **99**: 12795-12800

Tai H, Jaworski JG (1993) 3-Ketoacyl-acyl carrier protein synthase III from spinach (*Spinacia oleracea*) is not similar to other condensing enzymes of fatty acid synthase. Plant Physiol. **103**: 1361–1367

Takahashi Y, Shomura A, Sasaki T, and Yano M (2001) Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. Proc. Natl. Acad. Sci. U.S.A **98**: 7922-7927

Tapola N, Karvonene H, Niskanen L, Mikola M, Sarkkinen E (2005) Glycemic response of oat bran products in type 2 diabetic patients. Nutr. Metab. Cardiovasc. Dis. **15**: 255-261

The NCBI Handbook Glossary. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.glossary.1237, 2006

Thompson JD, Higgins DG, and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**: 4673-4680

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, and Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat.Genet. **28**: 286-289

Thro AM, Frey KJ, and Hammond EG (1985) Inheritance of palmitic, oleic, linoleic, and linolenic fatty acids in groat oil of oats. Crop Sci. **25**: 40-44

Till BJ, Colbert T, Tompa R, Enns LC, Codomo CA, Johnson JE, Reynolds SH, Henikoff JG, Greene EA, Steine MN, Comai L, and Henikoff S (2003) High-throughput TILLING for functional genomics. Methods Mol. Biol. (Clifton, N.J.) **236**: 205-220

Tinker NA and Mather DE (1995) MQTL: Software for simplified composite interval mapping of QTL in multiple environments. J. Ag.Genomics (available at http://www.cabi-publishing.org/jag/papers95/paper295/jqtl16r2.html)

Turner A, Beales J, Faure S, Dunford RP, and Laurie DA (2005) The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. Science **310**: 1031-1034

Vasemagi A and Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol.Ecol. **14**: 3623-3642

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial Analysis Of Gene Expression. Science **270**: 484-487

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. Cell **88**: 243-251

Vicient CM, Jaaskelainen MJ, Kalendar R, and Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol. **125**: 1283-1292

Vicient CM, Kalendar R, and Schulman AH (2001) Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. Genome Res. **11**: 2041-2049

Vikkula M, Metsaranta M, Syvanen AC, la-Kokko L, Vuorio E, and Peltonen L (1992) Structural analysis of the regulatory elements of the type-II procollagen gene. Conservation of promoter and first intron sequences between human and mouse. Biochem.J. **285** (Pt 1): 287-294

Voelker T and Kinney AJ (2001) Variations in the biosynthesis of seed storage lipids. Annu.Rev.Plant Physiol Plant Mol.Biol. **52**: 335-361

Wayne ML and McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. Proceedings of the National Academy of Sciences of the United States of America **99**: 14903-14906

Weber A and Jung K (2002) Profiling early osmostress-dependent gene expression in Escherichia coli using DNA macroarrays. J. Bacteriol. **184**: 5502-5507

Wight CP, Tinker NA, Kianian SF, Sorrells ME, O'Donoughue LS, Hoffman DL, Groh S, Scoles GJ, Li CD, Webster FH, Phillips RL, Rines HW, Livingston SM, Armstrong KC, Fedak G, and Molnar SJ (2003) A molecular marker map in 'Kanota' x 'Ogle' hexaploid oat (*Avena* spp.) enhanced by additional markers and a robust framework. Genome **46**: 28-47

Wolff GL (1955) The effect of environmental temperature on coat color in diverse genotypes of the Guinea pig. Genetics **40**: 90-106

Xu M (2005) Theoretically modeling microarray with the chemical equilibrium and thermodynamics. J. Bioinform.Comput.Biol. **3**: 477-490

Yao Y, Ni Z, Zhang Y, Chen Y, Ding Y, Han Z, Liu Z, and Sun Q (2005) Identification of differentially expressed genes in leaf and root between wheat hybrid and its parental inbreds using PCR-based cDNA subtraction. Plant Mol. Biol. **58**: 367-384

Yoshimura A, Ideta O, and Iwata N (1997) Linkage map of phenotype and RFLP markers in rice. Plant Mol.Biol. **35**: 49-60

Youngs VL (1972) Protein distribution in the oat kernel. Cereal Chem. **49**: 407-411

Youngs VL (1986) Oat lipids and lipid-related enzymes. *In* Oats Chemistry and Technology: 205-226. American Association of Cereal Chemistry, St. Paul, MN. Webster FH (ed)

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li S, Li X, Zheng H, Cong L, Lin L, Yin J, Geng J, Li G, Shi J, Liu J, Lv H, Li J, Wang J, Deng Y, Ran L, Shi X, Wang X, Wu Q, Li C, Ren X, Wang J, Wang X, Li D, Liu D, Zhang X, Ji Z, Zhao W, Sun Y, Zhang Z, Bao J, Han Y, Dong L, Ji J, Chen P, Wu S, Liu J, Xiao Y, Bu D, Tan J, Yang L, Ye C, Zhang J, Xu J, Zhou Y, Yu Y, Zhang B, Zhuang S, Wei H, Liu B, Lei M, Yu H, Li Y, Xu H, Wei S, He X, Fang L, Zhang Z, Zhang Y, Huang X, Su Z, Tong W, Li J, Tong Z, Li S, Ye J , Wang L, Fang L, Lei T, Chen C, Chen H, Xu Z, Li H, Huang H, Zhang F, Xu H, Li N, Zhao C, Li S, Dong L, Huang Y, Li L, Xi Y, Qi Q, Li W, Zhang B, Hu W, Zhang Y, Tian X, Jiao Y, Liang X, Jin J, Gao L, Zheng W, Hao B, Liu S, Wang W, Yuan L, Cao M, McDermott J, Samudrala R, Wang J, Wong GK, and Yang H (2005) The Genomes of *Oryza sativa*: a history of duplications. PLoS Biol. **3**: e38

Yun SJ, Martin DJ, Gengenbach BG, Rines HW, and Somers DA (1993) Sequence of a (1-3,1-4)-beta-glucanase cDNA from oat. Plant Physiol. **103**: 295-296

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, and Kruglyak L (2003) *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat.Genet. **35**: 57-64

Zeng ZB (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci . Proc.Natl.Acad.Sci. U.S.A **90**: 10972-10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics **136**: 1457-1468

Zhou S, Glowacki J, and Yates KE (2004) Comparison of TGF-beta/BMP pathways signaled by demineralized bone powder and BMP-2 in human dermal fibroblasts. J. Bone Miner.Res. **19**: 1732-1741

Zhu S and Kaeppler HF (2003a) A genetic linkage map for hexaploid, cultivated oat (*Avena sativa* L.) based on an intraspecific cross 'Ogle/MAM17-5'. Theor.Appl.Genet. **107**: 26-35

Zhu S and Kaeppler HF (2003b) Identification of Quantitative Trait Loci for Resistance to Crown Rust in Oat Line MAM17-5. Crop Sci. **43**: 358-366

Zhu S, Kolb FL, and Kaeppler HF (2003c) Molecular mapping of genomic regions underlying barley yellow dwarf tolerance in cultivated oat (*Avena sativa* L.). Theor.Appl.Genet. **106**: 1300-1306

Zhu S, Leonard KJ, Kaeppler HF. (2003d) Quantitative trait loci associated with seedling resistance to isolates of *Puccinia coronata* in oat. Phytopathology **93**: 860-866

Zhu S, Rossnagel BG, and Kaeppler HF (2004) Genetic Analysis of Quantitative Trait Loci for Groat Protein and Oil Content in Oat. Crop Sci. **44**: 254-260

Zou J, Katavic V, Giblin EM, Barton DL, MacKenzie SL, Keller WA, Hu X, and Taylor DC (1997) Modification of seed oil content and acyl composition in the brassicaceae by expression of a yeast sn-2 acyltransferase gene. Plant Cell. **9**: 909-923

Internal Licence Number/Numéro de permis interne: AAFC03-016

NUCLEAR SUBSTANCES / SUBSTANCES NUCLÉAIRES

The total quantity of an unsealed nuclear substance in possession shall not exceed the corresponding listed unsealed source maximum quantity. La quantité totale d'une substance nucléaire non scellée possédée ne doit pas excéder la quantité maximale qui est indiquée pour une source non scellée correspondante.

Item/Article: 5

Nucl. Subst./Subst. nucl.	*MaxQty/Qté*	*Room/Endroit*
Carbon 14	4 GBq	B69B; Bldg20: 2044, 4018, 4020, 4006 (Freezer), 4015 (Cold Room); Bldg21: 8, 11, 13
Hydrogen 3	40 GBq	B69B; Bldg20: 2032, 2034, 2035, 4018, 4020, 4006 (Freezer), 4015 (Cold Room); Bldg21: 8, 11, 13
Phosphorus 32	4 GBq	B69B; Bldg20: 1010, 2028, 2032,2034, 2036, 2039, 2041, 2043, 2044, 2048, 2048A, 4012, 4018, 4020, 4006 (Freezer), 4015 (Cold Room); Bldg21: 8, 11, 13; Bldg50: 4, 16
Phosphorus 33	1 GBq	B69B; Bldg20: 1010, 1018, 1020, 2028, 2036, 2039, 2041, 2043, 2044, 2048, 2048A, 4012; Bldg21: 8, 11, 13; Bldg50: 4, 16
Sulfur 35	2 GBq	B69B; Bldg20: 1010, 1026, 1028, 2028, 2032, 2034, 2035, 2036, 2039, 2041, 2043, 2044, 4018, 4020, 4006 (Freezer, 4015 (Cold Room), Bldg21:8, 11, 13

Centre: Ottawa

Internal Licence Number/Numéro de permis interne: AAFC03-016

USERS LIST /LISTE DES UTILISATEURS

Users/Utilisateurs: 64

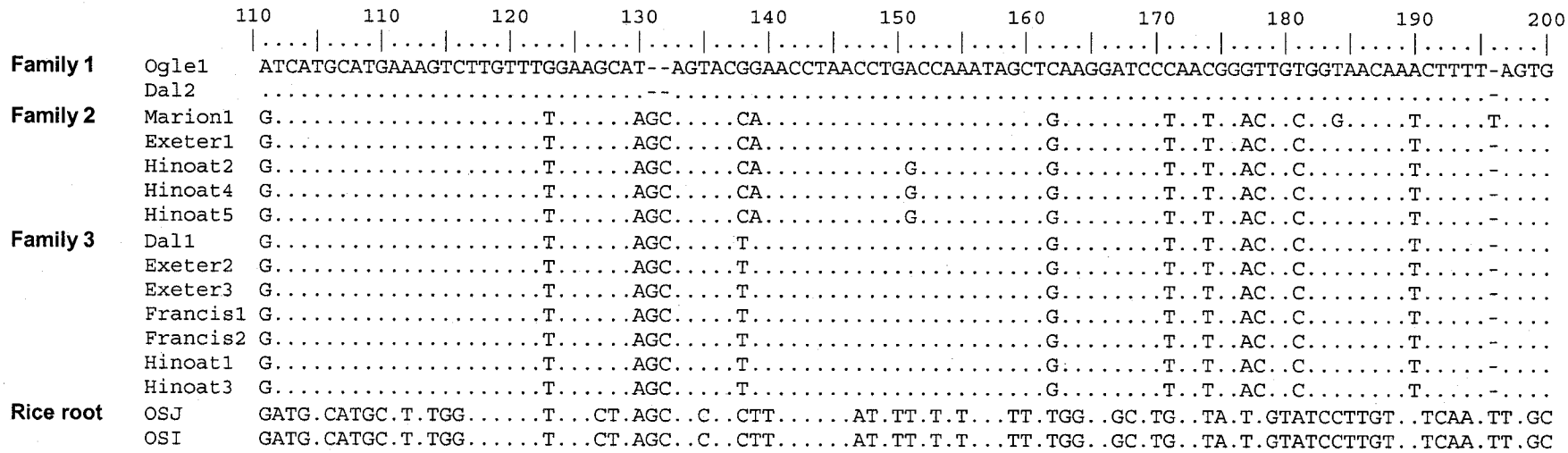
Users/ Utilisateur	*Nuclear Substance/Substance Nucléaire*
A. DeMoors	Phosphorus 32; Phosphorus 33; Sulfur 35
A. Hermans	Carbon 14; Hydrogen 3; Phosphorus 32; Phosphorus 33; Sulfur 35
A. Johnston	Carbon 14; Hydrogen 3; Phosphorus 32; Phosphorus 33; Sulfur 35
A. Koul	Phosphorus 32; Phosphorus 33; Sulfur 35
A. Léaustic	Phosphorus 32; Phosphorus 33; Sulfur 35
A. Locatelli	Phosphorus 32; Phosphorus 33
A. Lybaert	Phosphorus 32; Phosphorus 33
A. McKay	Carbon 14; Hydrogen 3; Phosphorus 32; Sulfur 35
A. Saparno	Carbon 14; Hydrogen 3; Phosphorus 32; Phosphorus 33; Sulfur 35
B. Watson	Carbon 14; Phosphorus 32; Phosphorus 33; Sulfur 35
C. Bordeleau	Phosphorus 32; Phosphorus 33; Sulfur 35
C. Gagnon	Carbon 14; Hydrogen 3; Phosphorus 32; Phosphorus 33; Sulfur 35
C. Gibbs	Phosphorus 32; Phosphorus 33; Sulfur 35
C. Piché	Phosphorus 32; Phosphorus 33; Sulfur 35
C. Sauder	Phosphorus 32; Phosphorus 33; Sulfur 35
C. Séguin	Nickel 63
C. Wight	Phosphorus 32; Phosphorus 33
D. A. Suleimani	Phosphorus 33
D. Luckert	Phosphorus 32; Phosphorus 33

Centre: Ottawa

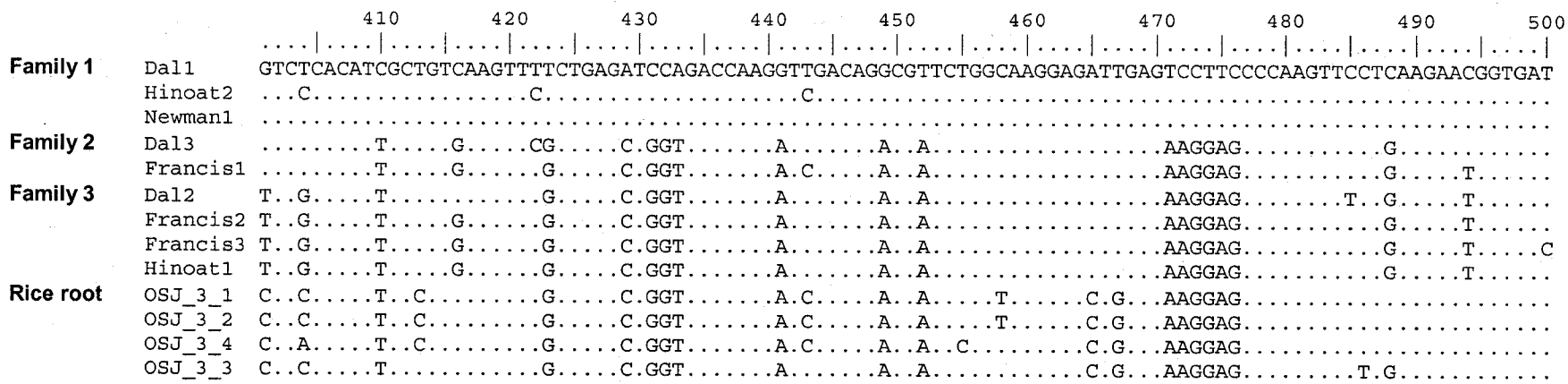
Pages 4-5-6-7

Appendix B: Sequence alignments

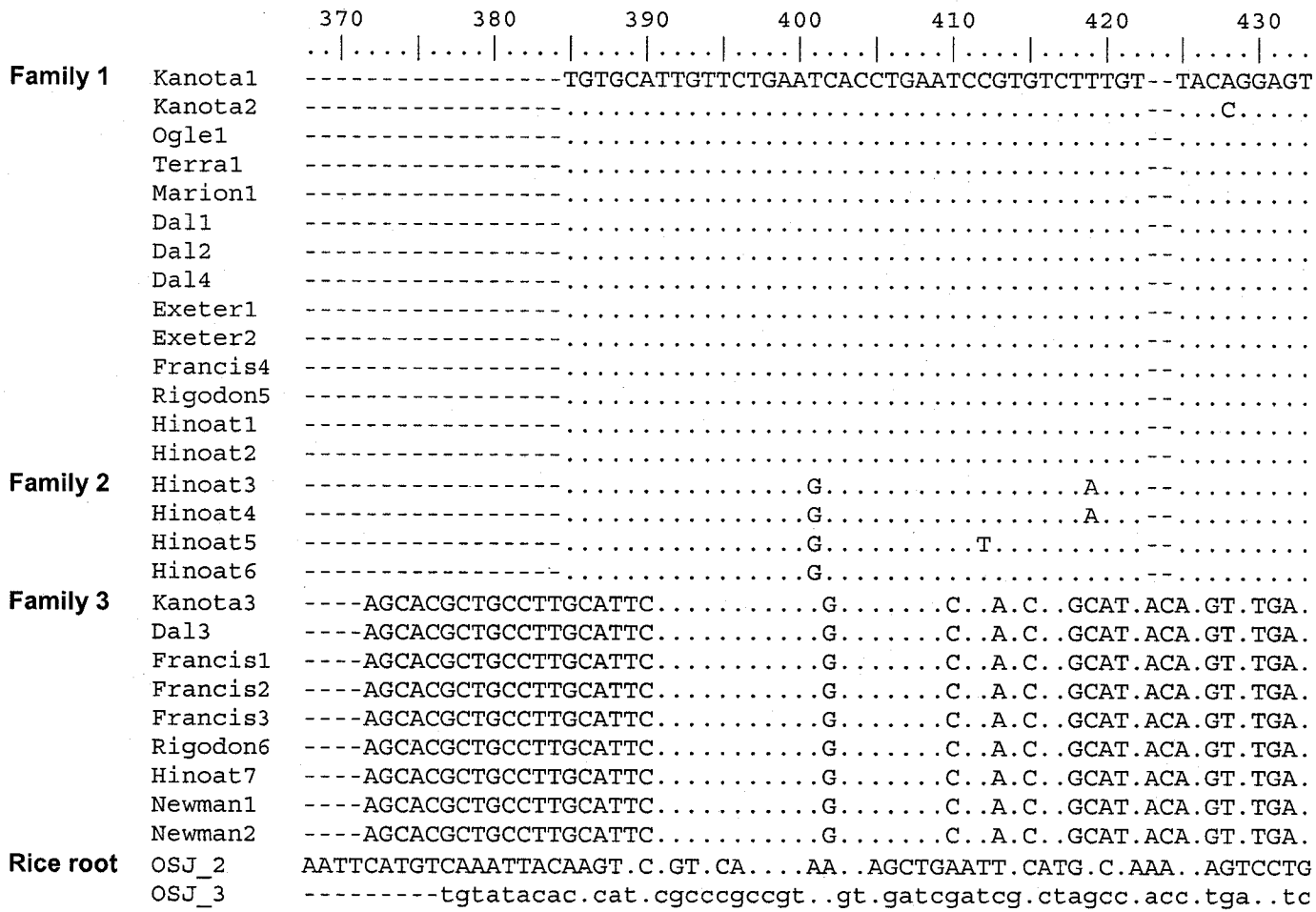
Partial alignments showing examples of nucleotide polymorphisms in oat sequences from six of the eight genes coding for products involved in lipid and protein biosynthesis, and not featured in Fig. 3.2. Identity to a standard sequence is shown by a dot and a missing nucleotide by a hyphen. Each sequence in these alignments is designated by the name of the cultivar from which the sequence was obtained, followed by a number to designate a particular sequence obtained for that cultivar. Family numbers on the left represent sequence families highlighted by the phylogenetic analysis of these sequences. Only a limited portion of the total alignment is represented for each gene. Only 44 of the 66 oat sequences included in the phylogenetic analysis of acetyl-CoA carboxylase are represented.



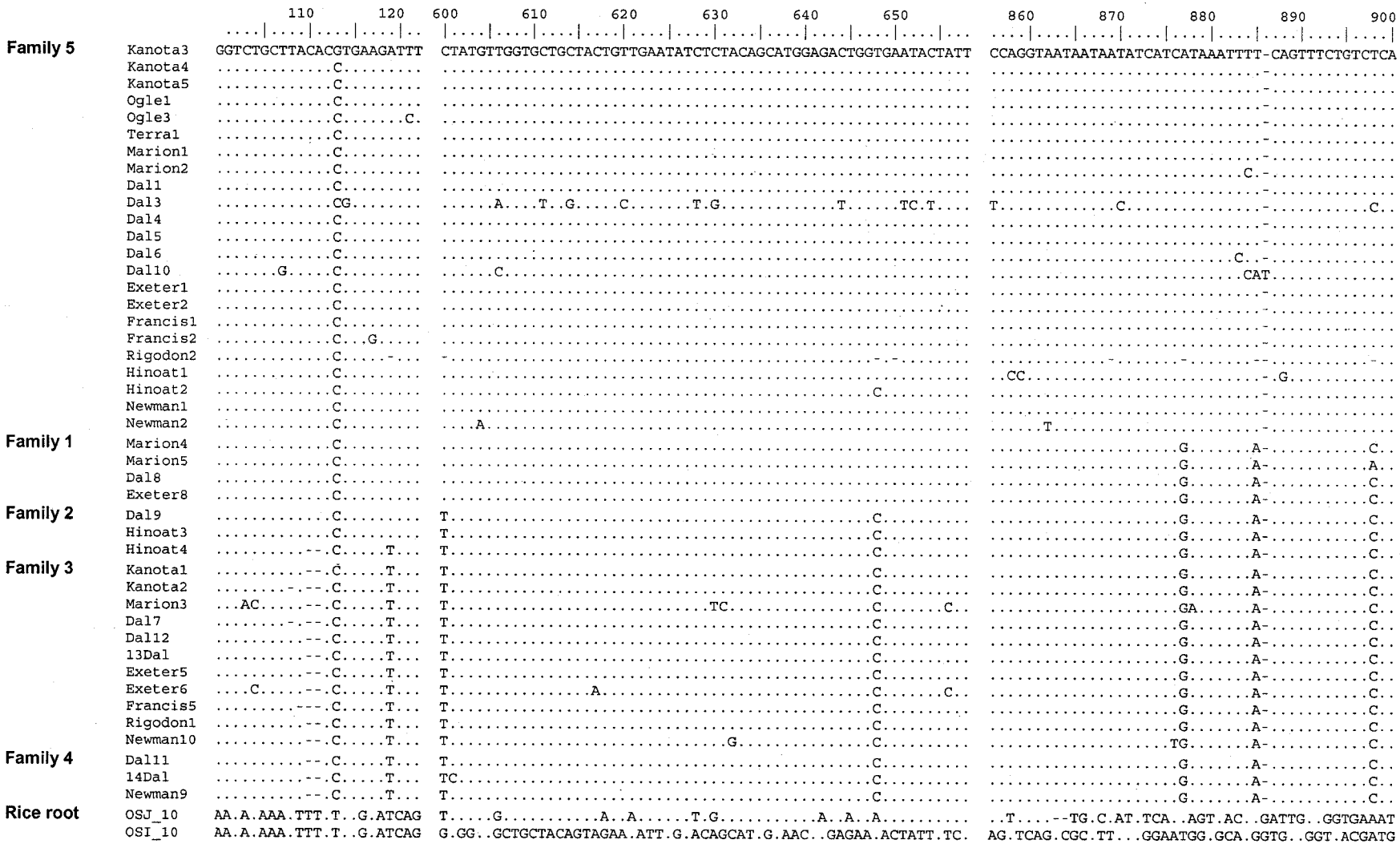
Acyl-carrier protein (part of intron 3)



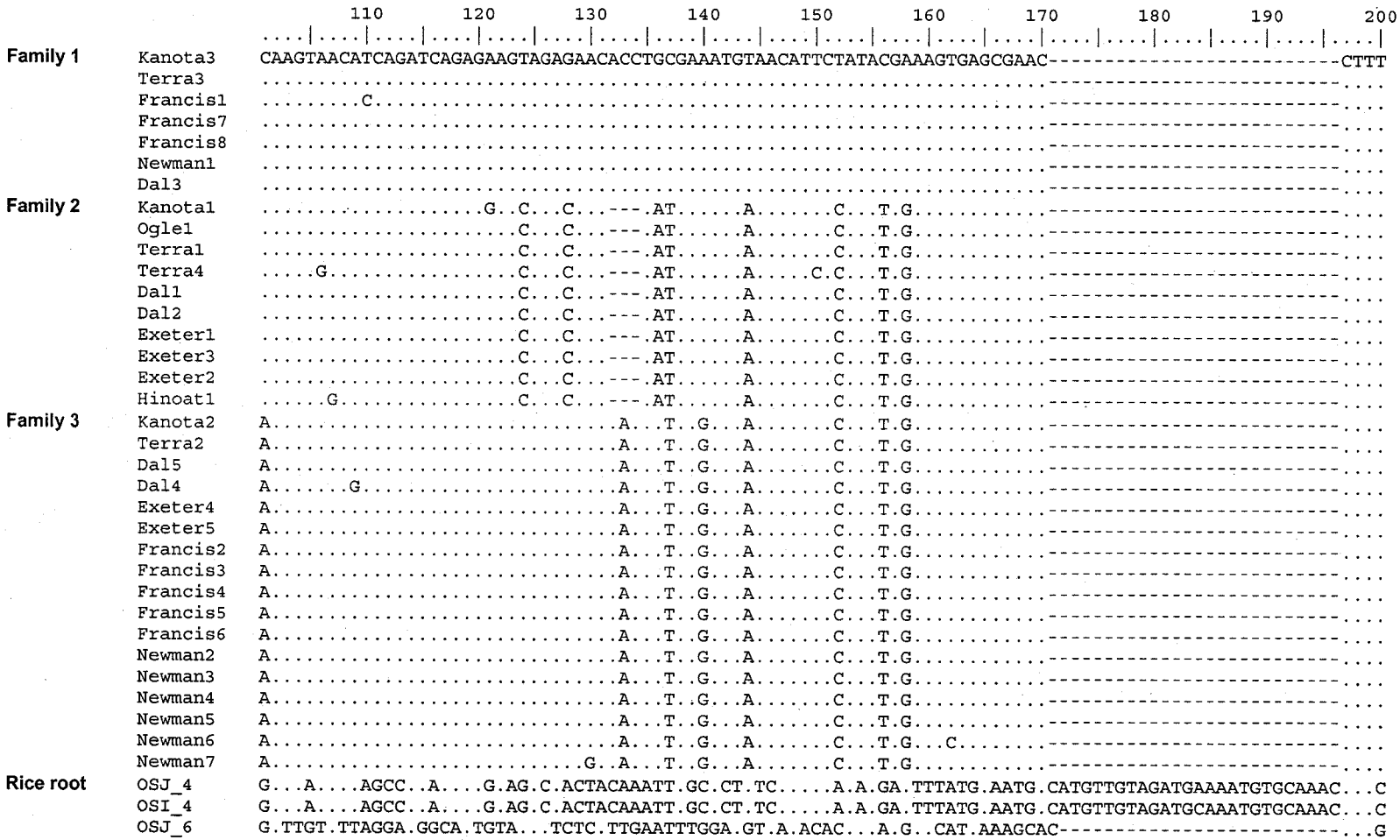
Protein elongation factor 1-*α* (part of exon 2)



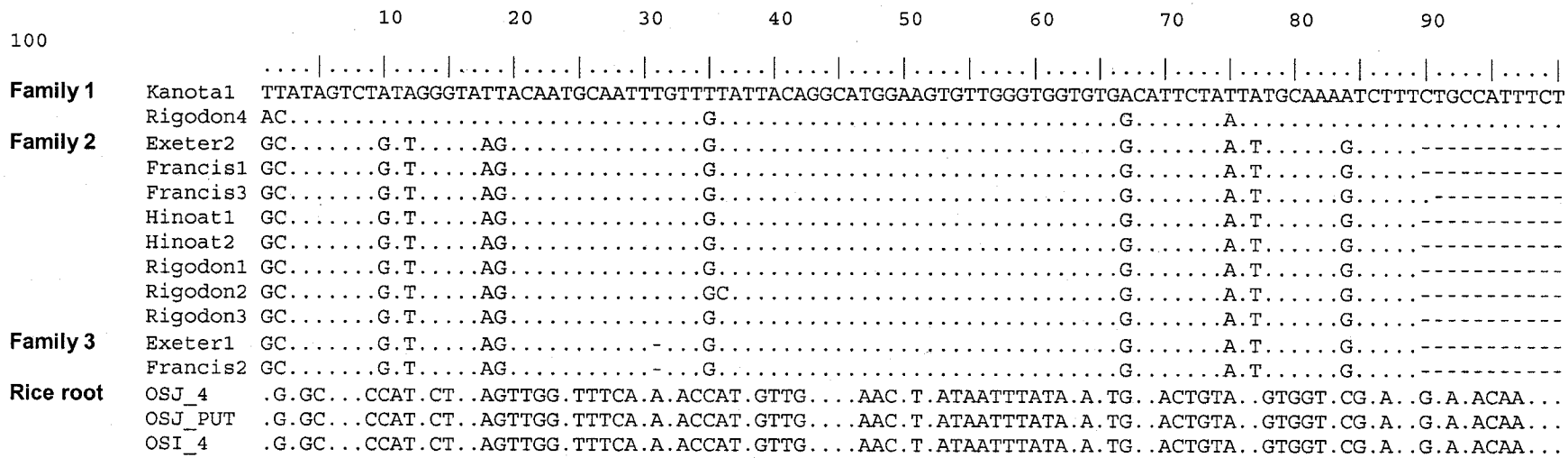
Glutamine synthase (part of intron 9)



Acetyl-CoA carboxylase (part of intron 5, exon 6 and exon 7)



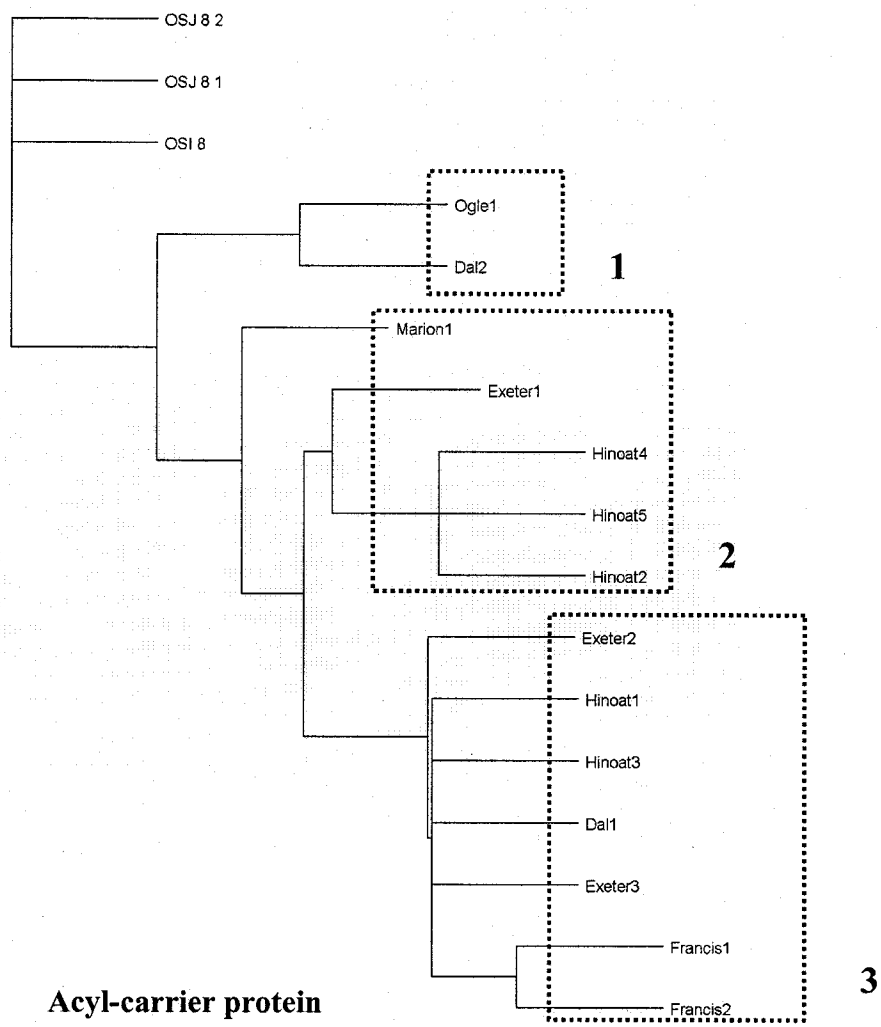
ß-ketoacyl-ACP synthase I (part of intron 6)



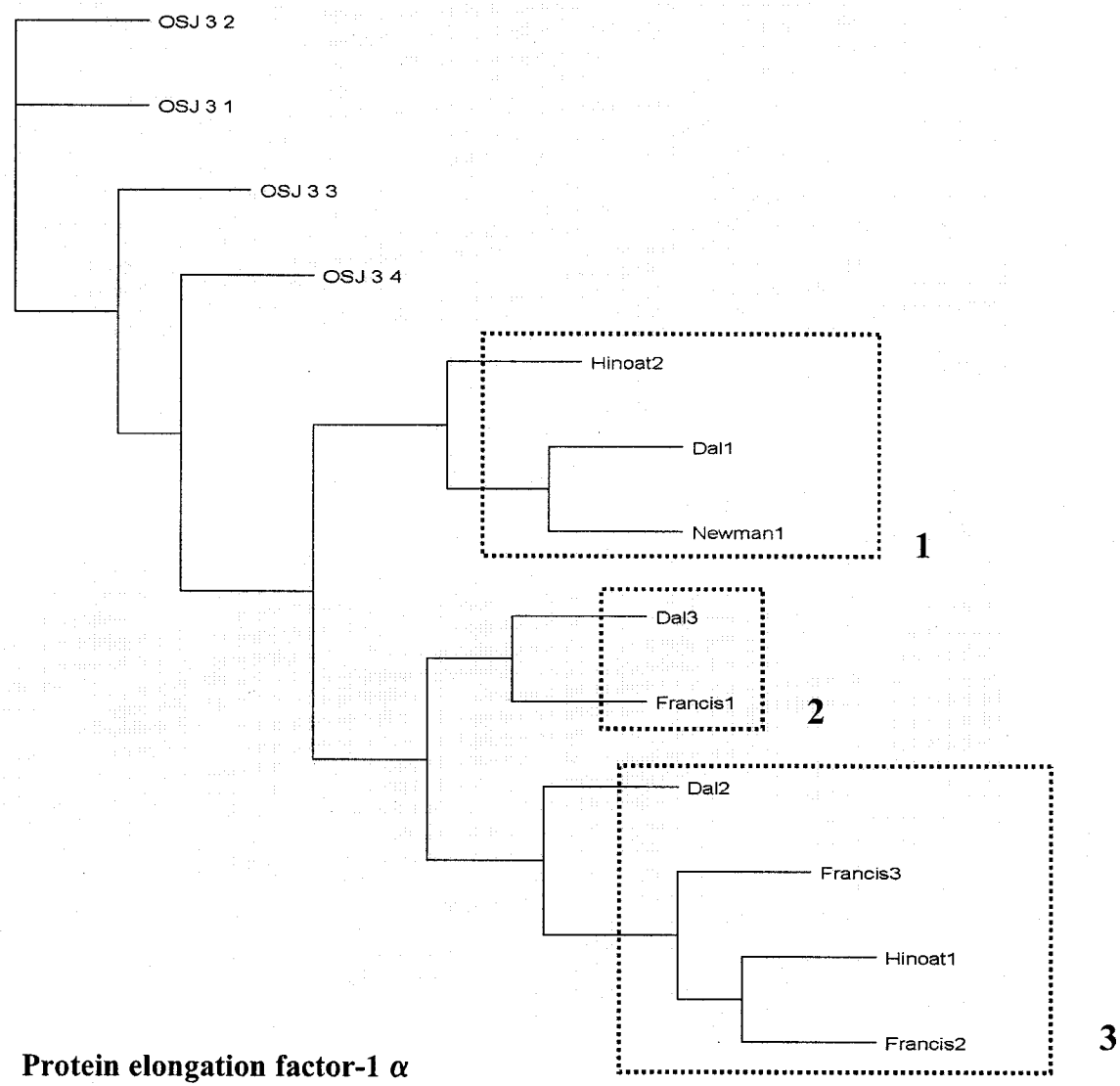
ß-ketoacyl-ACP synthase III (part of intron 5)

Appendix C: Phylograms

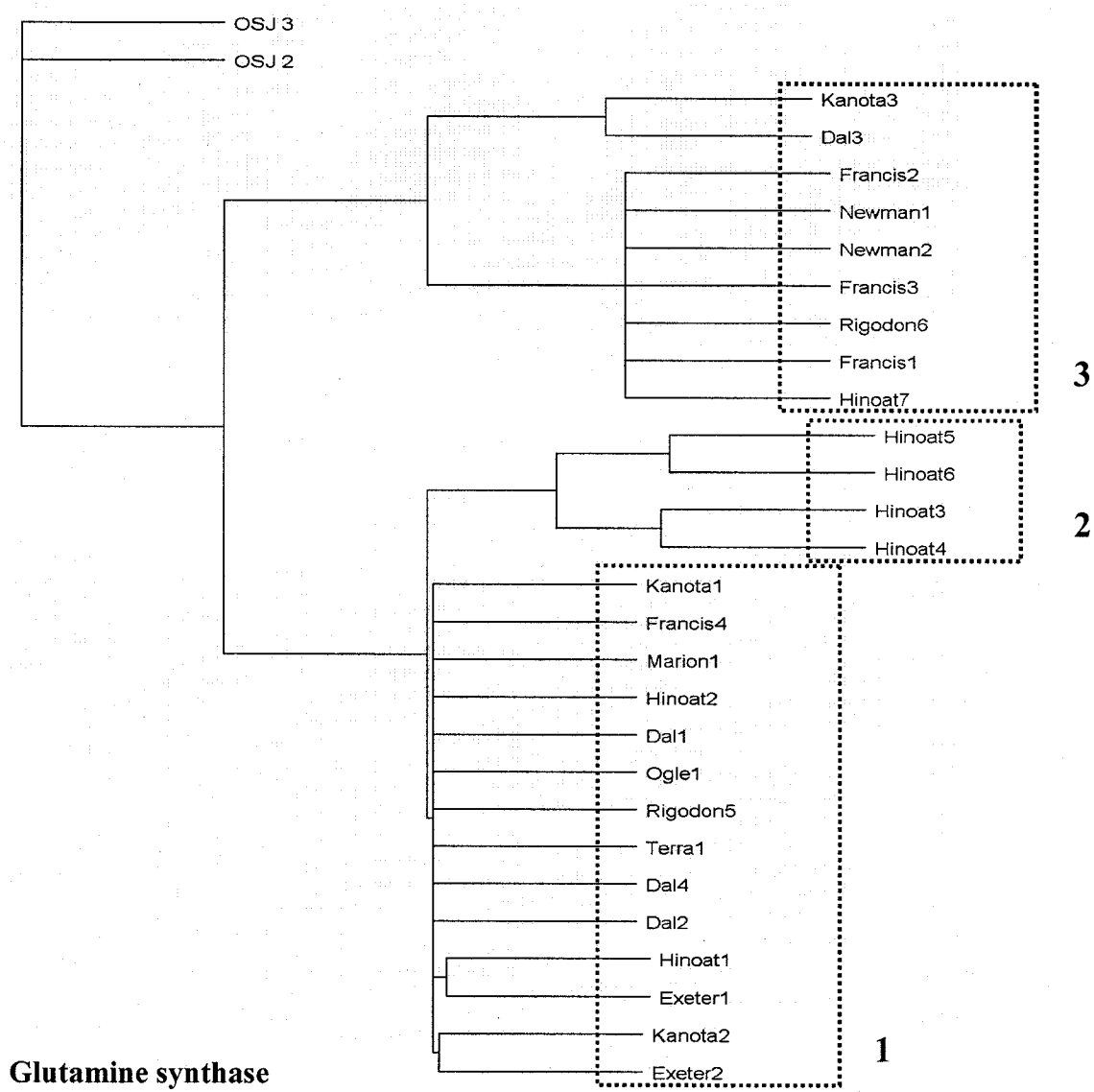
Phylograms of oat sequences for the six target genes coding for products involved in lipid and protein biosynthesis and not feathured in Fig 3.3. Each sequence in these alignments is designated by the name of the cultivar from which the sequence was obtained, followed by a number to designate a particular sequence obtained for that cultivar. Boxes represent sequence families. OSI and OSJ sequences are *Oryza sativa* sequences, respectively from cultivar groups *indica* and *japonica*, used as outgroups.



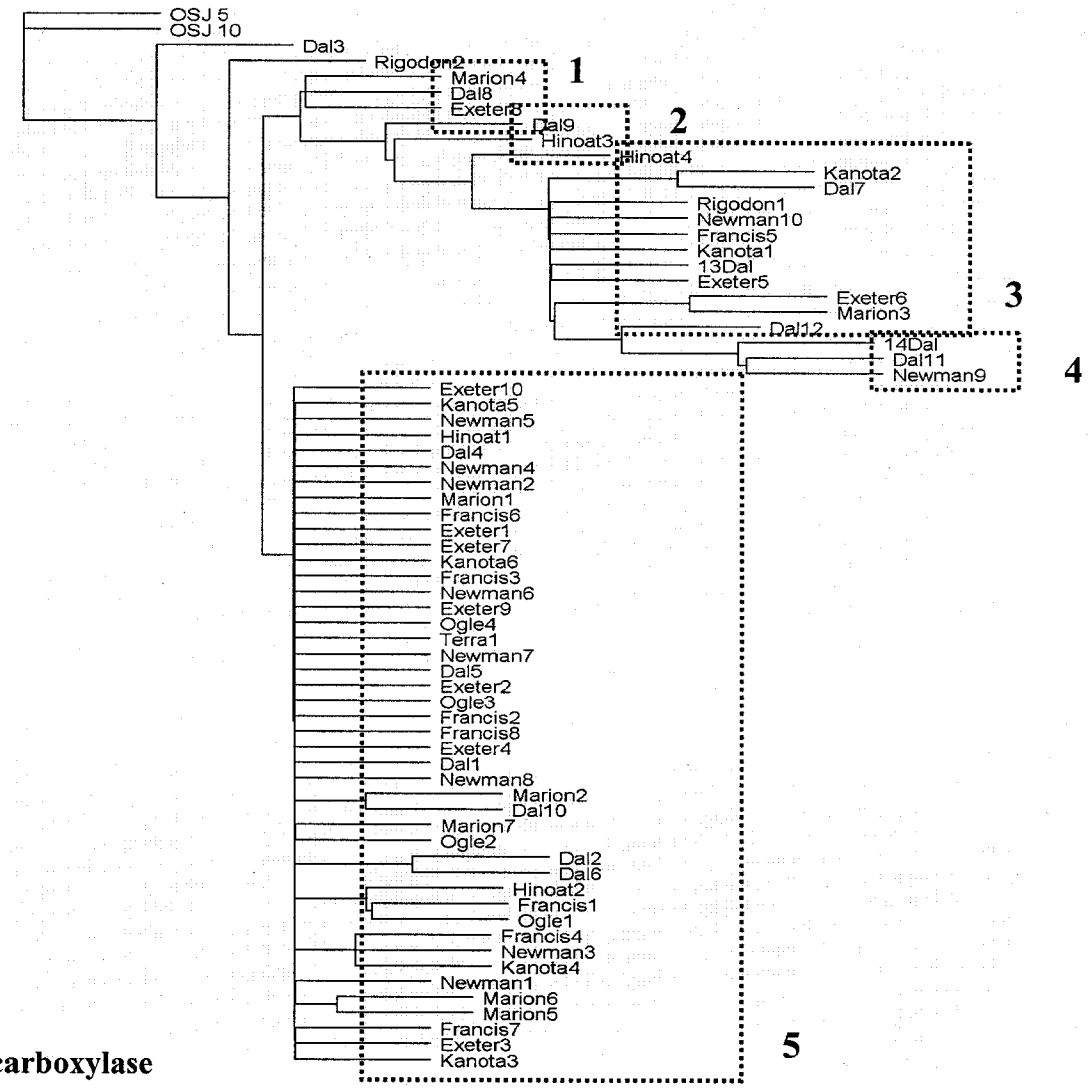
Acyl-carrier protein



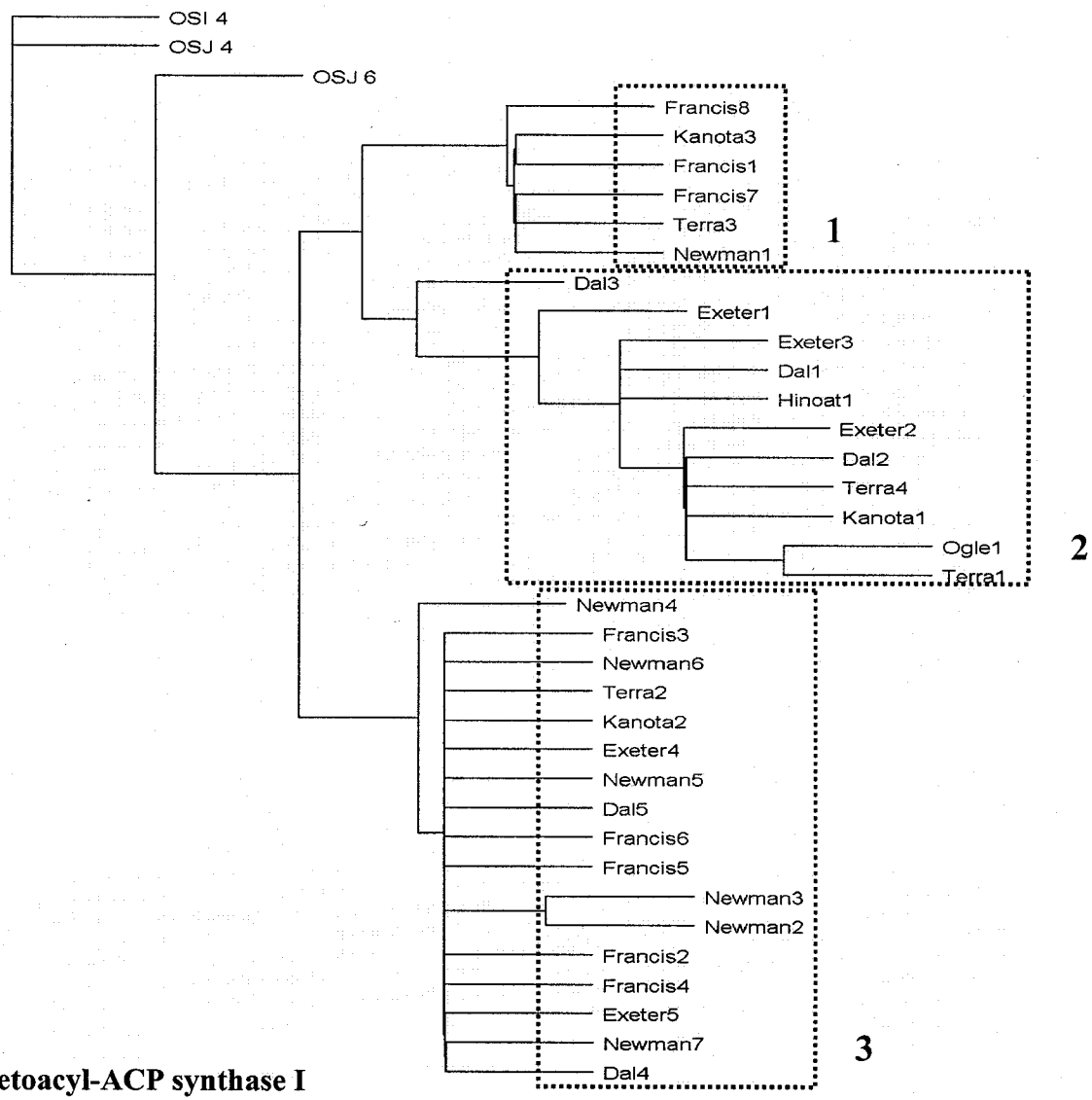
Protein elongation factor-1 α



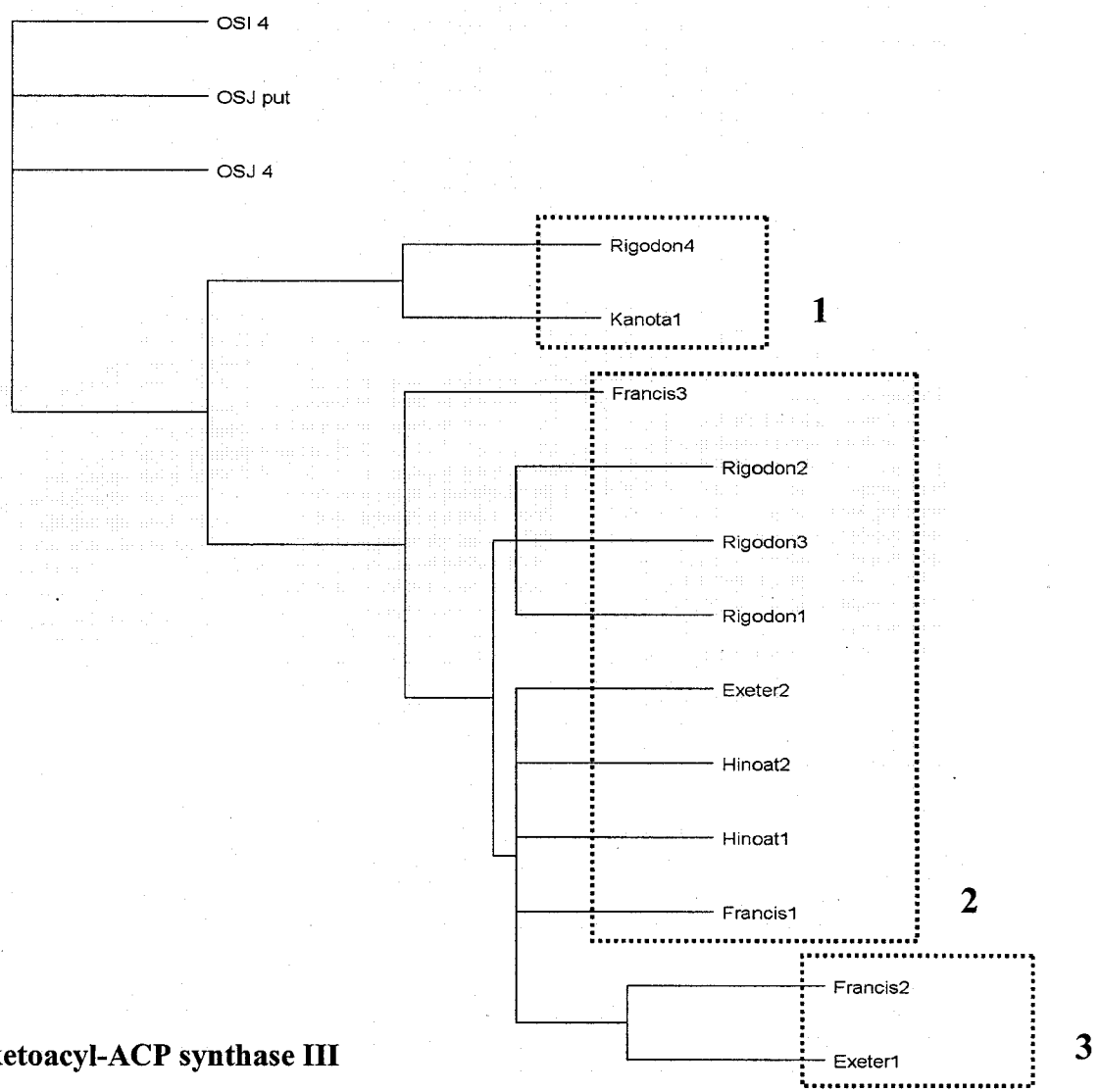
Glutamine synthase



Acetyl-CoA carboxylase



β-ketoacyl-ACP synthase I



β -ketoacyl-ACP synthase III