Comparative genomics of the relationship between gene structure and expression

Xin-Ying Ren

任新颖

Promotor:

Prof. Dr. W.J. Stiekema
Hoogleraar Genoominformatica
Wageningen Universiteit en Researchcentrum

Copromotor:

Dr. Ir. J.P. Nap
Senior onderzoeker
Plant Research International
Wageningen Universiteit en Researchcentrum

Promotiecommissie:

Prof. Dr. R.F. Hoekstra, Wageningen Universiteit
Prof. Dr. P. J. Weisbeek, Utrecht Universiteit
Prof. Dr. J.A.M. Leunissen, Wageningen Universiteit
Dr. R. C.H.J. van Ham, Plant Research International

Dit onderzoek is uitgevoerd binnen de onderzoeksschool “Experimental Plant Sciences”.

Comparative genomics of the relationship between gene structure and expression

Xin-Ying Ren

任新颖

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit
Prof. Dr. M.J. Kropff
in het openbaar te verdedigen
op woensdag 20 december 2006
des namiddags te half twee in de Aula

Ren, X.Y. (2006)

Comparative genomics of the relationship between gene structure and expression

PhD thesis, Wageningen University, the Netherlands
With summaries in English and Dutch

ISBN: 90-8504-554-1

Contents

Chapter 1

Introduction

Introduction

This thesis is about characterizing the relationship between gene structure and gene expression in genomes of higher eukaryotes. As part of the ‘Phytoinformatics, the added value from plants’ project within the Netherlands Organization for Scientific Research (NWO) Biomolecular Informatics program, the study was initially aimed at the functional annotation of the Arabidopsis genome focusing on the non-protein coding part. However, it soon became obvious that without the coding part of the genome and its expression, the function of the non-coding part of the genome cannot be properly analyzed. In addition, the analyses and conclusions would gain considerably in strength by adding the comparison with the genome of rice and other higher eukaryotes (worm, mouse, human). In all cases, extensive use has been made from publicly available genome annotation and expression datasets, notably the Massively Parallel Signature Sequencing (MPSS) data (mpss.udel.edu). Various parameters of gene configuration and structure, such as the presence of expressed neighboring genes or the number and size of introns, are correlated with expression data in several innovative ways in order to see if such relationships give more insight in the structure, function and possibly evolution of genomes of higher organisms.

The regulation of gene expression determines where, when and how all the activities inside the cells of a living organism are carried out. By studying and characterizing the relationships between the structure of genes and the regulation of gene expression with the help of bioinformatics approaches, we aim to contribute to the fundamentals of understanding genome organization and regulation. This increased understanding will contribute to the better use of these fundamentals in future wet-lab applications. It may help to find optimally expressed alleles in breeding populations or contribute to optimizing the expression of transgenes. A major issue in such studies is how to drive gene expression in the desired direction and how to control it properly. The comparison of widely-divergent genomes with respect to the correlation between gene structure and gene expression may help identifying the parameters and mechanisms that have helped to shape such correlations in evolution.

An introduction to the different topics related to the research presented in this thesis is given in the overview in **Chapter 2**. In the subsequent parts of the thesis, various aspects of gene expression are addressed in relation to genome organization and gene structure. In **Chapter 3,** the concept of local coexpression domains is introduced and defined. Expression of genes in eukaryotic genomes is known to cluster in domains, but domain size is generally loosely defined and highly variable. We introduce the strict requirement for a domain as a set of physically adjacent genes that are highly coexpressed with a pairwise Pearson's correlation coefficient larger than 0.7 to define local coexpression domains. The publicly available whole genome annotation and MPSS expression data of the dicotyledonous model plant *Arabidopsis thaliana* are used to analyze the occurrence of such domains. We identified 689 coexpression domains with the MPSS expression dataset. The domains consisted of two to four genes. A small (5%–10%), yet significant fraction of genes in the Arabidopsis genome is therefore organized into local coexpression domains. Genes in such local domains were for the major part not categorized in the same functional category (GOslim). Neither tandemly duplicated genes nor shared promoter sequence, nor gene distance explained the occurrence of coexpression of genes in such chromosomal domains. This indicates that other parameters in genes or gene positions are important to establish coexpression in local domains of the Arabidopsis genome.

Chapter 4 extends the approach to the occurrence of local coexpression domains in the genome of monocotyledonous model plant rice (*Oryza sativa*). Also in the rice genome, there are a small yet significant number of local coexpression domains that for the major part were not categorized in the same functional category (GOslim). Again, the various configuration parameters studies could not fully explain the occurrence of local coexpression domains.

Having identified local coexpression dmains in both rice and Arabidopsis, a comparative genomics approach was used to investigate the occurrence of syntenic coexpression domains between both genomes consisting of orthologous genes in both species. No such syntenic domains were formed between rice and Arabidopsis. This lack of microsynteny shows that maintenance of coexpression has not been an important driving force in evolution.

In **Chapter 5**, the relationships between the structure of the primary transcript and the expression level of the gene is investigated to describe the parameters and mechanisms that have helped shaping such correlations.

Combining whole genome annotations with expression datasets, we have developed a novel double ranking method to contrast higher and lower expressed genes. We show that in both rice and Arabidopsis, higher expressed genes have more and longer introns and a larger primary transcript than genes expressed at a lower level: higher expressed genes tend to be less compact than lower expressed genes. In animal genomes, it was reported to be the other way round. Therefore, the mechanisms explaining the relationship between gene configuration and gene expression based on animal data might be (or might have been) less important in plants. We speculate that selection, if any, on genome configuration has taken a different turn after the divergence of plants and animals.

A major issue in the conclusions with respect to gene structure in relationship to gene expression is the widely variable definitions and approaches used by the various research groups. Therefore, in **Chapter 6** we extend the approach documented in Chapter 5. An extensive comparative analysis of five widely diverse eukaryotic genomes (Arabidopsis, rice, worm, mouse and human), with the same definition of structural and expression parameters confirms that there is indeed a remarkable difference between plant and mammalian genes.

In the two mammalian genomes studied, higher expressed genes are more compact compared to their lower expressed counterparts. The difference in gene structure between mammals and plants is mainly due to the large differences in the length of introns. The possible explanations and consequences of this notable difference are discussed in terms of existing hypotheses and possible evolutionary steps.

In the final **Chapter 7**, the main findings of the research presented in this thesis are discussed in the context of the relationship between gene expression and gene configuration. We outline the necessary future work and future perspectives for this intriguing new aspect of gene regulation.

Chapter 2

Relationships between gene expression and gene structure in biological systems: an overview

Xin-Ying Ren & Jan-Peter Nap

Relationships between gene expression and gene structure in biological systems: an overview

With ever more genomes being sequenced and annotated and wealth of expression data being deposited in the public domain, the data is becoming available for detailed genome-wide analyses of the relationships, if any, between gene structure and position on the one hand, and gene expression on the other hand. This may reveal deeper levels of gene regulation and/or help elucidate the forces that shaped current genomes. Analyses of single or limited numbers of genes have shown that gene structure and position can affect the impact of that genetic information in a dynamic way. This introductory chapter presents an overview of the various parameters of structure and position that have been associated with gene expression. The material presented aims to define the more important parameters of gene structure and expression, also in an attempt to clarify and contribute to ongoing discussions. This sets the stage for the research on gene structure and expression presented in the subsequent chapters.

First, the structural components of DNA are described. The way of how DNA is organized in the nucleus into chromosomes is outlined and the functional determinants of genome sequences that can be distinguished in current genome descriptions are summarized. This is followed by an overview of current technology to analyze and study gene expression. At the end of the chapter, the combination of structure and expression studies in a genomics context is reviewed.

2.1. DNA to chromatin: structural considerations

A highly significant accomplishment in the history of biology was the discovery of the double helix structure of deoxyribonucleic acid (DNA) (Figure 2.1a) in 1953 by James D. Watson and Francis H.C. Crick (Watson and Crick, 1953). This double helix structure fulfilled all the requirements for the hereditary substance suggested as early as in 1944: the ability to store information and the ability to multiply or replicate (Griffiths et al., 1999). Genetic information is stored by the order, or sequence, of the nucleotides along each strand of the helix (Alberts et al., 2002) and part of this information is translated into protein through the process of gene expression (Griffiths et al., 1999). Replication is realized by strand separation and new synthesis is directed by the specificity of base pairing in a semi-conservative manner (Griffiths et al., 1999). The third requirement for the hereditary molecule is the possibility of change. A mutation, defined as the occasional replacement, deletion, or addition of one or more nucleotides, can result in a change of the encoded information. Mutations provide the variation evolutionary selection operates on (Griffiths et al., 1999) and are considered the driving force of evolution.

All DNA (and the information it carries) of an organism is called 'the genome' (Griffiths et al., 1999; Alberts et al., 2002). This complete set of DNA can comprise a lot of nucleotides (Stryer, 1999). For example, each human cell contains about 3 Gigabases (Gb; 3×10^9 bases) of DNA with a physical length of about 2 meters of DNA. This DNA needs to be packed into a nucleus of only about 0.006 mm in diameter (Griffiths et al., 1999). Therefore, an organized and compact way of packaging is needed and is accomplished. Recently, it was suggested that such dynamic packaging could be predictable (Richmond, 2006; Segal et al., 2006).

In eukaryotes, the DNA in the nucleus is generally distributed over a set of different molecules called chromosomes. Each chromosome consists of a single, linear DNA molecule associated with numerous proteins that fold and pack the DNA helix into a more compact structure (Figure 2.1, adopted from en.wikipedia.org/wiki/Chromosome). The structures that are cytologically known as 'chromosomes', are only visible around the time of cell division (metaphase). The complex of DNA and proteins is called chromatin (Alberts et al., 2002).

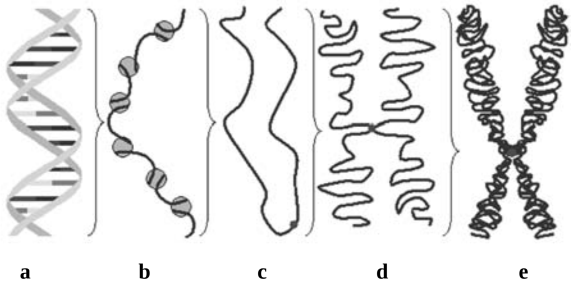


Figure 2.1: Different levels of DNA condensation.

a. Single DNA strand.
b. Chromatin strand (DNA with histones).
c. Chromatin during interphase with centromere.
d. Condensed chromatin during prophase. (Two copies of the DNA molecule are now present)
e. Chromosome during metaphase. The dot in the center represents centromere

In the first level of condensation, DNA is wrapped around several structural proteins called histones to form so-called nucleosomes. This becomes organized in the 10 nm "beads on a string" array (Figure 2.1b). The next level of chromatin organization forms the 30nm condensed chromatin fiber known as solenoid, consisting of nucleosome arrays in their most compact form (Figure 2.1c). Further packaging is supposed to consist of further coiling around a proteinaceous scaffold or matrix until the dense structure of a metaphase chromosome is attained (Figure 2.1e).

On the basis of cytological staining chromatin is subdivided into two types. One type is called heterochromatin. This is densely stained DNA that represents highly compact, gene-sparse and transcriptionally inactive regions that remain densely packaged through the cell cycle. The other type of chromatin is called euchromatin. This is less stained and represents the less condensed, gene-rich and transcriptionally active regions of the genome (Riddle and Elgin, 2006; Tremethick, 2006). The higher-order structure of the chromatin in the cell nucleus is being studied in considerable detail. For example, the three dimensional crystal structure of the tetranucleosome was elucidated recently (Schalch et al., 2005; Tremethick, 2006), supporting a zigzag model (Woodcock et al., 1984). In addition, it is generally assumed that chromatin is organized in close connection to the nuclear matrix, the filamentous protein network maintaining the overall size and shape of the nucleus (Martelli et al., 1996). Anchoring DNA elements known as matrix- or scaffold-associated (or attachment) regions (MAR, SAR or S/MAR) (Boulikas, 1993, , 1995) help to form higher-level chromatin structures such as chromatin loops. Such loops form either a transcriptionally open domain for easy access of transcription factors, or a closed domain that is less or inaccessible for transcription (Cremer et al., 2000). The loop size varies between organisms (Dillon, 2006) and between parts of the genome. It appears to depend also on the flexibility of the chromatin, which in turn is modulated by histone modification (Li et al., 2006b).

However, the above model of DNA condensation, albeit featuring in most if not all textbooks, is likely to suggest a too static sequence of events and structures. Although much has still to be learned about the functioning and dynamics of higher-order chromatin structures, more recent data generally indicate that higher chromatin structures are highly dynamic (Dillon, 2006; Luger, 2006; Tremethick, 2006). A dynamic loop domain is generally considered to represent the basic structural unit of the eukaryotic chromatin that is associated with gene expression (Heng et al., 2001).

2.2. Functional elements in eukaryotic DNA

A major challenge for nowadays bioinformatics is making sense of sequence: the full and reliable annotation of genome sequences. The ultimate goal is to describe for each and every nucleotide its role during the life span and reproduction of an organism (Fiers, 2006). It involves the correct identification and localization of distinct sequence elements such as genes, regulatory elements, repeats and many more, followed by a detailed description or prediction of the biological process in which it takes part. A major distinction is the difference between ‘coding’ and ‘non-coding’ DNA. This generally refers to the process of transcription in which RNA is made. Coding DNA can give rise to protein, but also DNA that only generates RNA is generally considered as ‘coding’. Recent data using whole genome tiling arrays indicate that much more of a genome may be transcribed (hence, be ‘coding’) than previously assumed (Bertone et al., 2004; Mockler and Ecker, 2005; Li et al., 2006a).

The operational part of any genome is the gene. For such a fundamental concept in genome function and structure, it is remarkable that the precise definition of ‘what is a gene’ is not agreed upon and a recent paper concludes that reaching consensus over the definition is likely to be virtually impossible in the near future (Pearson, 2006). In current textbooks, a gene is a region of chromosomal DNA, part of which can be transcribed into a functional RNA at the correct time and place during development (Griffiths et al., 1999). In this definition, the gene is comprised of the transcribed (and translated) region and the adjacent regulatory regions. As a modern, much looser definition of the concept of ‘gene’ was recently proposed: a gene is ‘a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions’ (Pearson, 2006).

A gene contains (or can contain) several functional regions. A eukaryotic protein-encoding gene generates its translatable mRNA by 5’capping, 3’adenylation and splicing of the primary transcript to yield the mature transcript. The most important functional regions of a gene are the protein-coding complements. Only non-protein-coding (hence RNA coding) genes do not contain these sequences, that are also known as exons. In addition to the exons, most eukaryotic genes contain non-protein coding regions called introns (Griffiths et al., 1999) that are excised from the initial transcript in a process called splicing. The functional role and the evolutionary history of introns are still being investigated and debated (see below). The phenomenon of alternative splicing, which involves an apparent choice of the splicing machinery to take out exon sequences as well, helps the cell to generate more (protein) diversity from a single gene. The frequency of occurrence, so the importance of alternative splicing in plant genomes, is on the rise (Kazan, 2003). In addition to introns and exons, there are stretches of DNA that are transcribed and retained in the mature transcript, but that are not translated. These sequences are called the 5’ and 3’ untranslated regions (UTRs). Many regulatory sequences are found in the UTRs, but in order to generate a functional transcript at the required time and place, a gene has regulatory regions at the 5’end and terminator signals at the 3’end. The regulatory region, or ‘promoter’, is DNA that can receive and respond to signals that can trigger the binding of regulatory proteins to initiate or regulate transcription (Griffiths et al., 1999). Transcription factor binding sites and enhancers are among the elements that can be present in the regulatory region of a gene, although enhancer sequences can also be located anywhere in the genome: close to the gene, far upstream or downstream of the coding sequence, or even on another chromosome

(Arnosti and Kulkarni, 2005). Between genes are various stretches of so-called intergenic or ‘spacer’ DNA, mostly with yet unknown functions. The distinction between regulatory (promoter) sequences and supposedly ‘neutral’ intergenic DNA is not easy to make. It may dynamically change depending on cell, tissue and/or environment. The length and nature of this intergenic DNA vary with the genome. Different types of repetitive DNA, such as microsatellites, are over-represented in intergenic regions. The function of such repetitive DNA is generally assumed to be largely structural (Griffiths et al., 1999).

2.3. Measurement of gene expression

Gene expression is the formation of messenger RNA (mRNA) molecules that are translated into the amino acid sequences of proteins that perform most of the critical functions of cells (Alberts et al., 2002). The formation of mRNA is the manifestation of many of the regulatory circuits that exist in cells. Understanding and interfering with the regulation of gene expression is a central research theme in molecular biology and, on a much larger scale, in current-day genomics. The study of gene expression involves detecting and analyzing the types and amounts of mRNA produced by a cell. The proper and coordinated expression of a large number of genes is a critical component of normal growth and development as well as the maintenance of health upon pathogen challenge. The patterns of gene expression tell how cells and organisms respond to environmental stimuli. Altered gene expression patterns, notably when compared with an appropriate control in either developmental stages or environment, makes it possible to elucidate the chain of events that is responsible for many diseases or other undesired or desired outcomes. Comparing the normal status of gene expression between and among different organisms gives insight into the fundamentals, similarities and differences in the regulatory systems of these organisms. It may reveal the impact of evolutionary selection on the regulation of gene expression.

Traditionally, gene expression was studied by isolating, cloning and analyzing individual or small groups of RNA molecules. Northern or RNA blot analysis allowe detecting a specific RNA molecule in a mixture of RNAs fractionated on a gel. Reverse transcription-polymerase chain reaction (RT-PCR), either qualitative or semi-quantitative, is a valuable alternative for the study of gene expression in relatively small samples. This technique is sensitive enough to enable quantitation of RNA from a single cell (www.ambion.com/techlib/basics/rtpcr/index.html). Both Northern blot analysis and RT-PCR are still essential for the confirmation of high-throughput expression data. Technological developments have gone fast and nowadays there are several technologies available for detecting and studying gene expression on a genome-wide scale (Pollock, 2002). Two types of approaches are distinguished: hybridization-based and sequence-based technologies. Hybridization-based technologies are extensions of the Northern blot approach. They are based on the ability of a given mRNA molecule to bind specifically, that is, hybridize, to the DNA from which it originated. The most developed technology is the microarray or DNA chip technology. With a glass slide (or silicon chips or nylon membrane) containing many immobilized DNA samples, the expression levels of thousands of genes within a cell or organism can be determined by measuring the amount of mRNA bound to each site on the array. The DNA samples are printed, spotted, or synthesized directly onto the support material. The samples themselves can be genomic DNA, cDNA, or oligonucleotides of various lengths (www.ncbi.nlm.nih.gov/About/primer/microarrays.html). DNA arrays permit the global analysis of gene expression in complex biologic systems in a high-throughput fashion at nowadays a very reasonable cost (Todd and Wong, 2002). However, the technology is still being refined (Pollock, 2002). The statistical interpretation of microarray data is developing into a field on its own. Advanced statistical techniques are necessary to generate reliable expression data from arrays and to compare data from different arrays. Data from different laboratories, even when using the same type of microarray, are not as comparable as one would wish, due to a multitude of largely experimental parameters. More work needs to be done to increase the reliability and sensitivity of the software available for interpreting the hybridization data (Pollock, 2002; Saluz et al., 2002).

In addition to hybridization technologies, several sequence-based technologies allow studying gene expression. The current sequence-based technologies tend to be more specialized and more expensive than microarray technologies. As a result, they are less widespread. The three sequence-based technologies that will be described here are: expressed sequence tag (EST) profiling, Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature Sequencing (MPSS).

Expressed Sequence Tags (ESTs) are small pieces of DNA sequence of usually 100 to 500 nucleotides that are read from either one or both ends of cloned mRNA molecules (Marra et al., 1998). With the appropriate laboratory infrastructure for mRNA isolation, cloning and sequening, ESTs can be generated in large amounts relatively inexpensive (Marra et al., 1998). As they are unedited single-pass sequencing reads, they are prone to error and attain at best a 97% accuracy rate. Redundancy of sequences is a general property of current EST data sets (Marra et al., 1998; Khan et al., 1999) and this is used for EST profiling. The relative occurrence of ESTs in a given sample, or in combined databases, indicates at what level the gene expresses in the tissue investigated. The occurrence of ESTs is somewhat biased in proportion to the abundance of the mRNAs in the tissues from which the library was prepared (Marra et al., 1998). Genes expressed at very low levels are not likely to be found within EST data sets, while abundantly expressed genes tend to be over-represented

(Khan et al., 1999). There are three other important uses of the ESTs. These are gene identification, gene-based physical-map construction, and the computer-assisted large-scale characterization of genomic sequences (www.ncbi.nlm.nih.gov/About/primer/est.html).

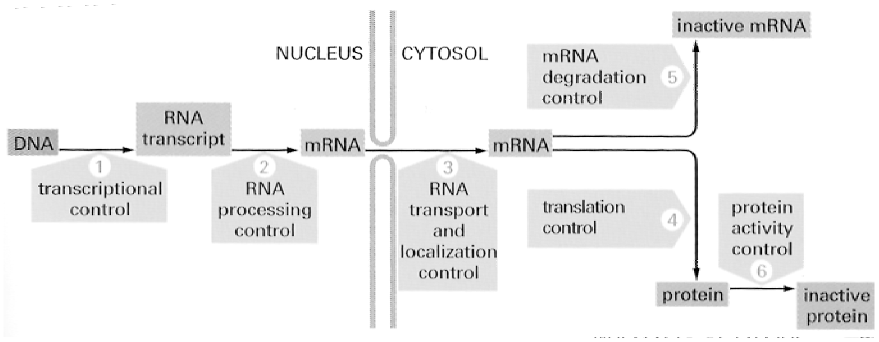
Serial Analysis of Gene Expression (SAGE) is based on the isolation of sequence tags from individual transcripts at the 3' end with the help of restriction enzymes and subsequent concatenation of those tags into long concatemer molecules. Sequencing of concatemer clones reveals individual tags and allows quantification and identification of cellular transcripts (Velculescu et al., 1995; Velculescu et al., 1997). SAGE allows a relatively rapid and detailed characterization of gene expression patterns and could be used to characterize the entire set of genes expressed from a eukaryotic genome (Velculescu, 1999). SAGE provides quantitative gene expression data, even for uncharacterized genes, without the prerequisite of a hybridization probe for each transcript (Velculescu et al., 1997; Pollock, 2002). It estimates the level of expression for a defined population of cells by counting tags without the need for advanced statistical normalization methods (Velculescu, 1999; Pollock, 2002). SAGE tags are normally 9-11 bp long (Velculescu et al., 1997). Therefore, the main disadvantage is that SAGE tags cannot always be unambiguously mapped to a unique gene, especially in larger genomes.

Massively Parallel Signature Sequencing (MPSS) is a technology similar in approach to SAGE. MPSS starts by first attaching unique fluorescent tag sequences (synthesized 32 bp long) to each cDNA molecule in a complex mixture, next amplifying the tagged library, then select with 5 μm microbeads that carry the complement (anti-tag) to each unique tag. In this way, a mixture of one million mRNA molecules can be converted into a library of about as many microbeads, each carrying about 100,000 copies of the templates (Brenner et al., 2000b). Next, a 17-base sequence (the 'signature') for each mRNA at a restriction site upstream of the poly-A tail (first *DpnII* site) is generated. The total number of signatures for a given mRNA is counted as the level of expression of any given gene (Reinartz et al., 2002). The mRNA abundance is expressed as transcripts per million (TPM) (Reinartz et al., 2002). As in SAGE, the tag-based expression allows for easy quantification of gene expression (Brenner et al., 2000a; Tetko et al., 2006) and for easy comparison among samples. The MPSS tags generated range from 17 to 20 bp. This facilitates the unique identification of transcripts. The transcripts measured are not pre-selected (except for the need of the presence of a *DpnII* site) and especially it allows detection of lowly expressed genes (Brenner et al., 2000a; Pollock, 2002; Reinartz et al., 2002; Coughlan et al., 2004; Tetko et al., 2006). A main disadvantage of the MPSS technology is that it requires advanced proprietory equipment and is (very) expensive.

2.4. Regulation of gene expression

In all cells, the expression of genes is regulated at several levels. Cells change the genes they employ in response to changes in their external or internal environment and their needs (Alberts et al., 2002). A cell does not produce all possible proteins at their full amount throughout its life cycle. In contrast, a cell adjusts the rate and amount of transcription and translation of different genes according to its need at a certain time and a certain place (Alberts et al., 2002). To accomplish such fine-tuning, gene expression is regulated at many different levels in the pathway from DNA to RNA to protein (Figure 2.2, adopted from (Alberts et al., 2002)). The past decade has seen a clear shift in focus from the protein-coding part of the genome to the analysis of non-protein coding part of the genome, where most of the regulatory machinery, if not all, is believed to reside in (Califano, 2001). The first level of regulation is transcriptional regulation. This determines the frequency of transcriptional initiation and as a result when and how often a given gene is transcribed. For most genes, the initiation of transcription is the most important point of control. A second level is the regulation of RNA processing. This determines how the primary RNA transcript is spliced or otherwise processed. The third level is the regulation of RNA transport and localization, that controls the access to or efficiency of transport channels, selects which mature mRNAs are exported from the nucleus to the cytosol and determines where in the cytosol these mRNAs are localized. Next is translational regulation. This selects which mRNAs in the cytoplasm are translated by ribosomes and speeds up or slows down protein synthesis. The rate of protein synthesis also depends on the availability of the various proteins and amino acids. The fifth level of control is the regulation of mRNA degradation. This modulates the speed by which transcripts are degraded. A final level of regulation is control over protein activity that selectively activates, inactivates, degrades, or compartmentalizes protein molecules after they have been made by a suite of post-translational modifications (Alberts et al., 2002). The various levels of regulation imply that there is no simple linear relationship between the various molecules involved, for example transcription levels cannot be equaled without precaution to protein abundance.

Figure 2.2: Six levels at which eukaryotic gene expression can be regulated (adopted from (Alberts et al., 2002)).



In addition to the levels of regulation of gene expression outlined above, gene expression is also regulated at the higher level of chromatin structure and position in the nucleus. The structure of chromatin is modified by a variety of means such as DNA methylation, histone modification, such as (de)methylation, (de)acetylation and more. These so-called epigenetic modifications play important roles in controlling transcription without changing the sequence of the DNA. In addition, topological dynamics such as (un)folding and chromatin remodeling play important roles in the regulation of gene expression (Dillon and Sabbattini, 2000; Dillon, 2006; Luger, 2006). Local and global chromatin domains are considered to be important regulators of transcription, replication, DNA repair and recombination (Dillon and Sabbattini, 2000; Chodaparambil et al., 2006; Dillon, 2006; Tremethick, 2006). The open or closed status of a chromatin domain determines the access of transcription factors and RNA polymerase to the promoter.

2.5. Patterns of gene expression

The expression of a gene also depends on the particular position of genes along the chromosome. There is a non-random distribution of genes in the genome (Caron et al., 2001; Lercher et al., 2002; Sankoff and Haque, 2006). Gene expression therefore is also not randomly distributed. Regions of increased gene expressions (RIDGEs) along chromosomes were detected in the human transcriptome (Caron et al., 2001). Further studies revealed chromosomal domains of highly and lowly expressed genes (Versteeg et al., 2003), indicating that RIDGEs (highly expressed domains) and anti-RIDGEs (lowly expressed domains) are likely to represent higher-order structures in the genome (Caron et al., 2001; Versteeg et al., 2003). The clustering of highly expressed genes into domains was due to the clustering of housekeeping genes, genes that are always expressed at a relatively constant and high level (Lercher et al., 2002). It was suggested that it might give a selective advantage to assemble housekeeping genes to an open chromatin conformation across all cells for easy access to the transcription machinery (Lercher et al., 2002).

Genes located in the same chromatin domain are supposed to exhibit coordinated regulation (Laemmli et al., 1992; Bode et al., 1996). Large scale chromosomal coexpression domains were found in many organisms ranging from yeast to mammals (Cohen et al., 2000; Roy et al., 2002; Spellman and Rubin, 2002; Lercher et al., 2003; Williams and Bowles, 2004; Ren et al., 2005) and often coincide with chromatin loops isolated by insulators that are anchored to nuclear attachment points (Burgess-Beusse et al., 2002; Labrador and Corces, 2002; Dillon, 2006). The creation of an artificially domain with two genes resulted in highly correlated expression of the two genes in tobacco (Mlynarova et al., 2002).

In the scientific literature, many different terms are used to describe genes based on different patterns of gene expression, such as the distinction between housekeeping genes and tissue-specific genes. Housekeeping (HK) genes are genes that are expressed in all cell types (Alberts et al., 2002) generally at a relatively constant and high level. They replicate early in S phase (Holmguist, 1987; Alberts et al., 2002). The products of HK genes are typically needed for maintenance of the cell. Examples include actin, GAPDH and ubiquitin. It is generally assumed that the expression of HK genes is unaffected by different experimental conditions. Therefore, HK genes are usually used as internal standard for the calibration of expressions of other genes (Thellin et al., 1999). Genes expressed in only a few cell types or in a certain tissue are often referred to as tissue-specific (TS) genes. TS genes generally replicate early in the cells in which they are expressed and later in cells where they are not (Alberts et al., 2002). Compring the group of TS genes with HK genes may give insight into the regulatory mechanisms that define and distinguish HK genes. Other terms used to describe patterns of gene expression are ‘constitutive’, ‘facultative’ or ‘inducible’. Such terms are not mutually exclusive. A constitutive gene is a gene that is transcribed continuously, whereas a facultative gene is only transcribed when needed (en.wikipedia.org/wiki/Gene_expression). An inducible gene is a gene whose expression is either responsive to environmental change or dependent on the phase of the cell cycle (en.wikipedia.org/wiki/Gene_expression).

2.6. Genomics of gene structure and gene expression

In August 1999, the GenBank repository of nucleic acid sequences contained about 3.4 billion bases (Benson et al., 2000). August 2006, there are well over 130 billion bases in GenBank (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide). Around 400 organisms have been fully sequenced, and about 10% concerns eukaryotes (genomesonline.org). Proper annotation is a major challenge (Fiers, 2006). Thanks to the high throughput expression platforms described above, databases containing expression data are growing with an even higher speed. The largest public repository with expression data is the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/projects/geo/) at the National Center for Biotechnology Information (NCBI) (Barrett et al., 2005). GEO currently holds over 30,000 submissions representing approximately half a billion individual molecular abundance measurements, spanning over 100 organisms (Barrett et al., 2005). The MPSS repository for plants contains 297,313 distinct 17 base-long signatures (tags) from the whole transcriptome of Arabodipsis genome and 274,096 from that of rice genome (Nakano et al., 2006) (mpss.udel.edu), which give expression data for about 72% of the well-annotated Arabidopsis genes and 41% of rice genes. The majority of the MPSS signatures are mapped to the intergenic regions representing previously unannotated or non-coding transcripts, particularly small RNAs (Nakano et al., 2006).

The further integration of sequence data and their annotations with expression data will provide valuable and possibly new insights into biological processes, as well as many new computational methodologies for the analysis of biological data (Califano, 2001). Expression data should assist annotation and will result in gene discovery and gene networking (Gaasterland and Oprea, 2001). Expression data show whether a predicted gene has biological relevance and can help to decide whether the predicted gene structure is correct. They outline under what conditions a gene is expressed and in association with what other genes (Gaasterland and Oprea, 2001). Expression studies have often focused on clustering genes based on similar expression profiles (Luscombe et al., 2001). Genes that are expressed in given conditions are supposed to share regulatory mechanisms (Gaasterland and Oprea, 2001; Luscombe et al., 2001). They may be functionally related either in the same pathway or carry out a function which is conserved during evolution (Lercher et al., 2002; Roy et al., 2002; Lee et al., 2004; Williams and Bowles, 2004; Zhan et al., 2006). Chromosomal coexpression domains consist of neighboring genes that have similar and synchronized spatial or temporal expression patterns in different cell types and/or under different developmental, environmental or experimental conditions (www.nature.com/nrg/journal/v5/n4/glossary/nrg1319_glossary.html). A reason that neighboring genes are coexpressed could be that they are in the same chromatin loop/domains, so that their expression patterns are regulated by changes at chromatin levels. Alternative explanations could be gene duplication (Lercher et al., 2003), shared promoter regions (West et al., 1984; Nakao et al., 1986; Osley et al., 1986; Kraakman et al., 1989; Kruglyak and Tang, 2000; Hurst et al., 2002) or genes that are located so close to each other that there is read-through transcription (Roy et al., 2002; Lercher et al., 2003; Semon and Duret, 2006). Chapter 3 and 4 in this thesis come to the conclusion that the higher-order chromatin structure of genes is the main factor that accomplishes significant coexpression of neighboring genes in Arabidopsis and rice.

2.7. Characteristics and functions of introns

With more and more genomes completed, the field of comparative genomics becomes the mainstream of research. The goal of comparative genomics is to unravel the nature of the relationships between gene structure and gene expression by focusing on specific differences and similarities in gene structure and organization across multiple genomes (Califano, 2001). A major topic for comparative analysis is the occurrence, position, number and length of introns in eukaryotic genes.

Introns are thought to have made crucial contributions to the evolution of complexity in multicellular organisms (Koonin, 2006). The origin of introns in eukaryotic genes has been intensively debated since their discovery in 1977 (Berget et al., 1977). Two major theories disagree on the origin of the introns: they are known as the introns early and introns late theories. The introns-early theory supposes that introns are ancient and have been lost in prokaryotes (Darnell, 1978; Doolittle, 1978; Gilbert, 1978; de Roos, 2005; Belshaw and Bensasson, 2006). The lack of introns in present-day prokaryotes is seen as the result of a ‘streamlining’ process due to selection for fast DNA replication (Gilbert, 1987; de Souza, 2003; Belshaw and Bensasson, 2006). The introns-late theory states that the introns were inserted into the eukaryote genes later in evolution (Cavalier-Smith, 1991; Palmer and Logsdon, 1991; Cho and Doolittle, 1997; Logsdon, 1998) and were related to the evolution of multi-modular proteins (de Roos, 2005). There is no conclusive evidence for either theory (de Roos, 2005). Also no consensus has been reached as to whether introns are under positive, negative or neutral selection (Roy and Gilbert, 2006), possibly because different scenarios apply to different introns.

Introns can harbor regulatory elements that probably function at the RNA level (Shabalina and Spiridonov, 2004). Especially proximal introns are implied in gene regulation (Majewski and Ott, 2002; Comeron, 2004; Seoighe et al., 2005; Kalari et al., 2006). There may be selection on longer introns to maintain pre-mRNA secondary structure (Kirby et al., 1995; Haddrill et al., 2005) and introns may contribute to the formation of an RNA secondary structure involved in gene expression (Liebhaber et al., 1992; Carlini et al., 2001). Introns may function as a “time-scheduler” for mRNA production, the time taken by transcription through a long intron can specify a functionally significant delay in the appearance of processed mRNA and protein products relative to the regulatory signals that activated the transcription of the gene (Shermoen and O'Farrell, 1991; Burnette et al., 2005). Longer introns are thought to be advantageous for recursive splicing, a mechanism that functions first at 3' splice sites and then regenerate 5' splice sites after ligation to an upstream exon (Hatton et al., 1998; Burnette et al., 2005). Another well-established biological role for introns is their involvement in nucleosome formation and chromatin organization (Shabalina and Spiridonov, 2004). Introns have higher potential for nucleosome formation than exons or Alu repeats (Levitsky et al., 2001; Shabalina and Spiridonov, 2004). In dipteran fly species (Beckingham and Rubacha, 1984), different chromatin states were observed for intron-free and intron-containing rRNA genes. S/MAR elements that are thought to anchor chromatin loops to the nuclear matrix (Glazko et al., 2003), are found to be abundant in introns (Rudd et al., 2004; Shabalina and Spiridonov, 2004; Tetko et al., 2006), supporting the involvement of introns in the formation of chromatin structures. The presence of introns can greatly increase proteome complexity by increasing the rate of recombination between exons via exon shuffling (Gilbert, 1978; Patthy, 1999; de Souza, 2003) and/or alternative splicing (de Souza, 2003). Intron retention, a form of alternative splicing in which introns can be retained, recently found to involve alterations in mRNA transport (Li et al., 2006c). Introns can also increase fitness by increasing intragenic recombination (Comeron and Kreitman, 2000; Roy and Gilbert, 2006) and boost transcript fidelity through nonsense-mediated decay (NMD) (Lynch and Kewalramani, 2003; Roy and Gilbert, 2006).

In view of the manifold potential functions of introns, it is remarkable that genomes differ considerably in the number and length of introns. Unicellular eukaryotes have few introns. Arabidopsis has less introns than exons, whereas nematodes have more introns than exons and in mammalian genomes the contribution of introns rises to comprise almost 95% of the total length of the primary transcript (Shabalina and Spiridonov, 2004). Mammalian introns are often up to tens of thousands of nucleotides long. The longest human intron known is 480 kb long (Maniatis and Reed, 2002). The reasons for such differences in number and size of introns are not clear. It is also not clear why mammalian systems would invest so much in sequences that are essentially non-coding. Transcription costs two ATP molecules per nucleotide and about 20 nucleotides can be transcribed each second (Castillo-Davis et al., 2002). Transcription of such large introns would seem an inefficient use of metabolic energy, especially when such a gene is transcribed often in many cell types. This consideration generated the transcriptional efficiency theory proposing that selection would favor shorter introns in highly and/or broadly expressed genes to reduce the burden of energy (Castillo-Davis et al., 2002). In almost all studies of animal genes it was indeed found that higher and/or broadly expressed genes have less and shorter introns, shorter exons, shorter UTRs and shorter intergenic regions when compared to lower expressed and/or tissue-specific genes (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Vinogradov, 2004, , 2005). This is taken as evidence that every aspect of the structure of genes is subject to selection for transcriptional efficieny. Other theories to explain the distribution of introns in relationship to expression chraracteristics are the genomic design theory (Vinogradov, 2004) and the regional mutational bias theory (Urrutia and Hurst, 2003). The genomic design theory hypothesizes that highly expressed genes are positioned into open chromatin domain made up by less intronic and intergenic noncoding DNA (Vinogradov, 2004). The regional mutational bias theory supposes that highly expressed genes are located in regions prone to deletions, so that sequences in these regions tend to be compact (Urrutia and Hurst, 2003).

Chapter 5 of this thesis investigates the relationships between gene structure and gene expression in Arabidopsis and rice, showing that in plants the relationship is apparently different than in animal systems. Chapter 6 extends the analyses to a broad-scale comparative analysis of genes in both plants and animal genomes and presents the apparent and obvious differences between plant and animal genomes in its evolutionary context.

2.8. Future outlooks

Organisms differ a lot in the numbers associated with genes and genomes. The average number and size of exons and notably introns are strikingly different. The amount of non-coding and repetitive DNA varies tremendously between species, as does the number of genes. Total genome size of species ranges several orders of magnitude, notably among the flowering plants. Also the number of pairs of chromosomes is hugely different between species, whereas chromosomes within a species can differ considerably in length and in the number of genes that they carry (Griffiths et al., 1999). Today, it is largely unclear whether any of these quantitative differences in genome structure and configuration has any functional role or biological significance. A better (re)definition of many of the parameters involved, such as organismal complexity, gene, coding part, noncoding part, domain etc. is becoming more and more important for future comparative genomics research into such quantitative differences. When the confusion about terms is solved, consensus about proper ways of analysis should be reached. This will help to avoid future confusion in the conclusions drawn from comparative analyses, within and among research disciplines.

Chapter 3

Local coexpression domains of two-to-four genes in the genome of Arabidopsis

Xin-Ying Ren[1, 2], Mark W.E.J. Fiers[1], Willem J. Stiekema[2, 3]& Jan-Peter Nap[1, 3]

[1]Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands
[2]Laboratory of Bioinformatics, Plant Sciences Group, Wageningen University and Research Center, 6703 HA Wageningen, the Netherlands
[3]Centre for BioSystems Genomics, 6700 AA Wageningen, The Netherlands

Published in *Plant Physiology* 138:923-934 (2005)

Local coexpression domains of two-to-four genes in the genome of Arabidopsis

Abstract

Expression of genes in eukaryotic genomes is known to cluster, but cluster size is generally loosely defined and highly variable. We have here taken a very strict definition of ‘cluster’ as sets of physically adjacent genes that are highly coexpressed and form so-called local coexpression domains. The *Arabidopsis thaliana* genome was analyzed for the presence of such local coexpression domains to elucidate its functional characteristics. We used expression data sets that cover different experimental conditions, organs, tissues and cells from the Massively Parallel Signature Sequencing (MPSS) repository and microarray data (Affymetrix) from a detailed root analysis. With these expression data, we identified 689 (MPSS) and 1481 (microarray) local coexpression domains consisting of 2 to 4 genes with a pair-wise Pearson’s correlation coefficient (R) larger than 0.7. This number is about 1 to 5-fold higher than the numbers expected by chance. A small (5-10%) yet significant fraction of genes in the Arabidopsis genome is therefore organized into local coexpression domains. These local coexpression domains were distributed over the genome. Genes in such local domains were for the major part not categorized in the same functional category (GOslim). Neither tandemly duplicated genes, nor shared promoter sequence, or gene distance explained the occurrence of coexpression of genes in such chromosomal domains. This indicates that other parameters in genes or gene positions are important to establish coexpression in local domains of Arabidopsis chromosomes.

Introduction

The combination of DNA sequence and expression data has revealed the existence of chromosomal domains of similarly expressed genes in several genomes, such as in yeast (Cohen et al., 2000), fly (Spellman and Rubin, 2002) worm (Roy et al., 2002; Lercher et al., 2003), human (Caron et al., 2001; Lercher et al., 2002; Versteeg et al., 2003), and more recently in the genome of the plant *Arabidopsis thaliana* (Birnbaum et al., 2003; Williams and Bowles, 2004). These analyses have focused on coexpression (Cohen et al., 2000; Spellman and Rubin, 2002; Lercher et al., 2003; Williams and Bowles, 2004), high expression (Caron et al., 2001), and so-called localized expression domains (Birnbaum et al., 2003), defined as spatial and/or temporal chromosomal domains of coordinated induction and repression in gene expression. Chromosomal domains (or clusters or regions) of similarly expressed (or coexpressed or co-regulated or correlated) genes have been identified using sliding windows of either a given sequence length (number of nucleotides) (Lercher et al., 2002) or of a given number of genes (Spellman and Rubin, 2002; Lercher et al., 2003; Williams and Bowles, 2004). Major experimental differences exist in the size of the window used for analysis and therefore the fraction of the genome evaluated as chromosomal domain. To determine the similarity between different expression profiles, the Pearson correlation coefficient (R) was used and the average of all pair-wise Rs over the expression values across experiments or tissues was evaluated over all windows and chromosomes (Cohen et al., 2000; Spellman and Rubin, 2002; Lercher et al., 2003).

In such genome-wide analyses, four different types of gene organization may account for high coexpression without giving evidence for the presence of chromosomal domains. These four types are (i) overlapping genes (Cohen et al., 2000), (ii) tandemly duplicated genes, (iii) homologous genes (Spellman and Rubin, 2002; Lercher et al., 2003), or (iv) genes in the same operon (Roy et al., 2002; Lercher et al., 2003). Generally, these four gene configurations have been analyzed separately for their contribution to coexpression. The remaining genes, if coexpressed, might be an indication of the existence of chromosomal domains. Housekeeping genes (Lercher et al., 2002; Roy et al., 2002; Lercher et al., 2003), genes with similar functions in different biological processes (Cohen et al., 2000; Spellman and Rubin, 2002), genes involved in the same metabolic pathway (Birnbaum et al., 2003), or genes involved in the same biological process (Williams and Bowles, 2004), have all been identified in these chromosomal domains. Therefore, there does not appear to be a clear functional classification of genes present in such chromosomal domains.

The molecular mechanisms responsible for coordinated expression of neighboring genes are not well understood (Hurst et al., 2004). Coexpressed adjacent genes in yeast could not be explained solely by upstream activating sequences and are not due to divergently transcribed promoter regions, although the extent of physical vicinity seems to be important

(Cohen et al., 2000). In worm, coexpression of genes could not be attributed to unrecognized operons or read-through transcription (Roy et al., 2002). Neither gene duplication nor common functionality was identified as the main cause for coexpression of neighboring genes in the Arabidopsis genome (Williams and Bowles, 2004). It is generally assumed that the coordinated expression of genes in chromosomal domains represents gene regulation at the level of specialized chromatin and chromosome structure. In Arabidopsis, limited chromosomal clustering of co-regulated genes associated genome organization with gene regulation (Birnbaum et al., 2003). Analyses of the phenomenon in transgenic plants also indicated the importance of chromosomal context for proper gene expression (Mlynarova et al., 1994; Mlynarova et al., 1995).

We here present the identification and analysis of local coexpression domains in the Arabidopsis genome. Local coexpression domains are here defined as chromosomal regions where physically adjacent genes have high correlated expression across all experiments. This definition focuses on the behavior of neighboring genes. Using the MIPS Arabidopsis genome annotation (Schoof et al., 2002) and two types of whole genome expression data, Massively Parallel Signature Sequencing (MPSS) (Meyers et al., 2004) and an Affymetrix microarray (MA) (Birnbaum et al., 2003), we have analyzed the coexpression of neighboring genes to identify local coexpression domains. Our results contrast with the genome-wide identification of more global coexpression domains, consisting of clusters up to 20 genes with a median cluster size of 100 kb. In such domains, coexpression was defined as a significant deviation from the averaged correlation coefficient (Williams and Bowles, 2004). This difference underlines the importance of distinguishing the size dimension of the chromosomal domains considered.

Results

Chromosomal coexpression maps reveal local coexpression domains

The combination of the MIPS annotation of the Arabidopsis genome with the available MPSS and MA data resulted in a collection of 16,144 gene pairs with MPSS expression values and 18,443 pairs with MA expression values that could be analyzed. A more detailed description of the data sets generated is given in Table 3.1 and in Materials and Methods.

For visualization purposes, the overall whole-genome coexpression data were plotted in chromosomal coexpression maps as introduced by Cohen et al. (2000) for each chromosome of Arabidopsis. Figure 3.1 shows the chromosomal coexpression map of an area of 80 genes on chromosome 1 for which both MPSS (Figure 3.1a) and MA (Figure 3.1b) data were available. Genes with correlated expression are indicated in green. Genes that have correlated expression and are physically close together form green regions along the diagonal of the chromosomal coexpression map. Examples of such regions are indicated with a blue box (Figure 3.1a, 3.1b). Comparison of the same genomic regions in the MPSS (Figure 3.1a) and MA (Figure 3.1b) data show that regions can have different coexpression patterns in different data sets. Subsets of neighboring genes having high coexpression in the MPSS data set (Figure 3.1a, blue box) showed low coexpression in the MA data set (Figure 3.1b, yellow box), while subsets of neighboring genes in MA having high coexpression (Figure 3.1b, blue boxes) have low coexpression in the MPSS data set (Figure 3.1a, yellow boxes). Such differences in coexpression are likely to reflect the biological differences between the MPSS and MA data sets, although it can not be fully excluded that technical differences between whole genome expression profiling with MPSS and Affymetrix MA have also contributed in part to the differences observed. The MPSS data cover plant tissues and organs, while the MA data refer to defined root cells. The averaged expression over the biological material sampled in a data set may influence coexpression patterns of neighboring genes.

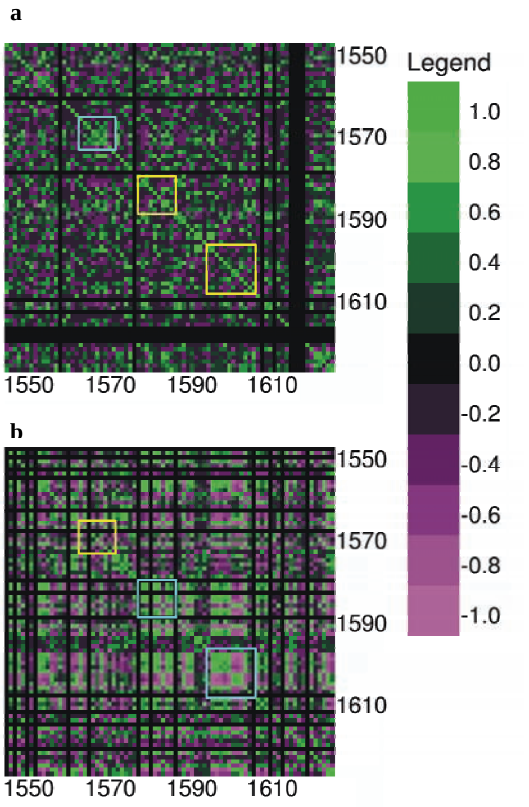


Figure 3.1 Chromosomal coexpression map of the Arabidopsis genome. The expression of each gene is correlated with all other genes on the same chromosome using a color coded representation of R. Green is positive correlation (R>0), magenta is anti-correlation (R<0) and black shows no correlation (R=0), no expression or missing data. **a,** Coexpression map of a small part of chromosome 1 using MPSS expression data, showing the 80 genes from At1g16240 (rank ID 1550) to At1g17090 (rank ID 1630). **b,** coexpression map of the same 80 genes on chromosome 1 using microarray (MA) expression data. The blue boxes in **a** and **b** indicate regions of blocks of coexpressed adjacent genes. The yellow boxes in **a** and **b** indicate the equivalent regions in the other data set.

Table 3.1 Description of expression data used for whole-genome analysis

	MPSS	MA
Genes with expression		
Total	20041	21940
Overlapping	39	34
Without expressed neighbor(s)	851	651
Represented in pairs	19151	21255
Adjacent pairs		
Total	16144	18443
Tandemly duplicated pairs (td)	1928 (11.9%)[a]	2278 (12.4%)[a]
Coexpressed	905 (5.6%)[b]	1800 (9.8%)[b]
Total excluding td	14216	16165
Coexpressed excluding td	689 (4.8%)[c]	1481 (9.2%)[c]
Coexpressed adjacent pairs		
Total	905	1800
Tandemly duplicated pairs	216 (23.9%)[d]	319 (17.7%)[d]
Tandemly duplicated pairs		
Total	1928	2278
Coexpressed	216 (11.2%)[e]	319 (14.0%)[e]

[a] The percentage of tandem duplicated pairs relative to the total number of adjacent pairs.

[b] The percentage of coexpressed adjacent pairs relative to the total number of adjacent pairs.

[c] The percentage of coexpressed adjacent pairs excluding td relative to the total number of adjacent pairs excluding tandemly duplicated pairs.

[d] The percentage of coexpressed tandemly duplicated pairs relative to the total number of coexpressed adjacent pairs.

[e] The percentage of coexpressed tandemly duplicated pairs relative to the total number of tandem duplicated pairs.

The number of local coexpression domains in the Arabidopsis genome and the number of genes involved were calculated. Two genes were considered to be adjacent, so present in a local coexpression domain, if their IDs (see Methods) were consecutive with a difference of one and their pair-wise correlation coefficient R exceeded 0.7. Notably tandemly duplicated genes are known to influence coexpression statistics (Zhu, 2003; Hurst et al., 2004). Therefore, the occurrence of tandemly duplicated pairs was determined with pair-wise protein BLAST using a cut-off of E< 2 x 10^{-1} (Lercher et al., 2003; Fukuoka et al., 2004; Williams and Bowles, 2004). This criterion has a false error rate of about 10% (Lercher et al., 2002; Williams and Bowles, 2004). In both the MPSS and MA expression data sets, only about ~12% of all adjacent pairs (1928 for MPSS and 2278 for MA) are tandemly duplicated (Table 3.1), of which only 11 – 14% are coexpressed (216 for MPSS and 319 for MA). This implies that in either expression data set only ~20% of the coexpressed pairs consist of tandemly duplicated genes. Only a minority of 11-14% of all tandemly duplicated gene pairs in the Arabidopsis genome is coexpressed (with R>0.7), reflecting gene divergence after duplication (Williams and Bowles, 2004). As about 5 – 9% from all adjacent pairs excluding the tandemly duplicated pairs (689 for MPSS and 1481 for MA) are coexpressed, a tandemly duplicated pair is about two-fold (that is, 11-14% relative to 5-9%) more likely to be coexpressed than a non-tandemly duplicated adjacent pair. Further analyses of the sub-population of tandemly duplicated gene pairs do not indicate that a particular transcriptional orientation of the tandemly duplicated genes has a significantly higher inclination to be coexpressed (data not shown). In subsequent analyses, the sub-populations were evaluated with and without tandemly duplicated genes. The results are summarized in Table 3.2.

Table 3.2 Number of local coexpression domains ranging 2 to 4 genes

	Arabidopsis genome		Random genome (100x)	
	Total[a]	Coexpressed[b]	Average[c]	P-value[d]
Pairs				
MPSS+td[e]	16144	905 (5.60%)	676 ± 25	1.52×10^{-18}
MPSS-td[f]	14216	689 (4.85%)	588 ± 24	2.88×10^{-6}
MA+td[g]	18443	1800 (9.76%)	1352 ± 33	1.80×10^{-34}
MA-td[h]	16165	1481 (9.16%)	1211 ± 31	5.96×10^{-16}
Triplets				
MPSS+td	13142	52 (0.40%)	22.6 ± 4.7	7.95×10^{-8}
MPSS-td	12392	42 (0.34%)	19.6 ± 4.1	6.33×10^{-6}
MA+td	15634	113 (0.72%)	70.9 ± 8.0	9.70×10^{-7}
MA-td	14493	107 (0.74%)	70.9 ± 8.8	1.39×10^{-5}
Quadruplets				
MPSS+td	10718	5 (0.05%)	0.76 ± 0.89	9.88×10^{-4}
MPSS-td	10403	5 (0.05%)	0.72 ± 0.92	7.84×10^{-4}
MA+td	13282	8 (0.06%)	4.39 ± 2.38	5.81×10^{-2}†
MA-td	12866	7 (0.05%)	4.50 ± 1.85	8.24×10^{-2}†

[a] Total number of pairs, triplets, quadruplets in each data set.
[b] Coexpressed pairs, triplets, quadruplets in each data set. Percentages in the brackets are coexpressed relative to the total.
[c] Average and standard deviation from 100 times randomizations.
[d] P-value according to the cumulative binomial distribution (Cohen et al., 2000) for obtaining such result by chance. P < 0.01 is considered significant; †, not significant
[e] MPSS data set including tandem duplicates.
[f] MPSS data set excluding tandem duplicates.
[g] MA data set including tandem duplicates.
[h] MA data set excluding tandem duplicates

Depending on the expression data set considered, 5 – 9 % of all non-duplicated gene pairs consist of coexpressed neighboring pairs (689 for MPSS and 1481 for MA). These pairs tend to be spread throughout the genome (Figure 3.2; MPSS data). The MA data set gave similar results (data not shown). Only 58 coexpressed pairs were common between the MPSS and MA sets, out of 11,144 total common pairs (excluding tandemly duplicated pairs). These common coexpressed pairs are also widespread throughout the genome (Figure 3.2).

In addition to the number of coexpressed gene pairs (duplets), the number of coexpressed triplets, quadruplets and pentaplets in the *A. thaliana* genome was determined (Table 3.2, Figure 3.2), using the strict criterion of highly correlated expression (R>0.7) of all members in a multiplet. Triplet and quadruplet coexpression domains were considerably rarer (Table 3.2), whereas coexpressed pentaplets did not occur in either the MPSS or the MA data set. To evaluate the significance of the observed numbers of the local coexpression domains in Arabidopsis, these numbers were compared with the numbers of pairs, triplets, quadruplets obtained from randomized sets using the cumulative binomial distribution (Cohen et al., 2000). Such comparisons indicated that in all cases examined, local coexpression domains ranging from 2 to 4 genes occur in the *A. thaliana* genome significantly more often than expected by chance alone (Table 3.2). Excluding tandem duplicates, coexpressed adjacent genes also occurred significantly more often than in random sets (Table 3.2). Tandem duplicates are therefore not an important explanation for the occurrence of local coexpression domains in the Arabidopsis genome.

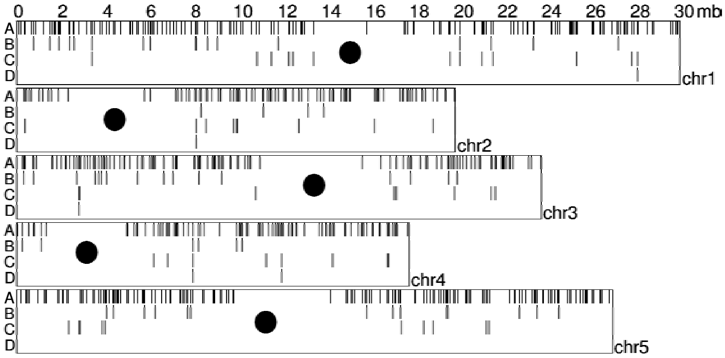


Figure 3.2 Distribution of local coexpression domains over the Arabidopsis chromosomes. Rectangles are schematic representations of chromosomes 1 to 5 from top to bottom, with black dots as centromeres. The numbers on the top show the scale in million bases along the chromosomes. Each gene is depicted as a black bar. Only the data sets excluding tandemly duplicated genes are shown. The letters on the left are: lane A, coexpressed pairs in the MPSS data set (689 pairs); lane B, common coexpressed pairs in both the MPSS and the MA data set (58 pairs); lane C, coexpressed triplets in the MPSS data set (42 triplets); lane D, coexpressed quadruplets in the MPSS data set (5 quadruplets).

Local coexpression domains are not solely explained by gene orientation and/or gene distance

Apart from tandem duplications, also gene orientation and gene distance could explain the occurrence of local coexpression domains. If promoter sharing is an important mechanism for coexpression in the Arabidopsis genome, divergently transcribed gene pairs (←*gene A gene B*→) should be over-represented in the sub-population of coexpressed pairs, compared to coexpressed pairs that are tandemly (*gene A*→*gene B*→ or ←*gene A*←*gene B*, so two possibilities) or convergently (*geneA* →←*gene B*) transcribed. For all three orientation groups, the number of pairs and the number of coexpressed pairs were determined (Table 3.3; Figure 3.3a, 3.3b). These results show that the Arabidopsis genome contains about twice as many tandemly transcribed pairs as divergently or convergently transcribed pairs. This is as expected, because the tandem orientation has two possibilities. For each orientation group, the fraction of coexpressed pairs relative to the total number of pairs in that group is plotted in Figure 3.3c. Expressed as a fraction relative to the total number of pairs in each orientation group, coexpressed divergently transcribed gene pairs occur in the same frequency as tandemly and convergently transcribed gene pairs (Figure 3.3c). There are no significant differences in the proportions of coexpressed pairs between tandem and divergent, tandem and convergent or divergent and convergent pairs (Table 3.3). These results demonstrate that divergently transcribed gene pairs are not over-represented in the subgroup of coexpressed gene pairs. Shared promoter sequences are therefore not a major explanatory variable for high coexpression between adjacent genes.

The closer two genes are, the higher the likelihood may be that their promoters influence each other, irrespective of gene orientation. Gene distance is here defined as the distance in nucleotides from the 5' start ATG of one gene to the 5' start ATG of the next gene. Thus defined, gene distance covers the distance between one coding sequence and promoter region for tandemly transcribed genes. For the other two gene orientations, this definition of gene distance results in the length of either the promoter sequence (in case of divergently transcribed genes) or the inclusion of two coding regions (in case of convergently transcribed genes). As a consequence, gene distance will favor divergently transcribed genes in the shorter distances and tandemly and convergently transcribed pairs in the larger distances. The subsequent distance analysis also distinguishes between gene orientations. All adjacent gene pairs (excluding the tandemly duplicated gene pairs) were sorted by gene distance and divided into consecutive bins of 1000 pairs. This way, any influence of unequal numbers of pairs in distance bins was prevented. For each 1000-pair bin, the number of tandemly, divergently and convergently transcribed adjacent pairs was counted and plotted against the average gene distance (Figure 3.4a). The average distance was calculated by averaging the gene distance of all pairs in each 1000-pair bin. In the same way, the number of coexpressed adjacent pairs in each orientation group was counted and plotted (Figure 3.4b). In both cases, it can be concluded that in the shorter gene distance classes divergently transcribed pairs occur more often than tandemly and convergently transcribed pairs, whereas in the larger gene distance classes divergently transcribed genes occur less often. Interestingly, coexpressed adjacent genes in any orientation could occur even at a gene distance as large as 12kb. To be able to compare the relative occurrence of the orientation groups among all the 1000-pair bins, the fraction of coexpressed pairs was plotted for each orientation group (Figure 3.4c). The fractions of coexpressed pairs stay similar among three orientation groups and also stay similar over large gene distance range. Basically identical results were obtained for the MA data set (Figure 3.4d-f). Similar results were obtained using distance bins of 1kb intervals or intergenic distances (data not shown). These results show that over a large gene distance range, the relative fraction of coexpressed pairs does not depend on gene distance, irrespective of gene orientation. Therefore, also gene distance and/or gene orientation are not important explanations for the occurrence of local coexpression domains in the Arabidopsis genome.

Table 3.3 Orientation of coexpressed gene pairs

Orientation groups[a]		Total[b]	Coexpressed[c]
MPSS			
	tan-td	6979	322 (4.61%)
	div-td	3541	191 (5.39%)
	con-td	3696	176 (4.76%)
MA			
	tan-td	7895	715 (9.06%)
	div-td	4127	396 (9.60%)
	con-td	4143	370 (8.93%)

[a]tan-td, div-td, con-td, respectively are the sub-groups of tandemly, divergently, convergently transcribed pairs excluding tandem duplicates. [b]Total number of pairs in each direction group. [c]Number of coexpressed pairs in each direction group. Percentages in the brackets are number of coexpressed pairs relative to the total number of pairs. None of the proportions are significantly different from each other according to the z test for comparing population proportions.

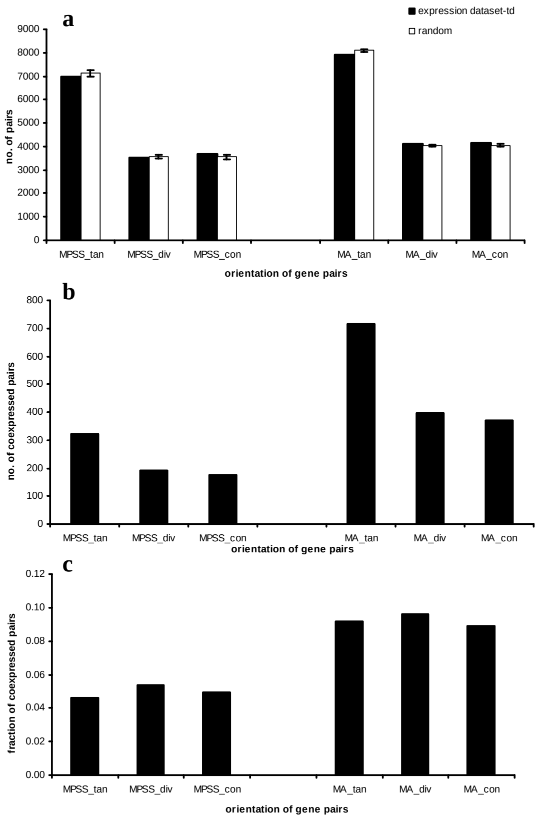


Figure 3.3 Orientation of genes in coexpressed pairs does not solely explain the occurrence of coexpression. The orientation groups based on the relative direction of transcription within a gene pair are tandem (tan), divergent (div) and convergent (con). Black bars are Arabidopsis expression data, white bars represent the averaged result from 100 randomizations. The x-axis gives the expression data set used, either MPSS or MA, without tandemly duplicated genes. **a,** the number of pairs in each orientation group; **b,** the number of coeexpressed pairs; **c,** the fractions of coexpressed pairs in each orientation group (given in **b**) relative to the total number of pairs in that corresponding orientation group (given in **A**). When corrected for the higher occurrence of tandemly oriented gene pairs, due to two possible orientations, none of the orientation groups is over-represented in coexpressed pairs.

Genes in local coexpression domains scatter over functional categories

Having estimated the number of local coexpression domains in the Arabidopsis genome, the nature of the genes involved in such chromosomal domains was analyzed. The Arabidopsis Information Resource (TAIR)’s Gene Ontology (GO) using the high-level ontology terms known as GOslims developed for plants (Berardini et al., 2004) were used to characterize the genes in local coexpression domains. Genes in coexpressed triplets and quadruplets were not examined separately and pairs consisting of tandemly duplicated genes were not included in this analysis. Using the plant GOslim terms, the genes in coexpressed pairs were classified into the divisions for molecular function (15 categories), biological process (15 categories), and cellular components (16 categories).

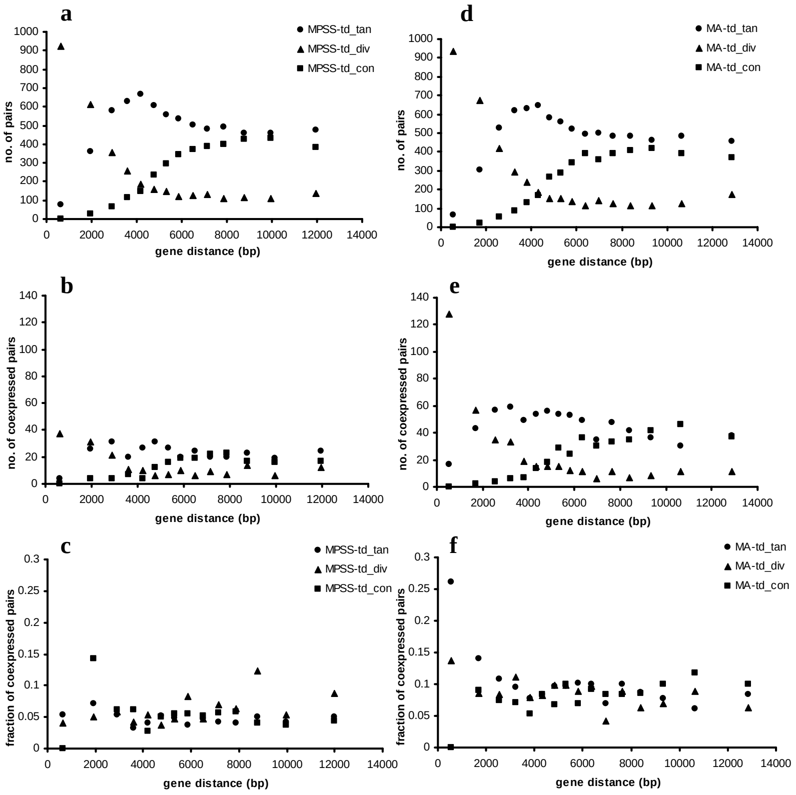


Figure 3.4 Gene distance of genes in coexpressed pairs does not solely explain the occurrence of coexpression. Gene distance, defined as start-to-start distance of adjacent gene pairs, is averaged for each 1000-pair bin and plotted as function of gene orientation, subdivided into tandem pairs (tan; rounds), divergent pairs (div; triangles) and convergent pairs (con; squares) for the MPSS data set **(a-c)** and the MA data set **(d-e)**. **a, d,** Number of pairs, **b, e**, Number of coexpressed pairs; **c, f,** the fraction of coexpressed pairs relative to the total number of pairs.

A pair was classified into a category if both members fell into the same category; otherwise the pair was classified as ‘not falling into the same category’. Pairs of which one or both genes could not be classified were not included in the analysis. About 90% of all pairs (out of 14,216 for MPSS, and 16,165 for MA; Table 3.4) or coexpressed pairs (out of 689 for MPSS and 1481 for MA; Table 3.4) could be assigned to at least one GOslim category. Classification using the MIPS Functional Catalogue (Wu et al., 2002) covered much less (about only 30%) of the genes in pairs (data not shown). In each GOslim division, there are GOslim terms for ‘unknown’ and ‘other’ (Berardini et al., 2004). These should be considered less informative for the classification of pairs of genes. Therefore, we have distinguished a subclass of genes falling into the well-defined categories, excluding all categories with ‘unknown’ and ‘other’. The results are summarized in Table 3.4. Considering the GOslim division for molecular function (GO_func), about 22% of the coexpressed pairs consist of genes that fall in the same functional category for both the MPSS and MA data sets (Table 3.4). For biological process (GO_proc), this is about 43% and for cellular component (GO_comp) this is 29%. When limited to the genes in categories that have no indication of ‘other’ or ‘unknown’, about 6-7% of the pairs have genes that classify in the same category. Compared to the distribution of the genes of all pairs, the percentages of pairs in the same functional category do not differ significantly (at $P<0.01$). Therefore, coexpressed pairs do not tend to fall more in the same GOslim category than other gene pairs (Table 3.4). Compared to what is expected on the basis of randomized genomes, the percentage of coexpressed genes falling in the same GOslim is not different from what is found in random genomes, with the notable exception of the percentage of genes that fall in the same category of well-defined biological processes. In both the MPSS and the MA data, about three times (6-7% versus 2% expected) more coexpressed pairs occur in this category than expected on the basis of a random distribution. Within the category of well defined biological processes, the category ‘protein metabolism’ is overrepresented in both data sets: 43% (18 from 43) for MPSS and 61% (48 from 79) for MA of the pairs fall in this particular GOslim category.

Discussion

For different organisms, it has been demonstrated that appreciable numbers of genes in a genome occur in clusters characterized by correlated expression. Averaging coexpression over size-based or gene number-based windows showed that about 20% of the Drosophila genome resides in coexpression clusters (Spellman and Rubin, 2002). Within the Arabidopsis genome, such window-based coexpression clusters may consist of up to 20 genes (Birnbaum et al., 2003; Zhu, 2003; Williams and Bowles, 2004), while some evidence

Table 3.4 Distribution of gene pairs over GOslim categories

	Genome			Random	
	All[a]	Coexpressed[b]	P-val [c]	coexpressed[d]	P-val [e]
MPSS					
GO_func					
Covered[f]	12920	624		537	
SameCat[g]	2662 (20.6%)	136 (21.8%)	0.48	114 (21.2%)	0.80
SameKnCat[h]	1041 (8.06%)	43 (6.89%)	0.26	43 (7.95%)	0.49
GO_proc					
Covered	12927	623		537	
SameCat	5366 (41.5%)	268 (43.0%)	0.46	225 (41.9%)	0.71
SameKnCat	867 (6.71%)	43 (6.90%)	0.86	11 (2.05%)	<0.0001 *
GO_comp					
Covered	13043	628		539	
SameCat	3314 (25.4%)	181 (28.8%)	0.07	138 (25.6%)	0.22
SameKnCat	789 (6.05%)	42 (6.69%)	0.41	29 (5.38%)	0.35
MA					
GO_func					
Covered	14804	1304		1117	
SameCat	3234 (21.8%)	286 (21.9%)	0.93	245 (21.9%)	1.0
SameKnCat	1147 (7.75%)	99 (7.59%)	0.83	99 (8.86%)	0.26
GO_proc					
Covered	14770	1316		1115	
SameCat	6132 (41.5%)	552 (42.0%)	0.73	470 (42.2%)	0.92
SameKnCat	931 (6.30%)	79 (6.00%)	0.66	24 (2.15%)	<0.0001 *
GO_comp					
Covered	14756	1317		1115	
SameCat	3806 (25.8%)	375 (28.5%)	0.04	289 (25.9%)	0.15
SameKnCat	811 (5.50%)	97 (7.37%)	0.012	77 (6.91%)	0.66

[a] Number of neighboring pairs included in the analysis.

[b] Number of coexpressed pairs included in the analysis.

[c] P value, the probability under the null hypothesis that the two population proportions are the same, derived from the standard normal tables of the z statistic for the difference of the population proportion between coexpressed pairs and all the pairs; *, significant (two-tailed; P < 0.01).

[d] Number of coexpressed pairs in random sets.

[e] P value, the probability under the null hypothesis that the two population proportions are the same, derived from the standard normal tables of the z statistic for the difference of the population proportion between coexpressed pairs of the Arabidopsis genome and coexpressed pairs in randomized sets; *, significant (two-tailed; P<0.01).

[f] Number of pairs of which both members are falling in a GOslim category.

[g] Number of pairs of which both members fall into the same GOslim category. Percentage is the number of pairs relative to the total number of pairs covered.

[h] Number of pairs of which both members fall into the same "known" GOslim category (excluding the categories with the indications 'unknown' and 'other'). Percentage is the number of pairs relative to the number of pairs covered.

from quantitative trait loci studies suggested that clusters may be much larger (Khavkin and Coe, 1997). These data support the notion of higher-level genome organization that may range over distances up to several mega-bases (Hurst et al., 2004). Yet the concept of large coexpression clusters in such studies is based on a loose definition of the term 'cluster' or 'chromosomal domain' and associated terms such as neighboring. The process of summing and averaging may obscure local effects and underrate the presence and/or role of individual genes with different expression levels or expression patterns in large clusters. Therefore, it is perhaps not surprising that coexpression clusters are often associated with the activity of housekeeping (Lercher et al., 2002; Roy et al., 2002; Lercher et al., 2003), or highly expressed (Caron et al., 2001; Versteeg et al., 2003) genes. Previous experience with transgene expression data indicated that the particular position of a single gene in a genome affects the expression of that gene. Depending on chromosomal context, two physically neighboring transgenes could be made to show correlated expression (Mlynarova et al., 2002). Therefore, we have here taken a very rigorous approach to the concept of 'cluster' and analyzed the coexpression characteristics of genes that are physically adjacent in the genome according to genome annotation data.

Whole-genome chromosomal coexpression maps indicate the existence of numerous cases of local coexpression (Figure 3.1), as was also shown in yeast (Cohen et al., 2000). Combining expression data and genome annotation, we identified 16,144 adjacent pairs of genes with sufficient expression data in the MPSS data set. The arbitrary criterion taken for inclusion of a gene in the analysis was detectable expression in at least one of the data libraries available. Although some genes may then have expression only in one library, around 80% of the genes have expression data in at least three different libraries and this is likely to yield reliable results. A major issue in such coexpression analyses is the occurrence of tandemly duplicated genes (Zhu, 2003; Hurst et al., 2004). Tandemly duplicated genes could be considered a trivial case of coexpression. All analyses, except when indicated, were performed with and without tandemly duplicated genes. From the 16,144 pairs, 12% were identified as tandemly duplicated genes. Only 11% of these tandemly duplicated gene pairs identified showed coexpression, which is 24% of all pairs with coexpression (Table 3.1). The MA data set corroborates the MPSS findings: only 14% of the tandemly duplicated pairs showed coexpression. This suggests that, in contrast to inferences made for other genomes (Lercher et al., 2003), tandemly duplicated genes in the Arabidopsis genome are not a major cause of correlated expression of adjacent genes. Also the particular orientation of the tandemly duplicated genes, either tandemly, divergently or convergently transcribed, was found to have no significantly higher inclination to be coexpressed, in contrast to the conclusions of the analyses of (Williams and Bowles, 2004). As the MA and MPSS data sets used in this study are biologically very different, their agreement with respect to the relative unimportance of tandemly duplicated genes in our analysis, suggests that the data sets used in the respective analyses need to be considered. Careful future comparisons of data sets, gene coverage and analytical methods used will have to reveal the cause of such differences.

From all non-tandemly duplicated pairs in the MPSS data set, 4.9% shows coexpression. They are distributed over the whole genome (Figure 3.2). Although this is a low percentage, randomization assays indicate that the number is significantly higher than to be expected by chance alone (Table 3.2), There is a small yet significant fraction of the Arabidopsis genome that shows correlated expression between neighboring genes. Enlarging such local clusters by looking for series of consecutive genes that are correlated in all pair-wise combinations reveals that there are few areas in the Arabidopsis genome that consist of more than two (up to four) genes (Table 3.2, Figure 3.2) with highly correlated expression. The size of these local coexpression domains is in agreement with local cluster sizes observed in yeast (Cohen et al., 2000) and worm (Roy et al., 2002). The microarray (MA) data set, despite its technologically different approach for obtaining expression data, and its biologically different experimental background, also showed local coexpression domains ranging from 2 to 4 genes distributed over the genome.

Over the whole genome, the two expression data sets show areas that have different coexpression patterns (Figure 3.1) and in total only 58 coexpressed pairs were shared between both data sets (Figure 3.2). These differences in coexpression and low number of shared pairs are likely to reflect the biological differences between the data sets. The MA data are well-defined root cells and tissues, while the MPSS data concern more broad tissues and organs. Such biological differences will influence correlations in gene activity. Any expression data set will present a fixed average of expression over the sampled cells, tissues, organs and experiments that should be taken into account when comparing such data sets.

To understand the possible causes for coexpression, the role of shared promoters and/or short gene distances was analyzed. The population of divergently transcribed genes does not contain a higher proportion of coexpressed genes compared to tandemly or convergently transcribed genes (Figure 3.3; Table 3.3). Promoter sharing is therefore not a likely explanation for the presence of local coexpression domains in the Arabidopsis genome, unlike the situation in the yeast genome (Cohen et al., 2000). Also gene distance does not offer an important explanation for the occurrence of local coexpression domains. When corrected for gene orientation, the fraction of coexpressed genes does not depend on either gene orientation or gene distance (Figure 3.4c, 3.4f). Short gene distances (<1kb) do not favor local coexpression and longer distances (up to 10 kb) need not necessarily be barriers to local coexpression. In this analysis, gene distance is defined as the distance from the 5' start ATG of one gene to the 5' start ATG of the next gene and includes the coding region of a gene (for tandemly transcribed gene pairs) or of both genes (for convergently transcribed gene pairs). Similar results were obtained when the intergenic distance, defined as the distance between the stop codon of one gene and the start codon of the next gene, was taken for analysis (data not shown). Therefore, the precise definition of gene distance in the analyses as presented does not affect the conclusions. The role between gene distance and correlation of expression has given different results in different studies. Some indicate that correlation declines with increasing distance (Cohen et al., 2000; Williams and Bowles, 2004), while others are less explicit and emphasize the role of relative genome compactness

(Fukuoka et al., 2004). Analyses of the MA data with the TIGR5 annotation of the Arabidopsis genome had no significant effect on trends and conclusions (data not shown). Previous studies suggested that clustering of functionally related genes might occur in all metazoans (including yeast, fly, worm, human) (Cohen et al., 2000; Lercher et al., 2002; Spellman and Rubin, 2002; Lercher et al., 2003). A recent study (Williams and Bowles, 2004) demonstrated a significant enrichment for coexpressed genes in the same metabolic pathway, although this appeared not to be an explanation for the neighboring coexpression. In this study, a loose definition of neighboring was used, defining two genes as neighboring when they were within ten genes of each other (Williams and Bowles, 2004). In worm, clusters of similarly expressed genes cover similar biological functions (Roy et al., 2002). In human, coexpression analysis over the whole genome was shown to correlate with functional relatedness (Lee et al., 2004). In the expression data sets here analyzed with a gene ontology developed for plants (GOslim; (Berardini et al., 2004)), there is however no evidence that coexpressed genes in pairs are enriched in the same functional category compared to all genes in pairs (Table 3.4). This is also the case when compared to the percentages of coexpressed genes in random sets (Table 3.4). When the GOslim categories without ‘unknown’ or ‘other’ are used, only in the GOslim division covering ‘biological process’ coexpressed gene pairs are about three times more frequently present than expected to occur by chance, notably in the GOslim biological process category protein metabolism. In the other GOslim divisions, no such trend is present: coexpressed gene pairs are as frequently present as all gene pairs (Table 3.4).

In our coexpression analyses of expression data, different libraries from either MPSS or MA data were combined irrespective of the biological characteristics of the material assayed. Therefore, the analyses have revealed the gene pairs that show stringent coexpression under a range of different (biological) conditions, cells and/or tissue types. Combining more and different data sets, such as the data in various Arabidopsis expression repositories now available at The Arabidopsis Information Resource (TAIR; (Rhee et al., 2003)), NCBI’s Gene Expression Omnibus (GEO; (Edgar et al., 2002)), Genevestigator (Zimmermann et al., 2004), Stanford Microarray Database (SMD; (Gollub et al., 2003)), or the *Arabidopsis thaliana* Tissue-Specific Expression Database (ATTED; (Obayashi et al., 2004)), will help to analyze the expression of genes over various conditions and cell types. Yet averaging more and different expression data sets would continue to favor the identification of gene pairs expressed under all conditions in as many cell and tissue types as available in expression repositories. Although this would reveal the expression potential of gene pairs in a genome, it would be much less informative for elucidating the whole-genome dynamics of coexpression. Local coexpression domains may be dynamic during growth and development of plants. In future analyses, it may therefore be worthwhile to analyze pre-chosen subsets of libraries and compare the local coexpression dynamics of different organs, tissues or cells to identify time- or tissue-specific local coexpression domains.

True neighboring pairs can form local coexpression domains of 2-4 genes irrespective of gene orientation or gene distance. Having eliminated such configuration factors, a role of either the gene sequence itself or the DNA sequences surrounding these genes is suggested. In the transgenic situation, it was shown before that the expression of two unrelated genes became correlated when their surrounding DNA was supplied with a chromatin boundary element (Mlynarova et al., 2002). A next step of genome analysis will therefore be the detailed analysis of the sequences next to local coexpression domains. These may consist of boundary elements such as matrix-associated regions (Boulikas, 1995; Bell et al., 2001), and help to further define the (sequence) characteristics of such elements.

Conclusion

Defining local coexpression domains as genome areas with physically neighboring genes showing tight coexpression, we have here shown that the Arabidopsis genome contains a small yet significant number of coexpression domains that range from two to four genes. Neither tandemly duplicated genes, nor divergently transcribed promoter regions, or short gene distances explain such local coexpression of adjacent genes. Either gene sequence or the surrounding DNA sequences are of importance for the coexpression pattern of such neighboring genes. Our study and the further unraveling of the relationships between local and global coexpression domains in relationship to surrounding DNA, gene regulation and chromosome structure will help to gain understanding of the molecular mechanisms that establish local chromosomal domains of genes with high coexpression characteristics.

Materials and Methods

Data retrieval and processing

The *Arabidopsis thaliana* genome annotation from the March 2003 version of Munich Information Center for Protein Sequences (MIPS; (Schoof et al., 2002)) has 26,439 annotated genes on 5 chromosomes. Mitochondria and chloroplast genes were not taken into account in this study. There are 6813, 4181, 5363, 3987 and 6095 genes on chromosome 1 to 5, respectively. The genes along each chromosome were sorted based on ascending start coordinates and were numbered consecutively. This established a rank number (rank ID) that helped to eliminate any discontinuity in the Arabidopsis Genome Initiative (AGI) numbers of the annotated genes and allowed analyzing physically adjacent genes. These rank IDs of genes were used to compare two different whole-genome expression data sets, Massively Parallel Signature Sequencing (MPSS) expression data and microarray (MA) expression data. Data are summarized in Table 3.1. The MPSS data was obtained from the Arabidopsis MPSS website (mpss.udel.edu/at/java.html; (Meyers et al., 2004)). The MPSS data set has 14 libraries covering 5 plant tissues: callus, inflorescence, leaf, root and silique. All MPSS 17bp signatures that had a normalized expression abundance of at least 1 transcripts per million (TPM) in at least one of the 14 libraries were retrieved manually. Genes without MPSS signature or with no expression value of at least 1 TPM in any of the 14 libraries were not taken into consideration. With Python scripts, the MPSS signatures were mapped onto the MIPS genome annotation, based on an exact match of 17 bp and assigned the corresponding chromosomal position. Each signature that was assigned more than once was removed from the data set. Each MPSS mapped signature was assigned to a class based on the genomic location and MIPS annotation. Seven different classes were defined according to the criteria on the MPSS website (mpss.udel.edu/at/java.html): class 1 (inside an annotated gene/feature); class 2 (within 250 bp 3' of the annotated gene/feature); class 3 (anti-sense to annotated gene/feature); class 4 (between gene/feature); class 5 (within intron, sense strand); class 6 (within intron, anti-sense strand) and class 7 (within 17 bp of an exon boundary; spliced). With the precedence ranking of classifications: 1 = 7 > 2 > 5 > 3 > 6 > 4 for signatures belonging to more than one possible class, every signature was assigned to only one class. The normalized expression values of both class 1 and class 2 signatures in the same library were summed and used as the expression value of the corresponding gene. Genes with neither class 1 nor class 2 signatures were considered to be not expressed and were not taken into consideration. This way, we obtained 20,041 genes having MPSS expression values, referred to as the MPSS data set.

The microarray (MA) expression data was obtained from the online supplementary material of a Science article (Birnbaum et al., 2003). The MA data set based on the ATH1 GeneChip (Affymetrix, Santa Clara, CA) has expression data only from Arabidopsis root tissue, encompassing 15 different zones of the root that correspond to different cell types and tissues at progressive developmental stages. Genes not on the array were not taken into consideration. Genes on the array of which the AGI numbers could not be mapped to the MIPS genome annotation were also discarded. After mapping these gene expression data by their unique AGI numbers to the MIPS annotation, we obtained 21,940 genes having MA expression values, referred to as the MA data set. Analyses of the MA data with the TIGRV annotation of the Arabidopsis genome had no major effect on trends and conclusions (data not shown)

In case of physically overlapping genes in either data set, the smaller one of the overlapping genes was removed from the data set, by which both gene and rank ID orders were maintained. For this reason, 39 and 34 genes were removed from the MPSS and the MA data set, respectively. The resulting data sets used for analysis consisted of 20,002 genes with MPSS expression data and 21,906 genes with MA expression data. Two genes were considered to be adjacent when their rank IDs were consecutive with a difference of one, and when the genome sequence had no long stretches of N's in-between. In six cases (three on Chr1 and three on Chr2), the genome sequence was interrupted by a stretch of 60 or 120 N's. These genes were included in the subsequent analyses. Adjacent genes were considered per chromosome. With these criteria, 16,144 adjacent gene pairs were identified in the MPSS data set. These pairs comprised 19,151 genes with expression values. A total of 851 (that is, the difference between 20,002 and 19,151) genes in the MPSS data set had no neighbors with expression data. In the MA data set, 18,443 adjacent gene pairs in the MA data set were identified, comprising 21,255 genes, and 651 isolated genes without expressed neighboring genes (Table 3.1).

Tandemly duplicated genes were identified by local pair-wise protein BLAST (BLASTP 2.2.6 [Apr-09-2003]; (Altschul et al., 1997)), on all gene pairs in both data sets. A gene pair was considered to be a tandemly duplicated (td) pair if BLASTP yielded $E < 2 \times 10^{-1}$ (Lercher et al., 2003; Fukuoka et al., 2004; Williams and Bowles, 2004)$^{.}$ This criterion, developed on the basis of duplicated human genes, removes about 90% of the related genes from a population and has a false positive rate of about 10% (Lercher et al., 2002; Williams and

Bowles, 2004). This way, 1928 and 2278 adjacent pairs were identified in the MPSS and the MA data set, respectively. Most analyses were done for data sets including td or excluding td. Such exclusion implied that the td pair was not included in the coexpression analysis, but the expression of each member of a tandemly duplicated gene pair was analyzed relative to its other neighbor.

Identification of local coexpression domains

Pearson's correlation coefficient (R) was calculated between all adjacent pairs (duplets) of genes using the expression data from all libraries available in each data set. If R was higher than 0.7, the gene pair concerned was considered to be coexpressed. The value of R>0.7 is generally considered a rule-of-thumb threshold (see for example bbc.botany.utoronto.ca/affydb/BAR_instructions and is used in various analyses (Cohen et al., 2000; Lee et al., 2004). When calculating the R values from a whole-genome all-against-all comparison (used to establish Figure 3.1), and plotting these as a histogram, the top 5% in this distribution may be used to derive a threshold for determining coexpression, analogous to the 5% upper tail in a normal distribution. For the MPSS data, the upper 5% cut-off is 0.65 and for the MA data 0.72. For convenience and comparability, the approximate average of R>0.7 was chosen for analysis. With a lower threshold value for R, such as for example R>0.5 (Blanc and Wolfe, 2004), the absolute numbers of the various categories of genes go up, but the relative results do not change dramatically from what is presented (data not shown).

The number of coexpressed adjacent pairs was counted. To evaluate the statistical significance of these numbers, they were compared with the number of coexpressed pairs from 100 randomizations of the population of expressed genes using the cumulative binomial distribution (Cohen et al., 2000). Preliminary analyses indicated that more than 100 randomizations did not result in significant changes in the numbers obtained (data not shown). In each round of randomization, non-adjacent pairs of genes were randomly selected with replacement from the list of expressed genes that have expressed neighbors till the same total number of pairs was obtained. For example, the MPSS data set has 16,144 gene pairs that are neighboring genes with expression values. One round of randomization on the MPSS data set consisted of 16,144 times of randomly picking two genes with replacement out of the list of genes represented in the 16,144 gene pairs, calculating R for each pair and counting the number of pairs having R>0.7. Similarly, coexpressed adjacent triplets, quadruplets and pentaplets were identified as series of genes with consecutive IDs in which all possible (that is, (n!/(n-2)!)*2; kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf) pair-wise Rs were above the cut-off of 0.7. The significance of results was evaluated with randomizations equivalent to the procedure used in case of duplets.

The role of gene direction and gene distance in Local coexpression domains

Adjacent gene pairs were separated into tandemly, divergently and convergently transcribed pairs according to their relative direction of transcription. The number of coexpressed pairs in each orientation group was expressed as percentage relative to the total number of adjacent pairs in that group. Random pairs were made by randomly picking two non-adjacent genes from the list of expressed genes represented in pairs, analyzed for their orientation and compared with the real genome using a variant of the two-sample t test for proportions for determining the significance of a difference between two population proportions (Ott and Longnecker, 2001). The test statistic is based on the z statistic from the normal distribution and is given by (p1–p2)/ √ (p1*(1-p1)/n1 + p2*(1-p2)/n2), with p1 and p2 the two sample proportions, n1 and n2 the two sample sizes, under the condition that n1*p1, n1*(1-p1), n2*p2 and n2*(1-p2) are all larger than 5. When | z | > 2.575, the two sample proportions are considered to be significantly different at the 1% level (P<0.01). The z value is converted to a p value using standard normal tables.

For gene distance, the length in nucleotides from the 5' start of one gene to the 5' start of the next gene was used. The data sets excluding the tandemly duplicated gene pairs were analyzed. This way, there were 14,216 pairs in the MPSS-td data set and 16,165 pairs in MA-td data set. For each data set, gene pairs were sorted based on gene distance and bins of 1000 pairs were taken and analyzed, excluding the last bin with less than 1000 pairs. Per 1000-pair bin, gene distance was calculated as the average over all 1000 pairs. In total, 14 bins of 1000 pairs for the MPSS-td data set and 16 bins of 1000 pairs in MA data set were analyzed. Within each 1000-pair bin, the numbers of tandem, divergent and convergent pairs were determined, as well as the numbers of coexpressed pairs within each orientation group. To be able to compare bins, the fraction of coexpressed pairs relative to the total number of pairs in each orientation group in each bin was calculated.

Functional categorization of genes represented in local coexpression domains

TAIR's GOslim, the Gene Ontology (GO) developed for plants (Berardini et al. 2004) was used to classify the genes present in local coexpression domains. The categories for molecular function (15 GOslim categories), biological process (15 GOslim categories) and cellular component (16 GOslim categories) were taken in consideration. With Python scripts, the number of pairs of which both members could be classified in GOslim was determined from the total number of coexpressed pairs. From this, the number of pairs of which both members fall in the same GOslim category was determined, also with the help of Python scripts. The GOslim categories include 'unknown' and 'other'. These were considered to give less information about functional categorization and were set apart from the genes falling into a well-defined category. The percentages obtained were compared with random sets using the z test for the significance of difference between two proportions (Ott and Longnecker, 2001) as outlined above.

Chapter 4

Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains

Xin-Ying Ren[1, 2], Willem J. Stiekema[2, 3] & Jan-Peter Nap[1, 3]

[1]Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands
[2]Laboratory of Bioinforamtics, Plant Sciences Group, Wageningen University and Research Center, 6703 HA Wageningen, the Netherlands
[3]Centre for BioSystems Genomics, 6700 AA Wageningen, The Netherlands

Submitted

Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains

Abstract

Chromosomal coexpression domains are found in a number of different genomes under various developmental conditions. The size of these domains and the number of genes they contain vary. We here define local coexpression domains as adjacent genes where all possible pair-wise correlations of expression data are higher than 0.7. In rice, such local coexpression domains range from 2 - 4 genes and make up ~5% of the genomic neighboring genes. The genes in local coexpression domains do not fall in the same ontology category significantly more than neighboring genes that are not coexpressed. Duplication, orientation or the distance between the genes does not solely explain coexpression. The regulation of coexpression is therefore thought to be regulated at the level of chromatin structure. The characteristics of the local coexpression domains in rice are strikingly similar to such domains in the Arabidopsis genome. Yet, no microsynteny between local coexpresion domains in Arabidopsis and rice could be identified. Although the rice genome is not yet as extensively annotated as the Arabidopsis genome, the lack of conservation of local coexpression domains indicates that such domains have not played a major role in evolution.

Introduction

The fast-growing data sets on genome annotation and genome-wide gene expression facilitate the study and comparison of gene activity between and among genomes. The genomic context of genes is supposed to play an important role in the regulation of gene expression (van Drunen et al., 1997). Non-random clusters of similarly expressed (co-regulated, coexpressed, highly expressed and/or broadly expressed) genes have been described in almost all organisms, ranging from prokaryotes to eukaryotes. In eukaryotes, from yeast to Arabidopsis to human, both short-range co-regulated/coexpressed clusters of two to five genes (Cohen et al., 2000; Ren et al., 2005; Zhan et al., 2006) and longer-range coexpression domains of up to 30 genes spanning up to 100 kb and more (Spellman and Rubin, 2002; Lercher et al., 2003; Zhan et al., 2006) have been described. Duplicated genes (Lercher et al., 2003), shared promoter regions (Kruglyak and Tang, 2000), shorter gene distance (Cohen et al., 2000; Roy et al., 2002; Williams and Bowles, 2004; Semon and Duret, 2006) and/or functional relatedness (Cohen et al., 2000; Spellman and Rubin, 2002; Lee et al., 2004; Williams and Bowles, 2004) have found to account for only part of the coexpression between genes. Therefore, most studies postulate that the occurrence of coexpression domains, small or large, is regulated on the level of higher-order chromatin structure (Cohen et al., 2000; Spellman and Rubin, 2002; Williams and Bowles, 2004; Hershberg et al., 2005; Ren et al., 2005), although alternative views exist (Semon and Duret, 2006).

Previously we have defined and demonstrated the existence of local coexpression domains in the genome of Arabidopsis. A local coexpression domain was defined as any set of physically adjacent genes that are highly coexpressed with a pairwise Pearson’s correlation coefficient larger than 0.7. It was shown that a small (5%–10%) yet significant fraction of genes in the Arabidopsis genome is organized in such local coexpression domains. Genes in such local domains were for the major part not categorized in the same functional category (GOslim). Neither tandemly duplicated genes nor shared promoter sequence nor gene distance explained the occurrence of coexpression of genes in such chromosomal domains. This indicates that other parameters in genes or gene positions are important to establish coexpression in local domains of Arabidopsis chromosomes. Here it is analysed whether a similar situation exist in the genome of the monocotyledonous model plant rice (*Oryza sativa*). We combined the whole genome rice annotation data (TIGR version 3; www.tigr.com) with the expression data from the MPSS platform (mpss.udel.edu/rice/) in a way similar to the analysis performed for the Arabidopsis genome (Ren et al., 2005). The results show that the characteristics of the two genomes with respect to the occurrence and configuration of local expression domains are remarkably similar.

Also in the rice genome, a small yet significant fraction of genes is organized in local coexpression domains that consist of 2-4 genes that are not categorized in the same functional category. The presence of tandemly duplicated genes, shared promoter sequence or gene distance is not fully explaining the occurrence of coexpression of genes in such chromosomal domains. Therefore, the regulation of local coexpression domains is postulated to the level of higher-order of chromatin structure.

Given the similarities in the characteristics and occurrence of coexpression domains between Arabidopsis and rice, we investigated whether the genes involved showed microsynteny between the two genomes. These analyses did not identify the presence of syntenic local coexpression domains between Arabidopsis and rice.

Results

Local coexpression domains consist of two to four neighboring genes

The current version of the rice genome annotation (TIGR version 3) has 57,915 predicted genes. This is about twice the number of genes predicted for Arabidopsis (28,952 genes; TIGR5 annotation). The current MPSS expression data coverage for the rice genome is 41% (Table 4.1), which is almost half of the expression coverage for the Arabidopsis genome (72% in TIGR5 update; (Ren et al., 2005)). This reflects the more advanced annotation of the Arabidopsis genome. The rice genome has more genes that are physically overlapping than the Arabidopsis genome. Excluding the smaller overlapping genes from the analyses, we were able to identify 12,920 gene pairs with expression in rice (Table 4.1; see also Materials and Methods). Of these, 584 (4.5%) were identified to represent a local coexpression domain as defined as being coexpressed with a pairwise Pearson’s correlation coefficient larger than 0.7 (Table 4.1). This percentage is similar to what we have found previously for Arabidopsis (Ren et al., 2005) and agrees well with other findings that ~3-5% of a genome is tightly coexpressed (Semon and Duret, 2006).

Notably duplicated genes are supposed to influence coexpression statistics due to their common origin (Lercher et al., 2003), although a somewhat surprising finding for the Arabidopsis coexpresssion domains was that only a minor fraction of duplicated genes were actually coexpressed (Ren et al., 2005). The occurrence of duplicated pairs in the rice set was determined with pair-wise protein BLAST using a cut-off of E< 0.2 (Lercher et al., 2003; Fukuoka et al., 2004; Williams and Bowles, 2004). This identified 1,663 (12.9%) duplicated rice gene pairs in the whole set of gene pairs.

Of these, only 146 (8.8%) were coexpressed (Table 4.1). Although this percentage is two times higher than the percentage of coexpression in non-duplicated pairs (3.9%), the majority of all duplicated pairs (91.2%) are not coexpressed. This shows that also in rice gene duplication does not correlate well with coexpression and suggests that expression divergence is a common phenomenon after duplication (Williams and Bowles, 2004). Excluding the duplicated pairs from the coexpressed set, there are 438 gene pairs coexpressed in rice. This accounts for 75% (=438/584) of all coexpressed pairs. Therefore, also in rice the occurrence of duplicated genes cannot explain all local coexpression domains.

Table 4.1 Description of expression data used for whole-genome analysis

		Rice MPSS
Total no. genes		57915
Genes with expression		
	Total	23510 (41%)[a]
	Overlapping	364
	Left	23146
	Without expressed neighbor(s)	5081
	Represented in pairs	18065
Adjacent pairs		
	Total	12920
	Coexpressed	584 (4.5%)[b]
	Tandemly duplicated pairs (td)	1663 (12.9%)[c]
	Total excluding td	11257
	Coexpressed excluding td	438 (3.9%)[d]
Coexpressed adjacent pairs		
	Total	584
	Coexpressed excluding td	438 (75%)[e]
Tandemly duplicated pairs		
	Total	1663
	Coexpressed	146 (8.8%)[f]

[a] The percentage of MPSS expression coverage of the whole genome
[b] The percentage of coexpressed adjacent pairs relative to the total number of adjacent pairs.
[c] The percentage of tandem duplicated pairs relative to the total number of adjacent pairs.
[d] The percentage of coexpressed adjacent pairs excluding td relative to the total number of adjacent pairs excluding tandemly duplicated pairs.
[e] The percentage of coexpressed excluding tandemly duplicated pairs relative to the total number of coexpressed adjacent pairs.
[f] The percentage of coexpressed tandemly duplicated pairs relative to the total number of tandem duplicated pairs.

Extending the size of the local coexpresion domain to triplets, quadruplets, pentaplets and so on, requiring that all pairwise combinations of genes have a tightly correlated expression, shows that few larger local coexpression domains exist (Table 4.2). No pentaplet domains could be identified. Local coexpression domains therefore consist of at most 4 genes, as was the case in the Arabidopsis genome (Ren et al., 2005). Excluding the presence of duplicated genes, there is even no local coexpression domain of four genes in the representation of the rice genome here analysed (Table 4.2).

To access the significance of the occurrence of the various local coexpression domains, we compared the number of coexpressed pairs, triplets and quadruplets with the average of such domains in 100 randomly generated genomes using the cumulative binomial distribution (Cohen et al., 2000). Such comparisons revealed that local coexpression pairs occur in the rice genome significantly more often than expected by chance alone (Table 4.2). However, when excluding the duplicated genes, triplets and quadruplets do not occur significantly more often than by chance (at a P value < 0.01). All subsequent analyses were focused on domains consisting of non-duplicated gene pairs, unless stated differently.

Table 4.2 Number of local coexpression domains ranging 2 to 4 genes

		Real genome		Random genome (100x)	
		Total[a]	Coexpressed[b]	Average[c]	P-value[d]
Pairs					
	+td	12920	584 (4.52%)	408 ± 17	1.46×10^{-17}
	-td	11257	438 (3.89%)	356 ± 21	2.17×10^{-6}
Triplets					
	+td	7775	23 (0.30%)	8.78 ± 2.9	2.95×10^{-5}
	-td	6831	13 (0.19%)	7.74 ± 3.0	0.025
Quadruplets					
	+td	4887	3 (0.06%)	0.24 ± 0.47	1.81×10^{-3}
	-td	4318	0 (0%)	0.18 ± 0.39	0.835

[a]Total number of pairs, triplets, quadruplets in each data set.
[b]Coexpressed pairs, triplets, quadruplets in each data set. Percentages in the brackets are coexpressed relative to the total.
[c]Average and standard deviation from 100 times randomizations.
[d]P-value according to the cumulative binomial distribution (Cohen et al., 2000) for obtaining such result by chance.
[e] MPSS data set including tandem duplicates.
[f] MPSS data set excluding tandem duplicates.

Orientation and distance do not solely explain the occurrence of local coexpression

In yeast, there are several examples that divergently transcribed promoter regions are the cause of co-regulated neighboring genes (Kruglyak and Tang, 2000; Korbel et al., 2004). If promoter sharing is an important mechanism for coexpression in the rice genome, divergently transcribed gene pairs should be over-represented in the sub-population of coexpressed pairs, compared to coexpressed pairs that are tandemly or convergently transcribed. For all three-orientation groups, the number of pairs and the number of coexpressed pairs in the rice genome were determined (Table 4.3). For each orientation group, the fraction of coexpressed pairs relative to the total number of pairs in that group was calculated (Table 4.3) and plotted in Figure 4.1. None of the fractions are significantly different from each other using a statistical test for comparing population proportions (Ott and Longnecker, 2001). The fraction of coexpressed divergent pairs is the lowest of the three groups (Figure 4.1 & Table 4.3). Therefore, shared promoter regions cannot solely explain the coexpression of adjacent genes.

Table 4.3 Orientation of coexpressed gene pairs

Orientation groups[a]	Total[b]	Coexpressed[c]
tan-td	5621	239 (4.25%)
div-td	2418	82 (3.39%)
con-td	3218	117 (3.64%)

[a]tan-td, div-td, con-td, respectively are the groups of tandemly, divergently, convergently transcribed pairs excluding tandem duplicates.
[b]Total number of pairs in each direction group.
[c]Number of coexpressed pairs in each direction group. Percentages in the brackets are number of coexpressed pairs relative to the total number of pairs. None of the proportions are significantly different from each other according to the statistical test for comparing population proportions.

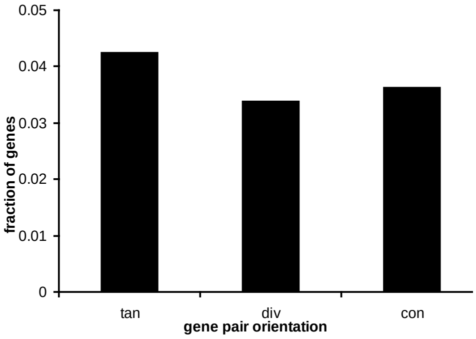


Figure 4.1 Orientation of genes in coexpressed pairs does not solely explain the occurrence of coexpression. The orientation groups based on the relative direction of transcription within a gene pair without duplications are tandem (tan), divergent (div) and convergent (con). The fractions of coexpressed pairs in each orientation group relative to the total number of pairs in that corresponding orientation group are plotted. None of the orientation groups is over-represented in coexpressed pairs.

The closer two genes are, the higher the likelihood is that they are coexpressed due to either *cis* or *trans*-activation (Hershberg et al., 2005). Therefore, we determined the intergenic distance, defined as the sequence length in nucleotides from the annotated end of one gene to the annotated start of the neighboring gene, including UTRs when known, otherwise taking the start and stop site for translation. This distance was used to investigate whether it would explain local coexpression domains. In Figure 4.2, the fraction of coexpressed pairs is plotted for each orientation and for each 1000-pair bin after sorting based on intergenic distance. The results show that the fraction of coexpressed pairs, irrespective of gene orientation, does not decrease with larger gene distance. When gene distance is defined as the sequence length from the start of one gene till the start of the next gene (Ren et al., 2005) the result is similar (data not shown). As a consequence, increasing intergenic distances do not seem to be a barrier for the occurrence of local coexpression and short intergenic distances do not favor coexpression. Therefore, intergenic distance does not solely explain local coexpression in the rice genome, as it did not in the Arabidopsis genome (Ren et al., 2005).

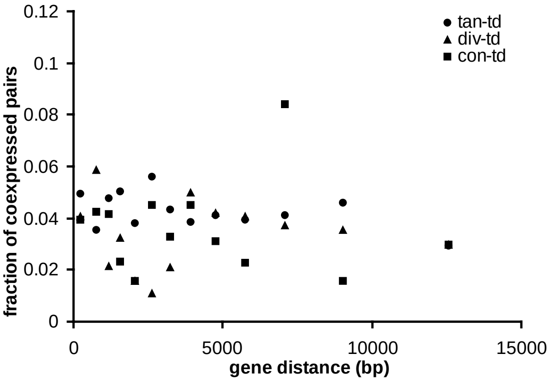


Figure 4.2 Gene distance does not solely explain the occurrence of coexpression. Gene distance, defined as the length in nucleotides from the annotated end of one gene to the annotated start of the next gene relative to the strand the genome that is given, with annotated start always smaller than the annotated end. X-axis is the averaged gene distance (in base pair) in each 1000-pair bin. Y-axis is the fraction of coexpressed pairs relative to the total number of pairs in each corresponding orientation (tan: tandem pairs; div: divergent pairs; con: convergent pairs) in each 1000-pair bin.

Functional categorization of coexpressed genes

To characterize the kind of genes that are present in the rice coexpression domains, the gene ontology (GO) developed for plants (GOslim; (Berardini et al., 2004)) was used. The GOslim ontology provides a controlled vocabulary to describe gene and gene product attributes in plants, focussing on three aspects of annotation: molecular function, biological process and cellular component. Each aspect has 15~16 categories with 4~5 categories having terms like “unknown” or “other”. To all pairs of genes, each aspect of the GOslim annotation was assigned. For each aspect, the number of pairs was determined for which both member genes were covered by a GOslim assignment. In addition, the number of pairs for which both member genes fall into the same well-defined categories (excluding the “unknown” and “other” subcategories) was determined. The fraction of the latter was compared between coexpressed pairs and non-coexpressed pairs to determine whether coexpressed pairs were enriched in the same categories (Table 4.4). The GOslim annotation coverage for both member genes in a pair is 22% for molecular function, 12% for biological process and only 3.4% for cellular component. Comparing these figures with the GOslim coverage of Arabidopsis genes ((Ren et al., 2005); using the TIGR5 update), which is ~94% for all 3 aspects, shows that the rice genome is currently considerably less well annotated than the Arabidopsis genome. When comparing the coexpressed and non-coexpressed pairs in rice for the fraction of gene pairs falling into the same well-defined GOslim catagory, there is no significant difference (Table 4.4). Therefore, coexpressed gene pairs are not enriched for the same functional category.

Table 4.4 Distribution of gene pairs over GOslim categories (Non-duplicated pairs)

	All[a]	Coexpressed[b]	Non-coexpressed[c]	P-val[d]
Rice 11,257 non-duplicated pairs in total				
GO_func				
Covered[e]	2502	100	2402	
SameKnCat[f]	365 (14.6%)	12 (12.0%)	353 (14.7%)	0.42
GO_proc				
Covered	1366	50	1316	
SameKnCat	144 (10.5%)	7 (14%)	137 (10.4%)	0.47
GO_comp				
Covered	383	17	366	
SameKnCat	113 (29.5%)	6 (35.3%)	107 (29.2%)	0.60

[a] Number of neighboring pairs excluding td in the analysis. Other kinds of pairs are all duplicates-free, if not otherwise mentioned

[b] Number of coexpressed pairs included in the analysis.

[c] Number of non-coexpressed pairs.

[d] P value from the standard normal tables of the z statistic for the difference of the population proportion between coexpressed pairs and non-coexpressed pairs in of the Arabidopsis genome; *, significant (two-tailed; P<0.01). P value here is the probability under the null hypothesis that the two population proportions are the same.

[e] Number of pairs of which both members are assigned (covered) with GOslim categories.

[f] Number of pairs of which both members fall into the same “known” GOslim category (excluding the categories with the indications ‘unknown’ and ‘other’). Percentage is the number of pairs relative to the number of pairs covered.

Microsynteny of local coexpression domains between rice and Arabidopsis

The structural characteristics of local coexpression domains in rice and in Arabidopsis (Ren et al., 2005) are remarkably similar. This prompts the question whether such domains also share functional characteristics and possibly consist of the same or related genes. Microsynteny in local expression domains of these two genomes would reflect conservation of such domains. The Inparanoid Eukaryotic Orthologous database (O'Brien et al., 2005) was used to retrieve the current list of genes that are supposed to be orthologous between Arabidopsis (14,753 genes) and rice (12,428 genes). The genes establishing coexpressed pairs in either Arabidopsis (data from (Ren et al., 2005); but updated to TIGR 5; 944 pairs including 116 duplicated pairs) or rice (584 pairs, including 146 duplicated pairs) were searched against these lists. This way, we aimed to identify the pairs of which both genes in the pair have an ortholog in the other plant and these orthologs are also coexpressed. The analyses showed that there was not a single coexpressed pair in either Arabidopsis or rice of which both genes are orthologous to a gene of a coexpressed pair in the other species. Therefore, given the current annotation of the two genomes, there are no syntenic local coexpresion domains between Arabidopsis and rice.

Partially syntenic local coexpression domains can occur by chance

In 34 cases though, one gene of a coexpressed pair in one plant was orthologous to at least one gene of a coexpressed pair in the other plant. That is 3.6% of all (944) coexpressed pairs in Arabidopsis and 5.8% of all (584) coexpressed pairs in rice. We will refer to such a case as a partially syntenic coexpression domain (PSCD). To assess the significance of such partially syntenic domains, we evaluated all the genes in non-coexpressed pairs, comparing Arabidopsis (15,629 pairs including 617 duplicated pairs) and rice (12,336 pairs including 1,517 duplicated pairs) to establish whether PSCDs are more enriched in the genome than non-coexpressed partially syntenic domains (PSD). We identified 4,488 PSDs (72 due to duplicated pairs) between all non-coexpressed pairs of genes in both plant genomes. This is 28.7% of all Arabidopsis non-coexpressed pairs and 36.4% of all rice non-coexpressed pairs. The percentages of PSDs among non-coexpressed pairs are 6-8 times higher than PSCDs from coexpressed pairs. Therefore, PSCDs do not seem to occur more often than expected by chance alone.

A complicating issue in the analysis of synteny is the occurrence of many-to-many orthologs. The Inparanoid database defines so-called inparalogs as paralogs arising through gene duplication after speciation. These can form a group of genes that together are orthologous to a gene in the other species. As a result, there can be many to many, many to one and one to one relationships. Individual member genes in many-to-many or many-to-one relationships may not be the main orthologs. Interestingly, there is one many-to-one case in which four Arabidopsis genes are all orthologs of the same single rice gene (Os07g43560.1). These 4 Arabidopsis genes are: At4g23140.2, At4g23150.1, At4g23230.1 and At4g23270.1. The first two, At4g23140.2, At4g23150.1, form a local coexpressed pair. The other two genes, At4g23230.1 and At4g23270.1, are not more than ten genes away from the previous two genes on the same chromosomal region. The latter two genes are separated from each other by a few genes. Further analysis shows that gene At4g23270.1 has a duplicated neighbor, At4g23280.1, but is not coexpressed with it. It is, however, coexpressed with its other neighbor At4g23260.1, but is not duplicated with it. Orthology is established between At4g23270.1 and the rice gene Os07g43560.1, but not between any of the neighbors of the Arabidopsis genes. The rice gene Os07g43560.1 is also coexpressed with one of its neighbour genes but is not duplicated with it, while this rice gene is duplicated with the other neighbor but not coexpressed with it. A schematic representation of the resulting gene configuration is given in Figure 4.3. Such detailed analyses may reveal local microsynteny in the twilight zone of statistical significance and evolutionary relevance.

Discussion

Local coexpression domains represent only a small part of the genome

Setting stringent criteria for coexpression, the rice genome was found to contain a significant number of local coexpression domains that range from 2-4 genes. This is similar to the situation in Arabidopsis (Ren et al., 2005) and agrees with other coexpression studies where strong coexpression was shown to occur only within close proximity of several genes (Cohen et al., 2000; Lercher et al., 2003; Hershberg et al., 2005; Semon and Duret, 2006). Although coexpression was shown to extend to regions covering up to 30 genes and more (Spellman and Rubin, 2002) and to cover a chromosomal up to 100 kb (Spellman and Rubin, 2002; Williams and Bowles, 2004) and more, there appears to exist a decrease in the strength of coexpression with increasing distance. The local coexpression domains described here represent ~4-5% of the potential coexpression fraction in the whole genome as found in other studies (Semon and Duret, 2006). Larger but looser coexpression domains might cover up to ~10% (Cohen et al., 2000; Williams and Bowles, 2004) – 20% (Spellman and Rubin, 2002) of the genome. The difference in occurrence between local and longer-range weaker but still statistically significant coexpression domains is highly dependent on the method used (Semon and Duret, 2006).

The terms cluster or chromosomal domain and associated terms such as neighboring are generally based on a (much) more loose definition compared to the definition used here to identify local coexpression domains. Local coexpression domains require a pairwise correlation between the expressions of all adjacent genes above 0.7. The larger domains are defined on the basis of the use of a sliding window of either a given sequence length (number of nucleotides) or of a given number of genes (Spellman and Rubin, 2002; Williams and Bowles, 2004). In such a window, the average correlation is calculated and compared with simulated sets. This allows the presence of genes within a domain that are not strongly (co)expressed but are “carried along for a ride” in the open chromatin domain (Spellman and Rubin, 2002).

Parameters shaping local coexpression domains

The existence of local coexpression domains in rice could not be explained solely by gene orientation, such as tandemly, divergently or convergently oriented gene pairs. No relative enrichment of the proportion of coexpressed pairs was seen. The fraction of coexpressed genes in the divergent orientation was even lower than for the other two orientations (Table 4.3 and Figure 4.1). So shared promoter regions (for divergent pairs) and transcriptional read-through (for tandem pairs) do not explain the local coexpression domains in rice,

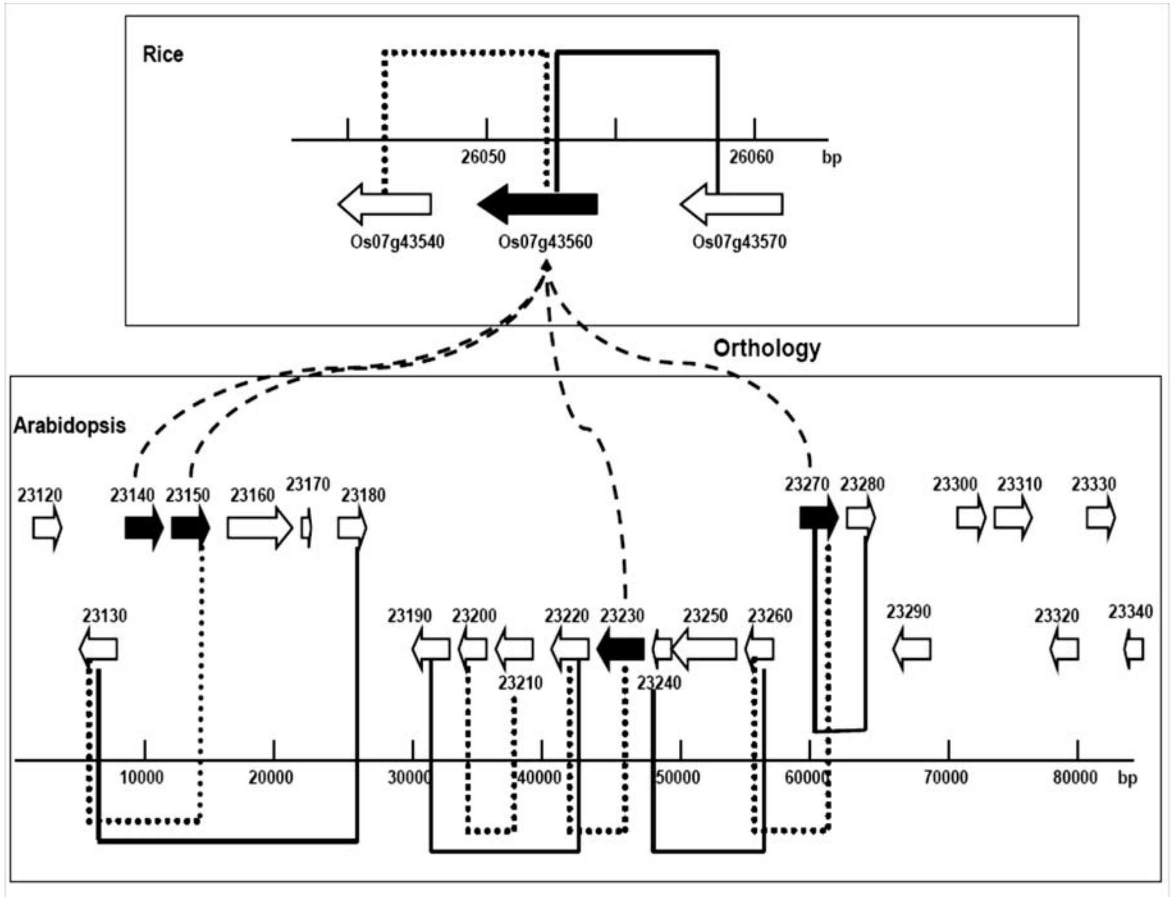


Figure 4.3 Schematic representation of the chromosomal regions covering genes involved in a four-to-one orthology between Arabidopsis and rice. Top part of the figure is the chromosomal region of rice (from gene locus Os07g43540 to gene locus Os07g43570) and bottom part of the figure is the chromosomal region of Arabidopsis representing 23 genes (from gene locus At4g23120 to At4g23340; the numbers in the picture skipped "At4g"). Black block arrows represent four Arabidopsis and one rice gene involved in this orthology, and dashed curved connecting lines show the orthology relationship. Black bracket-like lines depict duplication and the genes connected and included within are duplicated to each other. Dotted lines depict coexpression relationship and genes connected and included within are coexpressed with each other.

similar to what we have concluded for Arabidopsis (Ren et al., 2005). This is in contrast to some other studies in which shared promoter region (for divergent pairs) and transcriptional read-through established coexpression domains (Semon and Duret, 2006). Whereas we do not detect any preferred orientation to result in coexpression in rice, other studies show a higher degree of coexpression in divergent and tandemly oriented gene pairs (Williams and Bowles, 2004; Zhan et al., 2006). The differences in conclusions are most likely due to the different methods used, such as the definition of coexpression of neighboring genes as well as the dataset and/or expression platform used.

Gene distance is not the explanatory factor for the occurrence of local coexpression domains in rice. No significant decrease in the fraction of coexpressed genes was observed with increasing intergenic distance (Figure 4.2). The fraction of coexpressed pairs does not decrease even with gene distances up to 12kb. It shows that at a relatively large distance neighbouring genes can still be coexpressed. Another study reported that when genes >12kb apart were taken into account, the negative correlation between coexpression and gene distance was gone (Williams and Bowles, 2004). In a comparative study of 6 eukaryotic genomes, coexpression was shown to vary at chromosomal distances above 100kb (Fukuoka et al., 2004). This suggests that considerable coexpression of neighbouring genes can occur even at large gene distance, although the coexpression may not be related to the physical distance anymore. While gene distance itself is not predictive for coexpression (Cohen et al., 2000; Kruglyak and Tang, 2000), the likelihood of coexpression would favour short gene distances (Hurst et al., 2002; Lercher et al., 2003; Hershberg et al., 2005; Semon and Duret, 2006). However, it should be kept in mind that the rice data set now analysed is far from complete in terms of its annotation, so what are now far-apart neighbouring genes may be no longer directly neighbouring the moment the annotation is improved.

With the gene ontology developed for plants (GOslim), there is no evidence that coexpressed genes are more enriched in the same functional category in comparison to non-coexpressed genes (Table 4.4). Previous studies suggested that clustering of functionally related genes would occur in all metazoans (Cohen et al., 2000; Lercher et al., 2003). Recent studies demonstrated a significant enrichment for coexpressed genes in the same metabolic pathway (Williams and Bowles, 2004) or the same biological processes (Zhan et al., 2006), although this appeared not to be the explanation for the coexpression of neighboring genes (Williams and Bowles, 2004). In worm, clusters of similarly expressed genes cover similar biological functions (Roy et al., 2002). In human, coexpression over the whole genome was shown to correlate with functional relationships between the genes (Lee et al., 2004). Our study found no enrichment of coexpressed gene pairs in the same functional category than non-coexpressed pairs, suggesting that it is not necessarily true that the natural selection maintained regional coexpression by keeping genes with similar functions in adjacent positions (Cohen et al., 2000; Semon and Duret, 2006).

The genomic context of genes is supposed to play an important role in the regulation of gene expression (van Drunen et al., 1997). In a number of coexpression studies in varies organisms, the occurrence of coexpression domains, whether small (local) or larger (global), sometimes independent of gene orientation and gene distance, were all supposed to be regulated at the level of higher-order chromosomal structure (Cohen et al., 2000; Spellman and Rubin, 2002; Williams and Bowles, 2004; Hershberg et al., 2005; Ren et al., 2005; Zhan et al., 2006). Previous experience with transgene expression data indicated that the particular position of a single gene in a genome affects the expression of that gene considerably (Mlynarova et al., 1994; Mlynarova et al., 1995). Depending on the chromosomal context, two physically neighboring transgenes could be made to show correlated expression (Mlynarova et al., 2002). Our study of local coexpression domains in rice that are independent of duplication, gene orientation, or gene distance strengthen the notion that the regulation of genes in such domains resides at the level of higher-order chromatin structures.

Lack of microsyntenic coexpression

From an evolutionary point of view, syntenic regions between species reveal genes for conserved and important traits. Macrosynteny is generally not easily detectable after a long evolutionary time, as colinearity erodes by various mechanisms, such as transposon activity, intra or inter-chromosomal rearrangements, duplications, translocations, inversions and/or individual divergence after speciation (Salse et al., 2002). While macrosynteny may not be detectable any more for genomes that diverged more than 100 mya, microsynteny, i.e. conservation of local gene order and orientation, may still exist and be informative (Devos et al., 1999; Salse et al., 2002). Arabidopsis and rice are thought to have diverged about 120-200 million years ago (mya) (Salse et al., 2002). Microsyntenic local coexpression domains between Arabidopsis and rice would indicate the importance of the evolutionary conservation of regulatory systems beyond sequence similarity after the divergence of dicotyledonous and monocotyledonous plants. Analyses show that there is not a single coexpressed pair in either Arabidopsis or rice of which both genes are orthologous to a gene in a coexpressed pair in the other species. Therefore, there are no syntenic local coexpression domains between Arabidopsis and rice. Maintenance of coexpression has apparently not been an important driving force in evolution during or after the divergence of dicotyledonous and monocotyledonous plants. Although individual genes in local coexpression domains in either rice or Arabidopsis may have an ortholog in the other species, establishing so-called partially syntenic coexpression domains (PSCDs), this does not seem to occur above chance in the context of whole-genome configurations. Without statistical significance, the occurrence of such PSCDs is unlikely to have any evolutionary relevance on a genome-wide scale. Detailed analyses of individual cases and gene locations may suggest the occurrence of local microsynteny and point to chains of evolutionary events in which the conservation of coexpression could be involved. However, more detailed studies are required to assess the functional relevance, if any, of such genomic constitutions.

Material and Methods

Genome data

The rice (*Oryza sativa*) genome was downloaded from the website of The Institute of Genomic Research (TIGR; www.tigr.org/). The rice TIGR version 3 [Jan. 2005] annotation has 57,915 gene loci. In case of alternative splicing, the longest variant of the gene was used. The genes along each chromosome were sorted based on ascending start coordinates and were numbered consecutively. These rank numbers (rank ID) helped to eliminate any discontinuity in the unique Os gene identifiers of the annotated genes and facilitated analyzing physically adjacent genes. In case of overlapping gene loci, the smaller one of the overlapping genes was removed from the data set. This way the order of both gene and rank ID numbers was maintained.

Expression data

The expression data for rice was obtained from rice MPSS database (mpss.udel.edu/rice/). Only the unique MPSS tags (mapping to the genome only once) and those mapping to unique gene identifiers in TIGR v3 were used in our analyses. The expression values in libraries representing the same tissues under the same experimental conditions in different replicates (for example, 60 days mature leaves replicate A, 60 days mature leaves replicate B) were averaged. This way, 18 different libraries were generated, that cover expression in 9 tissues (callus, panicle, leaves, root, germinating seed and seedling meristem, ovary and stigma, pollen, stem) under different experimental treatments or in different developmental stages.

Identification of local coexpression domains

Pearson's correlation coefficient (R) was calculated between all adjacent pairs (duplets) of genes using the expression data from all 18 libraries. If R was higher than 0.7, the gene pair concerned was considered to be coexpressed. The value of R>0.7 is generally considered a rule-of-thumb threshold (see for example bbc.botany.utoronto.ca/affydb/BAR_instructions and is used in various analyses (Cohen et al., 2000; Lee et al., 2004; Ren et al., 2005). The number of coexpressed adjacent pairs was counted. To evaluate the statistical significance of these numbers, they were compared with the number of coexpressed pairs from 100 randomizations of the population of expressed genes using the cumulative binomial distribution (Cohen et al., 2000). Previous analyses indicated that more than 100 randomizations did not result in significant changes in the numbers obtained (Ren et al., 2005). In each round of randomization, non-adjacent pairs of genes were randomly selected with replacement from the list of expressed genes that have expressed neighbors till the same total number of pairs was obtained. Similarly, coexpressed adjacent triplets, quadruplets and pentaplets were identified as series of genes with consecutive IDs in which all possible (that is, (n!/(n-2)!)*2; where n is the number of genes involved) pair-wise R's should be above the cut-off of 0.7. The significance of results was evaluated with randomizations equivalent to the procedure used in case of duplets.

Duplicated genes

Duplicated genes were identified by local pair-wise protein BLAST (BLASTP 2.2.6 [Apr-09-2003]; (Altschul et al., 1997)), on all gene pairs in the rice genome. A gene pair was considered to be duplicated (dup) if BLASTP yielded an E-value < 0.2 (Lercher et al., 2003; Fukuoka et al., 2004; Williams and Bowles, 2004). To determine duplicated triplets, quadruplets and pentaplets, it was required that any pair of the genes concerned had a BLASTP E-value <0.2.

Analyses of gene orientation and gene distance

Adjacent gene pairs were separated into tandemly, divergently and convergently transcribed pairs according to their relative direction of transcription. The number of coexpressed pairs in each orientation group was expressed as percentage relative to the total number of adjacent pairs in that group. Random pairs were made by randomly picking two non-adjacent genes from the list of expressed genes represented in pairs, analyzed for their orientation and compared with the real genome using a variant of the two-sample t test for proportions for determining the significance of a difference between two population proportions (Ott and Longnecker, 2001). The test statistic is based on the z statistic from the normal distribution and is given by (p1–p2)/ √ (p1*(1-p1)/n1 + p2*(1-p2)/n2), with p1 and p2 the two sample proportions, n1 and n2 the two sample sizes, under the condition that n1*p1, n1*(1-p1), n2*p2 and n2*(1-p2) are all larger than 5. When | z | > 2.575, the two sample proportions are considered to be significantly different at the 1% level (P<0.01). The z value is converted to a p value using standard normal tables.

To determine the gene distance, the intergenic distance is used. This distance is defined as the length in nucleotides from the annotated end of one gene to the annotated start of the next gene, including the UTRs when known, otherwise the translation start and stop sites were taken. The data sets excluding the duplicated gene pairs were analyzed. For each data set, gene pairs were sorted based on gene distance from short to long and bins of 1000 pairs were taken and analyzed, excluding the last bin with less than 1000 pairs. The advantage of using equal pair bin is that it avoids unequal number of gene pairs in different distance categories. Per 1000-pair bin, gene distance was calculated as the average over all 1000 pairs. For each 1000-pair bin, the fraction of coexpressed pairs relative to the total number of pairs in each orientation group in each bin was calculated and plotted.

Functional categorization of genes

TAIR's GOslim, the Gene Ontology (GO) developed for plants (Berardini et al., 2004) was used to classify the genes present in local coexpression domains. The three aspects of GOslim, molecular function, biological process and cellular component, were analyzed in parallel. With Python scripts, the number of pairs of which both members could be classified in GOslim was determined, and the number of pairs of which both members fall into the same well-defined GOslim category was also determined. The GOslim categories of 'unknown' and 'other' were not included into well-defined categories, because they give less (or no) information about functional categorization. The percentage of coexpressed pairs falling into the same well-defined category was compared with that of non-coexpressed pairs to determine whether coexpressed genes are more enriched in the same functional category than non-coexpressed genes.

Assessing synteny between Arabidopsis and rice

The Inparanoid Eukaryotic Ortholog groups (O'Brien et al., 2005) (inparanoid.cgb.ki.se/) was used to download all known orthologous and inparalogous clusters between Arabidopsis and rice. Inparanoid defines inparalogs as paralogs that arose through gene duplication after speciation. Inparalogs can form a group of genes that together are orthologs to a gene in another species. There are 9,044 orthologous clusters between Arabidopsis (from Ensemble) and rice (from the Model Organism database) and all of them were taken into account. These clusters were downloaded on Dec. 12, 2005. In the orthologous clusters, 15,544 sequences (proteins) from Arabidopsis are inparalogs and 14,807 sequences (proteins) from rice are inparalogs. More than half of Arabidopsis and rice inparalogs are many-to-many or many to-one orthology cases. Less than a half of the cases are one-to-one orthology cases.

The Ensembl protein IDs (for Arabidopsis) and the Model organism database protein IDs (for rice) in Inparanoid were first translated to their unique gene identifiers in the respective TIGR annotation by BLASTP using an E-value < $1e^{-20}$. This yielded 14,753 unique Arabidopsis genes and 12,428 unique rice genes as inparalogs. The pairs of genes in local coexpression domains were analyzed to determine which genes in a rice local coexpressed pair have orthologs in an Arabidopsis local coexpressed pair, and *vice versa*. As coexpressed triplets and quadruplets are always combinations of coexpressed pairs, they were not further analyzed. For comparison, the pairs of genes that are not coexpressed were analyzed to determine how many non-coexpressed pairs, or one of their member genes, have orthologs in the other plant species. The numbers were then compared between coexpressed pairs and non-coexpressed pairs to determine the significance of occurrence of syntenic local coexpression domains.

Chapter 5

In plants, highly expressed genes are the least compact

Xin-Ying Ren[1, 2], Oscar Vorst[1], Mark W.E.J. Fiers[1], Willem J. Stiekema[2, 3] & Jan-Peter Nap[1, 3]

[1]Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, 6708 PB Wageningen, The Netherlands
[2]Laboratory of Bioinforamtics, Plant Sciences Group, Wageningen University and Research Center, 6703 HA Wageningen, the Netherlands
[3]Centre for BioSystems Genomics, 6700 AA Wageningen, The Netherlands

Published in *Trends in Genetics* 22(10): 528-532

In plants,
highly expressed genes are the least compact

Abstract

In both the monocot rice and the dicot Arabidopsis, highly expressed genes have more and longer introns, as well as a larger primary transcript than lowly expressed genes: higher expressed genes tend to be less compact than lower expressed genes. In animal genomes, it is the other way round. Although the length differences in plant genes are much smaller than in animals, these findings indicate that plant genes are in this respect different from animal genes. Explanations for the relationship between gene configuration and gene expression in animals may be (or may have been) less important in plants. We speculate that selection, if any, on genome configuration has taken a different turn after the divergence of plants and animals.

Introduction

A major issue in relating genome structure to gene expression is the relationship between the relative activity of genes and their position and/or structure. In organisms as diverse as human (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Vinogradov, 2004) and *Caenorhabditis elegans* (Castillo-Davis et al., 2002), highly expressed genes have less and shorter introns, shorter coding sequences, as well as shorter intergenic regions (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Vinogradov, 2004, , 2005). This compact nature of highly expressed genes is explained by a selection for either transcriptional efficiency to reduce time and energy (Castillo-Davis et al., 2002), a regional mutation bias that positions highly expressed genes in domains more prone to deletions (Urrutia and Hurst, 2003) or by a genomic design into open chromatin (Vinogradov, 2004). We here present a whole genome analysis of the relationship between gene structure and gene expression for two widely diverged plant species, the monocot rice (*Oryza sativa*) and the dicot *Arabidopsis thaliana* with data from two different expression platforms, MPSS and microarrays. In both plant genomes, highly expressed genes have more and longer introns, as well as a longer primary transcript. In short, they are less compact than the lowly expressed genes. This contrasts with the relationship between gene expression and gene structure in human and *C. elegans*, although the absolute differences between plant genes are considerably smaller than for human genes. These findings could suggest that the outcome of selection has been different between animals and plants.

Results

Analysis of plant gene expression in relationship to gene structure

The public domain Massively Parallel Signature Sequencing (MPSS) expression data for Arabidopsis (Meyers et al., 2004) (see mpss.udel.edu/at/) and rice (Nakano et al., 2006) (see mpss.udel.edu/rice/) offer a good genome-wide expression coverage in a range of different expression libraries and allow easy quantification. To correlate expression data with gene structure, we obtained Arabidopsis and rice genome sequences and annotations from The Institute of Genomic Research (TIGR). All genes annotated as either (retro)transposon or pseudo gene were excluded from the analysis and in case of alternative splicing, the longest variant was used in the analyses. We mapped the MPSS expression data to their position in the Arabidopsis (TIGR5) and rice (TIGR version 3) genome and all 17-b MPSS tags with a unique position were taken into account. Genes without expression data were not included in the analysis.

To compare the levels of expression of genes in different expression libraries, we sorted the expression values in each library in an ascending order, then divided them into 5 groups each containing 20% of the population and assigned an expression rank from 1 (lowly expressed) to 5 (highly expressed). In case the cutoff caused equal expression values to be in different rank groups (happening notably with zero expression), the expression values were placed in the lower rank group. For each gene, we averaged the expression ranks over all libraries. This averaged expression rank (rE) indicates the relative expression level of each gene under all conditions analyzed. Alternative methods of expression analysis (supplementary material) give similar results as found for rE. As the rE can be influenced in part by the number of libraries in which transcription is detected (the so-called breadth of expression) (Urrutia and Hurst, 2001), the analyses were also performed with the highest rank of the gene over all libraries, the peak expression rank (pE) (supplementary material).
We correlated the rE parameter with various structural characteristics of each gene, such as the number of introns per gene, the total length of the introns per gene, and others. The rE values were sorted in an ascending order and equal quantiles were taken from the two tails of the population. The top and bottom 1%, 5%, 10%, 20%, 30%, 40% and 50% quantiles were compared for the structural characteristic evaluated to avoid discussions over what genes should be considered ‘highly expressed’ and ‘lowly expressed’. For comparison, the data of the whole population (100%) is also given. The quantile comparisons for the parameters are shown in Figure 5.1 for Arabidopsis (Figure 5.1ace) and for rice (Figure 5.1bdf). The corresponding quantitative data for the 40% quantile, representing 80% of all genes analyzed, is given in Table 5.1.

The differences between the means and medians (Table 5.1) indicate that the various parameters are not normally distributed, that is why we used the non-parametric Mann-Whitney test for comparisons. Analyses of the logarithmically transformed gene parameters confirmed the conclusions (supplementary material). Both plant species have the same average number of introns per gene: 4.7 ± 0.04 s.e.m. (standard error of the mean). In both plants, and for each quantile analyzed, the higher expressed genes have significantly ($P<10^{-4}$) more introns than the lower expressed genes (Figure 5.1ab) and the total intron length per gene is significantly ($P<10^{-4}$) longer (Figure 5.1cd) in the higher expressed genes. In plants, therefore, highly expressed genes have not only more, but also longer introns than lowly expressed genes. Therefore, also the average intron length per gene is larger for highly expressed genes (Table 5.1).

Excluding the genes without introns, or removal of up to the first four introns, to correct for the tendency of introns to become smaller towards the 3'end (Seoighe et al., 2005), all confirmed the same relationship between expression and gene characteristics, as did the analysis based on pE (supplementary material). Therefore, the positive correlation between high expression and the number or length of introns is not due to the first introns only, nor can the correlation be an artifact of the averaged ranking (rE) over libraries.

To investigate the potential importance of transcription on expression, we analyzed the correlation between rE and the length of the primary transcript, including all introns and UTR sequences as annotated. For this structural parameter as for those analyzed previously, the highly expressed genes in both Arabidopsis (Figure 5.1c) and rice (Figure 5.1f) are significantly ($P<10^{-4}$) longer than the lowly expressed genes for all quantiles analyzed. All variations of the analyses described above, did not affect the results (supplementary material). Notably in current genome annotations, UTRs may be missing from the gene model. Limiting the analyses to all genes with both 5' and 3' UTR sequences given in their gene model again confirmed the results (supplementary material).

The length of the coding sequence per gene is larger in higher expressed genes than in lower expressed genes (Table 5.1), owing to the higher number of introns – and consequently also exons – in higher expressed genes, although the average exon length correlates negatively with expression level. Excluding all genes that have alternative splicing forms in their annotation, also did not affect the results, ruling out alternative splice variants as explanation (supplementary material). No positive correlation was found between high expression and either short introns or short flanking intergenic regions (supplementary material), whereas in human such a correlation motivated the regional mutation bias model (Urrutia and Hurst, 2003) and the genomic design model (Vinogradov, 2004). Similar analyses on plant expression data from a microarray platform (Birnbaum et al., 2003) used in other analysis (Ren et al., 2005) confirmed the above results (supplementary material). In both Arabidopsis and rice, highly expressed genes have larger primary transcripts with more and longer introns than lowly expressed genes. In these plants, higher expressed genes are, in other words, less compact than lower expressed genes.

Table 5.1

	Arabidopsis			Rice		
Expression level	Highly	Lowly	All	Highly	Lowly	All
No. genes included	n=7358	n=7358	n=18394	n=8572	n=8572	n=21431
Number of introns	5.5 ± 0.07[a] (4)[b]	3.8 ± 0.06 (2)	4.7 ± 0.04 (3)	5.9 ± 0.06 (4)	3.6 ± 0.05 (2)	4.7 ± 0.04 (3)
Average intron length per gene (bp)	164 ± 1.8 (133)	140 ± 2.1 (106)	152 ± 1.2 (120)	416 ± 4.4 (333)	359 ± 4.6 (250)	387 ± 2.9 (298)
Total intron length per gene (bp)	876 ± 11 (684)	603 ± 9.0 (367)	740 ± 6.3 (533)	2204 ± 23 (1818)	1405 ± 20 (816)	1805 ± 14 (1368)
Average exon length per gene (bp)	372 ± 5.0 (212)	479 ± 5.7 (293)	430 ± 3.5 (251)	329 ± 3.8 (203)	474 ± 5.4 (298)	405 ± 3.0 (244)
Total coding sequence length per gene (bp)	1396 ± 12 (1173)	1284 ± 9.9 (1113)	1350 ± 6.8 (1152)	1400 ± 11 (1164)	1251 ± 9.4 (1071)	1339 ± 6.6 (1128)
Length of primary transcript (bp)	2692 ± 20 (2313)	2105 ± 17 (1822)	2411 ± 12 (2082)	3988 ± 30 (3420)	2842 ± 25 (2277)	3435 ± 18 (2895)

[a]average ± standard error; [b]median
All genes that have a unique 17-b MPSS tag in at least one library and a protein translation in their annotation were taken into account. All parameters for higher expressed genes are significantly ($P < 10^{-4}$) different from lower expressed genes according to the z value approximation (www.texasoft.com/winkmann.html) of the non-parametric Mann-Whitney test for the comparison of two samples.

Discussion

Are animal genes different from plant genes?

In animals, highly expressed genes have smaller primary transcripts with less and smaller introns (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Vinogradov, 2004). The more compact nature of highly expressed animal genes is explained by transcriptional efficiency (Castillo-Davis et al., 2002), regional mutation bias (Urrutia and Hurst, 2003) or genomic design (Vinogradov, 2004). In plants, our data indicate that highly expressed genes tend to be significantly less compact than lowly expressed genes, although the absolute difference is much smaller than in animals. As highly expressed plant genes are not more compact than lowly expressed plant genes, there is no need to hypothesize the existence of selection for such compactness in high expression. Neither transcriptional efficiency, nor regional mutational bias or genomic design favoring open chromatin seems necessary, or appropriate, to explain the relationship between gene structure and gene expression in Arabidopsis and rice. Interestingly, in pollen-expressed genes of Arabidopsis, evidence for the efficiency hypothesis was documented (Seoighe et al., 2005). These results may indicate that expression in the male gametophyte of plants is more prone to selection on intron length than expression in the sporophyte (supplementary material).

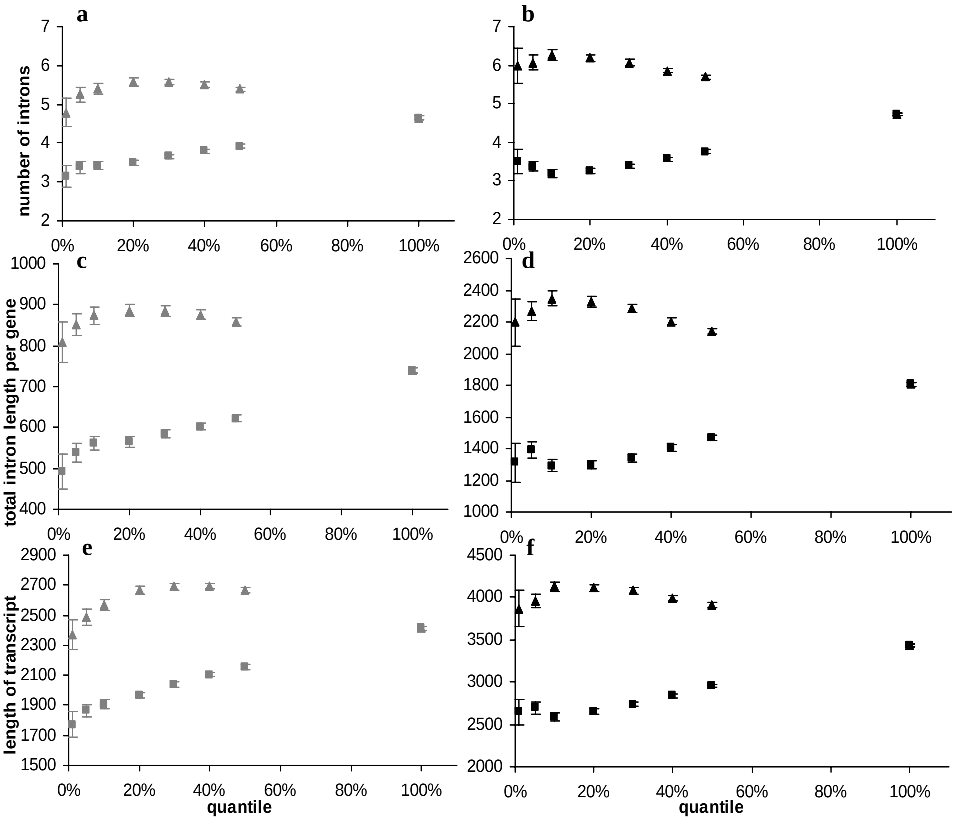


Figure 5.1 Relationships between the structural characteristics of plant genes and their expression. The panels show a structural parameter ± standard error of the mean versus the average expression rank (rE) in the 1%, 5%, 10%, 20%, 30%, 40% and 50% quantiles from both tails of the ranked population, as well as the value for the whole population (100%). Relationships are shown for three different parameters: the number of introns in Arabidopsis (**a**) and rice (**b**), the total intron length per gene (**c**: Arabidopsis; **d**: rice) and the length of primary transcript (**e**: Arabidopsis; **f**: rice). The rice data is indicated in black and the Arabidopsis data in grey. Higher expressed genes are plotted with a triangle and lower expressed genes with a square. In total, 18394 Arabidopsis genes with expression in 14 different libraries (callus, inflorescence, leaf, root, and silique under different experimental conditions and developmental stages) and 21431 rice genes with expression in 18 different libraries (callus, leaf, root, seed, panicle, meristem, pollen, stem, seedlings, ovary and stigma under different experimental conditions and developmental stages) were included in these analyses. Additional data are available in the supplementary material.

An important parameter to consider for the interpretation of these data is the relative length of introns per gene. The average intron length per gene in the human genome is about 5.5 kb (Sakharkar et al., 2004), which is considerably larger than the average intron length per gene in the plant genomes here analyzed (Arabidopsis: 152 b; rice: 387 b; Table 5.1). Human genes have on the average also more introns (7.7 introns per gene (Sakharkar et al., 2004)), so the total intron length per gene in human is about 42 kb, compared to 0.74 kb (1.8% of human) for Arabidopsis and 1.8 kb (4.3% of human) for rice. In contrast, the total exon length per gene in human (1.49 kb, with 8.7 exon per gene (Sakharkar et al., 2004)) is of the same order of magnitude as the total exon length per gene in Arabidopsis (1.35 kb, with 5.7 exon per gene; Table 5.1) or rice (1.34 kb, also with 5.7 exon per gene; Table 5.1). Therefore, in plant genomes, not all gene parameters are smaller than in the human genome, but it is the intron size per gene (either average or total) that is very different and makes the configuration of plant genes different from animal genes. More genomes will have to be analysed to show whether plant introns are under selection to stay relatively small or to become relatively small. The difference in total intron length between highly and lowly expressed genes in the 40% quantile class is about 273 b for Arabidopsis and 799 b for rice (Table 5.1). This is between about 11% (Arabidopsis) and 23% (rice) of the average length of the primary transcript of the genes (averaged over both classes). When the 10% quantile is considered, these figures go up to 14% for Arabidopsis and 31% for rice (data not shown). These differences in total intron length per gene are significant ($P<10^{-4}$), also when the first four introns are removed (supplementary material).

The hypothesis of selection for efficiency in pollen using Serial Analysis of Gene Expression (SAGE) data was based on a (significant) difference of 16 b per intron and about 140 b in total (Seoighe et al., 2005). Therefore, it seems reasonable to assume that the similar small differences here reported have also biological relevance. If so, they point to a different outcome of selection in plants and animals with respect to intron length and expression characteristics. It is feasible that the much larger differences in total intron length in the human genome cause the primary transcripts to be subject to other selective forces than the overall much smaller plant transcripts. Possibly, the difference in intron length between highly and lowly expressed genes in plants is not -or much less- relevant for a selection based on length. Introns are involved in a variety of regulatory phenomena such as RNA stability (Kirby et al., 1995; Shabalina and Spiridonov, 2004; Haddrill et al., 2005), post-transcriptional gene regulation (Liebhaber et al., 1992; Carlini et al., 2001; Shabalina and Spiridonov, 2004), nucleosome formation and chromatin organization (Zuckerkandl, 1997; Mattick and Gagen, 2001; Shabalina and Spiridonov, 2004; Vinogradov, 2005), and/or separating functional domains of proteins (Duester et al., 1986; Choi et al., 1991). Any or a combination of such phenomena could have shaped the structural configuration of highly expressed plant genes in comparison to lowly expressed plant genes. Possibly, in plants longer introns with regulatory roles were necessary to achieve high(er) expression. Such a regulatory role of plant introns may have favored additional selective forces to keep plant introns relatively small to reduce the likelihood of interruption by transposons. There could be a preferred intron length for high expression, whereas selection, if any, for low expression would have been different between human and plant.

Highly expressed genes in various yeasts and other unicellular organisms also have longer introns (Vinogradov, 2001). Although these analyses were based on relatively low numbers of genes, they also suggest a functional role for intron length in gene expression (Vinogradov, 2001). A recent study on the evolution of intron number in a set of orthologous genes showed that Arabidopsis and human were equally exceptional in intron gain over intron loss (Roy and Gilbert, 2005). Unfortunately, rice genes were not covered and neither intron length nor expression characteristics was considered. Our results show that it may be worthwhile to include intron length and expression characteristics in further studies on the evolution of eukaryotic gene structure. Whatever selection, if any, has been responsible for more and longer introns in highly expressed plant genes, it must have been selective forces that took a different turn after the split of plants and animals, some 1,600 million years ago (Sanderson et al., 2004).

Chapter 6

Comparative genomics of the relationship between gene structure and gene expression in a range of higher eukaryotes

Xin-Ying Ren[1, 2], Rudi Alberts[3], Willem J. Stiekema[2, 4], Ritsert C. Jansen[3] & Jan-Peter Nap[1, 3, 4]

[1]Applied Bioinformatics, Plant Research International, Wageningen University and Research Centre, 6708PB Wageningen, The Netherlands
[2]Laboratory of Bioinformatics, Plant Sciences Group, Wageningen University and Research Centre 6703HA Wageningen, The Netherlands
[3]Groningen Bioinformatics Centre, University of Groningen, 9751NN Haren, The Netherlands
[4]Centre for BioSystems Genomics, 6700AA Wageningen, The Netherlands

Manuscript in preparation

Comparative genomics of the relationship between gene structure and gene expression in a range of higher eukaryotes

Abstract

In various biological systems, higher expressed genes are reported to be the more compact in terms of introns and length. We here show that this general view of the relationship between gene structure and gene expression should be reconsidered. The relationships between gene structure and gene expression were analyzed for five genomes (Arabidopsis, rice, worm, mouse, human), using public domain MPSS and affymetrix microarray (for worm) expression data sets that cover a wide variety of tissues and conditions. Five different parameters of gene structure were examined with the help of rank-based methods: the number of introns, as well as the total length of introns, combined untranslated regions, coding sequence and the combined total length of the primary transcript. In addition, the broadness or breadth of expression is evaluated. The methods of analyses were identical for all genomes considered. It is found that tissue specific genes, defined as genes that are expressed in only one (or at most a few) tissues/conditions, are among the more compact genes in all genomes evaluated. Moreover, in plants the higher expressed genes tend to be longer and less compact than the lower expressed genes, whereas in the mammalian genomes analyzed the trend is the opposite. Worm takes an intermediate position. The different genomes differ markedly in the details of the relationship between expression and structure of the genes that are in the middle class of expression level. As the major difference in genome configuration is the absolute length of introns, possible explanations for the contrasting trends in plant and mammalian genomes question the role and evolutionary history of introns. Possibly there is a threshold amount of intron number and/or size upon which selection acts as to give different outputs between genomes. Or some groups of plant introns have possibly been introduced in plant genomes well after the split between animals and plants.

Introduction

In various organisms, higher expressed genes have shorter introns, shorter intergenic regions, as well as code for shorter proteins (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Comeron, 2004; Vinogradov, 2004, , 2005). This more compact nature of higher expressed genes is explained by either a selection for transcriptional efficiency to reduce time and energy (Castillo-Davis et al., 2002), a regional mutation bias that positions highly expressed genes in domains more prone to deletions (Urrutia and Hurst, 2003) or a genomic design into open chromatin (Vinogradov, 2004). In plants, selection for short distal introns was observed in genes higher expressed in haploid pollen (Seoighe et al., 2005). However, we have recently shown that genes higher expressed in diploid somatic plant cells are not more compact than other genes (Ren et al., 2006), suggesting major differences between plants and animals in a fundamental aspect of genome organization.

Unfortunately, there is not yet a uniformed or standardized way of defining higher and lower gene expression on a genome-wide scale. The various studies have used different structural parameters for a gene or have defined structural parameters in different ways, for example either averaging over the whole genome, over subgroups or per gene. Also, studies were based on vastly different genome representations in terms of numbers of genes and/or annotation. These differences hamper comparisons and keep the possibility open that differences between genomes, if any, reflect the methods of analyses rather than meaningful differences in gene or genome organization. To be able to compare the relationships between gene expression and gene structure in different genomes, a more comparative way of data analysis is needed.

In this study, we gathered and analyzed whole genome and expression data for five eukaryotes that span a wide taxonomic range, including two plants, one invertebrate and two vertebrates: Arabidopsis (*Arabidopsis thaliana*), rice (*Oryza sativa*), worm (*Caenorhabditis elegans*), mouse (*Mus musculus*) and human (*Homo sapiens*). This range of organisms is thought to represent a trend of increasing biological complexity, if only from the point of view of theoretical information definition (Taft and Mattick, 2003). We have analyzed the relationships between gene structure and gene expression of these five genomes using the same methodology and five parameters to define gene structure. The structural parameters chosen for analysis are three parameters of length that are defined per gene: (1) the total length of all introns, (2) the total length of the coding sequence (CDS) and (3) the total length of the combined (5’ plus 3’) untranslated regions (UTRs). These three parameters allow assessing gene structure in relationship to expression primarily from the point of view of primary transcriptional costs (Castillo-Davis et al., 2002). The total length of the primary transcript is included in the analysis as a fourth parameter, although it is the sum of the previous three parameters and therefore not independent. The fifth parameter examined is the number of introns per gene. Combined, these parameters allow comparing the length and compactness of genes without formally defining ‘compactness’ mathematically. Alternative splicing, however important in biological systems, is not taken into account: the longest transcript known is taken for analysis. Expression data is, as much as possible, taken from public domain repositories with Massively Parallel Signature Sequencing (MPSS) data. Only for worm, microarray data are included.

Due to the way the sequence tag-based expression is generated, MPSS data are directly comparable across different genomes and datasets. The MPSS technology has no hybridization issues, the transcripts measured are not known to be pre-selected and the technology is sensitive to lowly expressed genes (Coughlan et al., 2004). MPSS will only miss the few genes that do not have the recognition site of the restriction enzyme used in the procedure. Current MPSS data already offers large coverage of biological tissues. Moreover, MPSS data permits relatively easy quantification.

With these data, we analyzed the distribution of expression levels by rank-based methods to reduce the influence of potential outliers and to be able to compare within and between genomes. In addition, we analyzed the broadness or breadth of expression, defined as the number of different tissues in which a given gene is expressed. This analysis allows determining the relationships between gene structure and for example tissue specificity of expression. This markedly extends our earlier findings (Ren et al., 2006) that plant and animal genomes differ in their relationships between structural characteristics of genes and their expression. The possible reasons for the existence and characteristics of such differences are discussed.

Results

Descriptive and comparative statistics of five genomes and their gene structures

To be able to compare the relationships between gene structure and gene expression among different genomes, we present the descriptive statistics and comparisons for the five genomes (Table 6.1) to highlight the overall similarities and differences. The latest genome sequence and annotations for Arabidopsis, rice, worm, mouse and human were downloaded from their appropriate source and were combined with public domain expression data. We calculated the whole-genome averages of the five structural parameters of genes defined above and we compared these using all expressed genes. The MPSS expression coverage of the human genome (27%) is somewhat lower than the expression data coverage for the other four organisms, but simulations show that this coverage is high enough to represent the whole-genome trend for the subsequent analyses (data not shown).

The data in Table 6.1 indicate that the non-protein coding part of genes makes a dramatic difference in gene structure between plants and vertebrates. Genes in mouse and human carry on average twice as many introns per gene as genes in plants. Especially the size of vertebrate introns is considerably larger, up to 59-fold larger when compared to Arabidopsis. In addition, the vertebrate UTR is up to 5-fold longer. Worm genes tend to be closer to plant genes than to animal genes, with the exception of a remarkably (2-fold) shorter UTR. In contrast, the average length of the protein-coding part of genes (CDS) is within the same range of ~1300 to ~1700 bp for all five genomes (Table 6.1). The notable increase in noncoding sequences, especially in the total intron length per gene, from Arabidopsis and rice to mouse and human appears to reflect the transition from plant to animal. The much smaller differences in the length of the CDS among these eukaryotic organisms is thought to reflect the much more stringent selective forces on protein coding sequences (Rubin et al., 2000; Liu and Rost, 2001; Zhuang et al., 2003; Brocchieri and Karlin, 2005). With a similar CDS length and a lower number of introns (therefore also less exons) in plant genes than animal genes, the average exon length of plant genes is longer than that of animal genes. Therefore, not all the structural parameters of genes are more compact in plants than in animals.

Table 6.1 Descriptive and comparative statistics of the five genomes studied

Organism	Arabidopsis	rice	worm	mouse	human
[a] Genome size (Mb)	119	377	100	2,848	3,108
(+ratio)	(1)	(3.16)	(0.84)	(23.9)	(26.1)
[b] Total number of gene loci	28,952	57,915	23,254	19,774	25,108
[c] Number of genes	18,394	21,431	17,751	10,998	6,680
(+%)	(64%)	(37%)	(76%)	(56%)	(27%)
[d] Average number of introns (±SE)	4.7 ± 0.04	4.7 ± 0.04	5.4 ± 0.03	10.2 ± 0.09	9.6 ± 0.11
(+ratio)	(1)	(1.02)	(1.16)	(2.17)	(2.04)
[e] Average total UTR length (±SE)	268 ± 1.6	290 ± 2.6	115 ± 1.4	1,248 ± 10	1,372 ± 14
(+ratio)	(1)	(1.08)	(0.43)	(4.66)	(5.12)
[f] Average CDS length (±SE)	1,350 ± 6.8	1,339 ± 6.6	1,311 ± 9.6	1,730 ± 15	1,705 ± 19
(+ratio)	(1)	(0.99)	(0.97)	(1.28)	(1.26)
[g] Average total intron length (±SE)	740 ± 6.3	1,805 ± 14	1,624 ± 22	41,510 ± 782	43,746 ± 1,000
(+ratio)	(1)	(2.44)	(2.19)	(56.1)	(59.1)
[h]Av. primary transcript length (±SE)	2,411 ± 12	3,435 ± 18	3,050 ± 27	44,488 ± 788	46,822 ± 1,008
(+ratio)	(1)	(1.42)	(1.27)	(18.5)	(19.4)
(+ % intron sequence)	(26%)	(42%)	(41%)	(79%)	(79%)

[a] Genome size in megabases (Mb). In brackets is the ratio relative to the genome size of Arabidopsis.

[b] Total number of gene loci in each genome according to the annotation data used.

[c] Number of genes with expression data in the expression set used. In brackets is the percentage of expressed genes relative to the total number of gene loci.

[d] The genome average of the number of introns per gene over all genes with expression data ± standard error of the mean (SE). In brackets is the ratio relative to the equivalent parameter of the Arabidopsis genome.

[e] The genome average of the total UTR length per gene over all genes with expression data ± SE. In brackets is the ratio relative to the equivalent parameter of the Arabidopsis genome.

[f] The genome average of CDS length per gene over all genes with expression data ± SE. In brackets is the ratio relative to the parameter of the Arabidopsis genome.

[g] The genome average of the total intron length per gene over all genes with expression data ± SE. In brackets is the ratio relative to the equivalent parameter of the Arabidopsis genome.

[h] The genome average of the primary transcript length per gene over all genes with expression data ± SE. In the first brackets is the ratio relative to the equivalent parameter of the Arabidopsis genome. In the second brackets is the whole-genome percentage of the fraction of the total intron length per gene as fraction of the primary transcript length per gene over all genes with expression data.

Breadth of expression and gene structure differ between organisms

The distribution of the expression of a gene over a range of tissues gives information about the relationships between gene expression and structure. The broadness or breadth of expression is defined as the number of tissues in which a gene is expressed. This expression parameter was used to demonstrate the relative compactness of genes in relationship to the presumed energy costs of transcription and translation (Eisenberg and Levanon, 2003; Vinogradov, 2004). The breadth of expression only considers whether the gene is expressed or not and does not take the actual level of expression into account. Based on this expression parameter, genes can be classified as tissue-specific (TS) genes when they are expressed in only one tissue or developmental stage, or as tissue-specific-like (TS-like) genes when they are expressed in a few tissues or developmental stages. Likewise, housekeeping (HK) genes are expressed in all tissues and developmental stages and housekeeping-like (HK-like) genes are expressed in most tissues or developmental stages. What constitutes ‘a few’ or ‘most’ in these definitions often depends on the dataset and number of tissues considered and seems largely a matter of choice. Genes were grouped according to the number of tissues they were expressed in, using an MPSS tag count > 0 and a threshold of $\log_{10}$(expression) > 1.6 for the microarray data. For each group of genes, the average of all five structural parameters considered was calculated (see Materials and Methods for details). In Figure 6.1, the averages of the structural parameters are plotted against the breadth of expression. Within each sub-panel, the TS genes (one tissue) are on the left and the HK genes (all tissues) are on the right.

These results show that both in plants and in worm, TS and TS-like genes are more compact than the HK and HK-like genes in the same genome. The TS and TS-like genes have less introns, shorter UTRs, shorter CDS lengths, shorter total intron lengths and shorter primary transcripts. The situation in human is the opposite: TS and TS-like genes have longer primary transcripts and longer total intron lengths than HK and HK-like genes, although the former have more introns and longer UTRs than the latter. In mouse, HK genes are overall as compact as TS genes, although the genes with intermediate breadth of expression show notable differences with either extreme.

Plant and animal genomes differ in the relationship between expression and structure

The breadth of expression does not consider the level of expression, whereas the actual level may reveal more about the relationship between expression and structure. We first combined expression with annotation using the double ranking approach developed earlier (Ren et al., 2006) . The average (arithmetic mean) of the grouped ranks over all tissues is defined as the rank of expression (rE). The whole genome is divided into 10 quantiles based on sorted rE from low to high. Ranking based on the geometrical mean of ranks gave the same result (see Supplementary Material). In Figure 6.2, the relationships between the expression level and the five structural parameters considered are shown. As reported earlier (Ren et al., 2006), higher expressed genes in both Arabidopsis and rice have more introns, longer UTRs, longer total intron length per gene and, consequently, a longer length of the primary transcript than lower expressed genes. This is not the case in the vertebrate genomes analyzed with the same methodology. Although the highest expressed vertebrate genes (10th quantile) have more introns and longer UTRs than the lowest expressed genes (1st quantile), they have much shorter total intron length and, as a consequence, a much shorter primary transcript (Figure 6.2). In contrast to plant genes, the higher expressed genes in mouse and human are more compact than the lower expressed genes, when considering the length of primary transcript. The genes in worm take an intermediate position: higher expressed genes in worm are as compact as lower expressed genes (Figure 6.2).

To analyze the five structural parameters between the five genomes, we compared the values for the 1st, 5th and 10th expression quantiles in a pairwise manner. These values represent the low, medium and high expression groups, respectively. The results are presented in Table 6.2. They show that in all five genomes evaluated, the genes in the medium expression group are less compact (larger) than the genes in the low expression group: all ratios in the column labeled 5th/1st are larger than 1 (Table 6.2). In plants and worm, the high expression group is also less compact (larger) than the low group. In contrast, the high expression group in vertebrates is significantly more compact (shorter) than the medium group (Table 6.2, values marked with a grey box in the column labeled 10th/5th).

Detailed comparison of five structural parameters in five genomes

For further analyses with a larger coverage of the genes considered, we will refer to the quantiles 1-3 as the lower expressing gene class, the quantiles 4-7 as the middle expressing gene class and the quantiles 8-10 as the higher expressing gene class. It should be noted, however, that this division is arbitrary. In both plants and vertebrates, higher expressing genes have more introns than middle and lower expressing genes, with a Pearson correlation coefficient (R) over all quantiles well above 0.9 (Arabidopsis: R=0.94; rice: R=0.99; mouse: R=0.91 and human: R=0.90). Similar results are found when intron density per kilobase of CDS length is taken as parameter (see Supplementary Material). In worm, the higher expressing genes have less introns than all other genes. In all five genomes, the total length of the UTR also increases with higher gene expression, with correlation coefficients around 0.9 (Arabidopsis, rice, worm and human), whereas the situation in mouse is less straightforward (R=0.53). Also the relationships between CDS length and expression level

Table 6.2 Pair-wise comparison of structural parameters of genes for the 1st, 5th and 10th expression quantile

Organism	Parameter	Quantile comparison		
		5th/1st [a]	10th/1st [b]	10th/5th [c]
Arabidopsis	No.of introns	1.16 [d]	1.43	1.24
	Total UTRs	1.55	2.27	1.47
	Total CDS	1.11	*0.99* [e]	0.89 [f]
	Total intron length	1.17	1.46	1.24
	Length of transcript	1.17	1.25	1.07
Rice	No.of introns	1.36	2.02	1.48
	Total UTRs	2.08	3.08	1.48
	Total CDS	1.12	1.14	1.02
	Total intron length	1.23	1.81	1.47
	Length of transcript	1.23	1.57	1.29
Worm	No.of introns	1.16	*0.95*	0.81
	Total UTRs	1.88	3.88	2.06
	Total CDS	1.17	1.17	1.00
	Total intron length	1.57	*0.96*	0.61
	Length of transcript	1.38	1.13	0.82
Mouse	No.of introns	1.45	1.55	*1.07*
	Total UTRs	1.23	*1.10*	*0.89*
	Total CDS	1.23	1.11	*0.90*
	Total intron length	1.28	0.71	0.55
	Length of transcript	1.28	0.73	0.57
Human	No.of introns	*1.04*	1.23	1.18
	Total UTRs	1.19	1.27	1.07
	Total CDS	*1.06*	*1.02*	0.97
	Total intron length	*1.02*	0.60	0.58
	Length of transcript	*1.03*	0.63	0.61

[a] The ratio of gene parameters in the 5th expression quantile to the gene parameters in the 1st quantile

[b] The ratio of gene parameters in the 10th expression quantile to the gene parameters in the 1st quantile

[c] The ratio of gene parameters in the 10th expression quantile to the gene parameters in the 5th quantile

[d] Values in normal style indicate the gene parameter in the higher expression quantile is significantly larger than the gene parameter in the lower expression quantile. A p value < 10^{-4} was taken as threshold for significance and no corrections for multiple testing were applied. The non-parametric Mann-Whitney test (with large sample size z value approximation, www.texasoft.com/ winkmann.html) was used for pair-wise comparison.

[e] Values in Italics indicate there is no significant ($p>10^{-4}$) difference between the gene parameters in the two quantiles compared.

[f] Values marked in grey indicate the gene parameter in the higher expression quantile is significantly ($p < 10^{-4}$) smaller than the gene parameter in the lower expression quantile.

are similar across all five genomes.

The CDS length increases from the lower expressing genes to the middle expressing genes, but decreases in the higher expressing genes. As a result, there is a similar CDS length in both higher expressing and lower expressing genes in Arabidopsis, mouse and human. In contrast, there is still a longer CDS length in the higher expressing genes compared to lower expressing genes in rice and worm (Figure 6.2). These genome-wide data show that higher expressing genes do not code for smaller proteins than lower expressing genes, in contrast to previous suggestions (Urrutia and Hurst, 2003; Comeron, 2004).

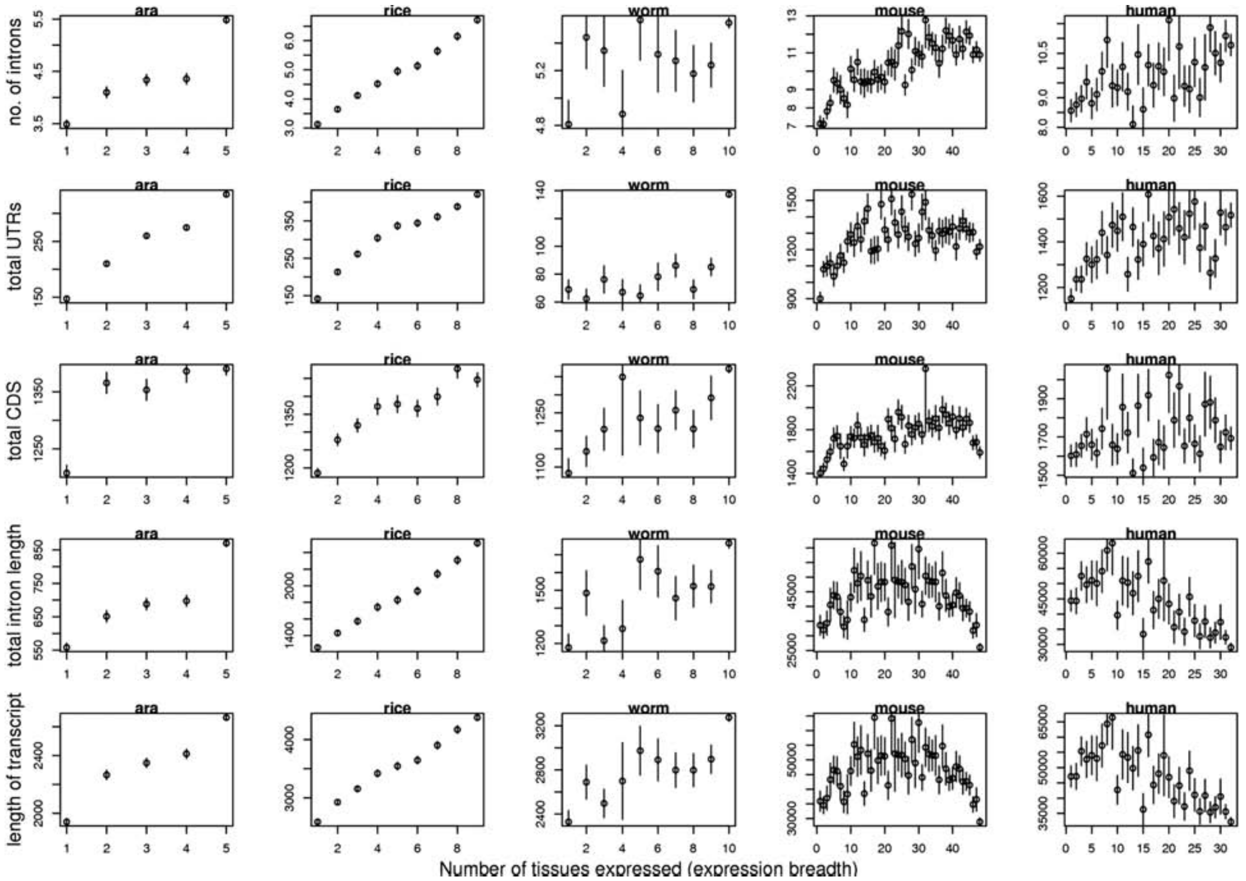


Figure 6.1 The relationship between breadth of gene expression and gene structure. The averages of the five gene structural parameters in each expression breadth were plotted against the number of tissues in each organism. The X-axis depicts the number of tissues and the Y-axis depicts the number of introns (1st row), total UTRs (2nd row), total CDS length (3rd row), total intron length (4th row) and the length of the primary transcript (last row). The averages of gene parameters are shown in rounds and the se (standard error of the mean) in vertical lines.

In the plant genomes, the total intron length per gene increases as gene expression gets higher (Arabidopsis: R=0.95; rice: R=0.98), but in the other three genomes the behavior of intron length is more complex. Intron length tends to increase from the lower expressing to the middle expressing genes. After that, there is a clear decrease, yet the decrease occurs at different quantile groups. In worm, the decrease in intron length starts from the 8th expression quantile. In mouse, it starts from the 6th quantile and in human the decrease starts as early as from the 4th quantile. Notably the middle expressing genes behave differently in the various genomes.

Because the length of the primary transcript is a direct result of the length of the introns (Table 6.1), all trends found for introns also apply to the primary transcript (Figure 6.2). Therefore, the three classes of organisms differ considerably in the way their genomes have dealt with the relationships between gene structure and gene expression during evolution.

Impact of the absence of expression

In the ranking method used to generate the results depicted in Figure 6.2, all expression data are given equal weight. However, absence of expression of a gene in a given tissue could be considered less informative, because possibly due to many reasons, than actual expression of that gene.

This issue is related to the parameter known as 'peak expression', in which the highest expression level of a gene is taken irrespective of the tissue or developmental stage, to establish the relationship with the structural parameters of a gene (Urrutia and Hurst, 2003). To investigate the relative role of the absence of gene expression, an alternative way of ranking was employed. The geometric mean of the expression ranks is only taken from the tissues in which the gene is expressed (MPSS tag count > 0 for the MPSS data; $\log_{10}$(expression)>1.6 for microarray data). Each dataset is subsequently sorted based on this rank that we indicate as geo_exrE (see Material and Methods), and divided into 10 quantiles from low to high expression. Figure 6.3 shows the relationships between expression based on geo_exrE and structure for the same five structural parameters as used above. Due to the alternative way the ranks are calculated, genes belong to different quantile groups. This is mainly the case for the lower and middle expressing genes that show tissue specific expression so absence of expression in the majority of tissues (see Supplementary Material). The graphs depicted in Figure 6.3 show that the differences between the plant and animal genomes are further emphasized when absence of expression is taken into account. In plant genomes, genes tend to become larger and less compact with increasing expression, while genes in animal genomes show a largely negative correlation between expression and size or compactness.

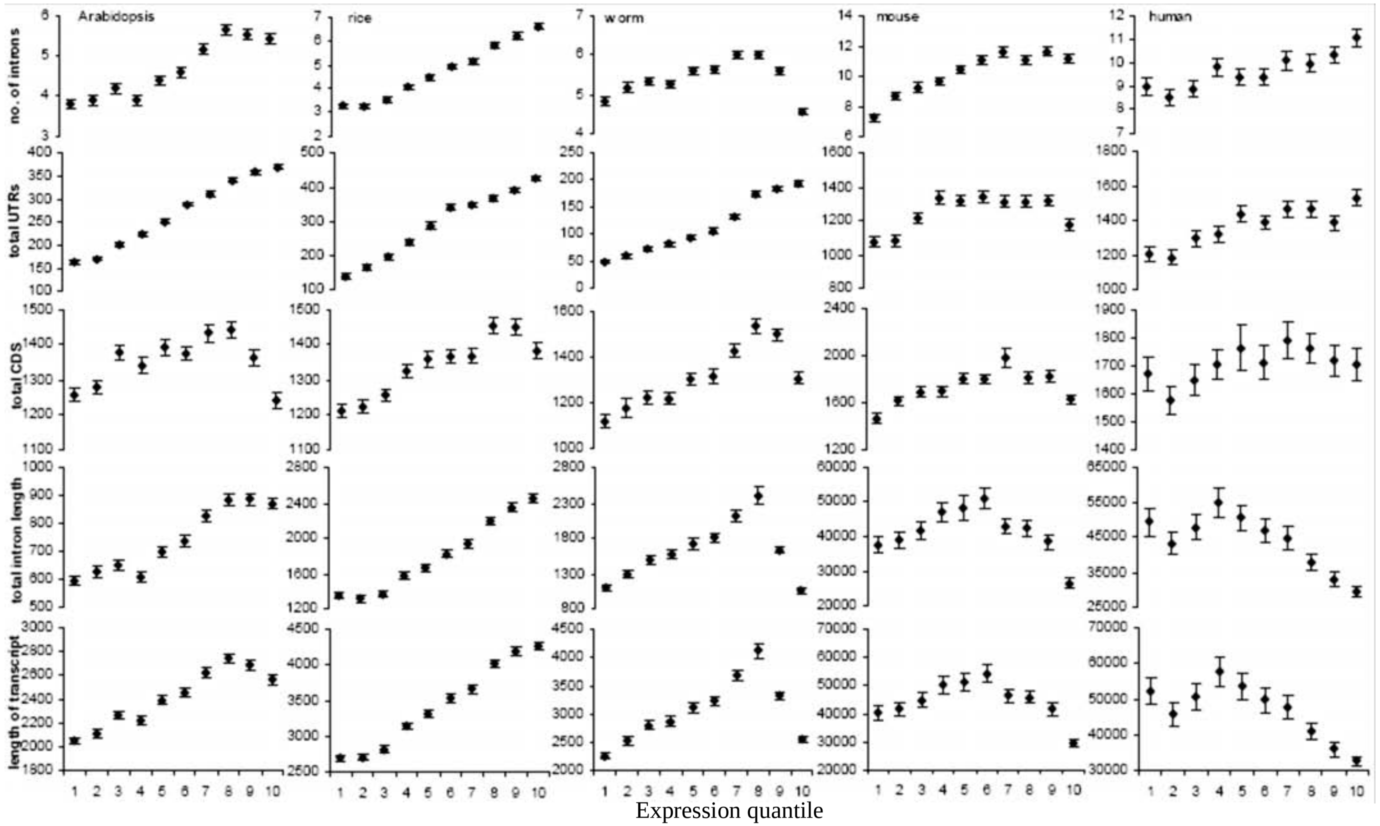


Figure 6.2 The relationship between gene expression level and gene structure. The averages of the five structural parameters of genes (5 rows) in each expression quantile based on the arithmetic means of ranks (rE) were plotted against the 10 expression quantiles for each organism. The X-axis depicts the expression from low (1: 1st quantile) to high (10: 10th quantile). The Y-axis depicts the number of introns (1st row), total UTRs (2nd row), total CDS length (3rd row), total intron length (4th row) and the length of transcript (last row). From left to right, first column subfigures are for Arabidopsis, 2nd column for rice, 3rd for worm, 4th for mouse and 5th column is for human. There are 1,839, 2,143, 1,775, 1,100 and 668 genes in each quantile in Arabidopsis, rice, worm, mouse and human respectively.

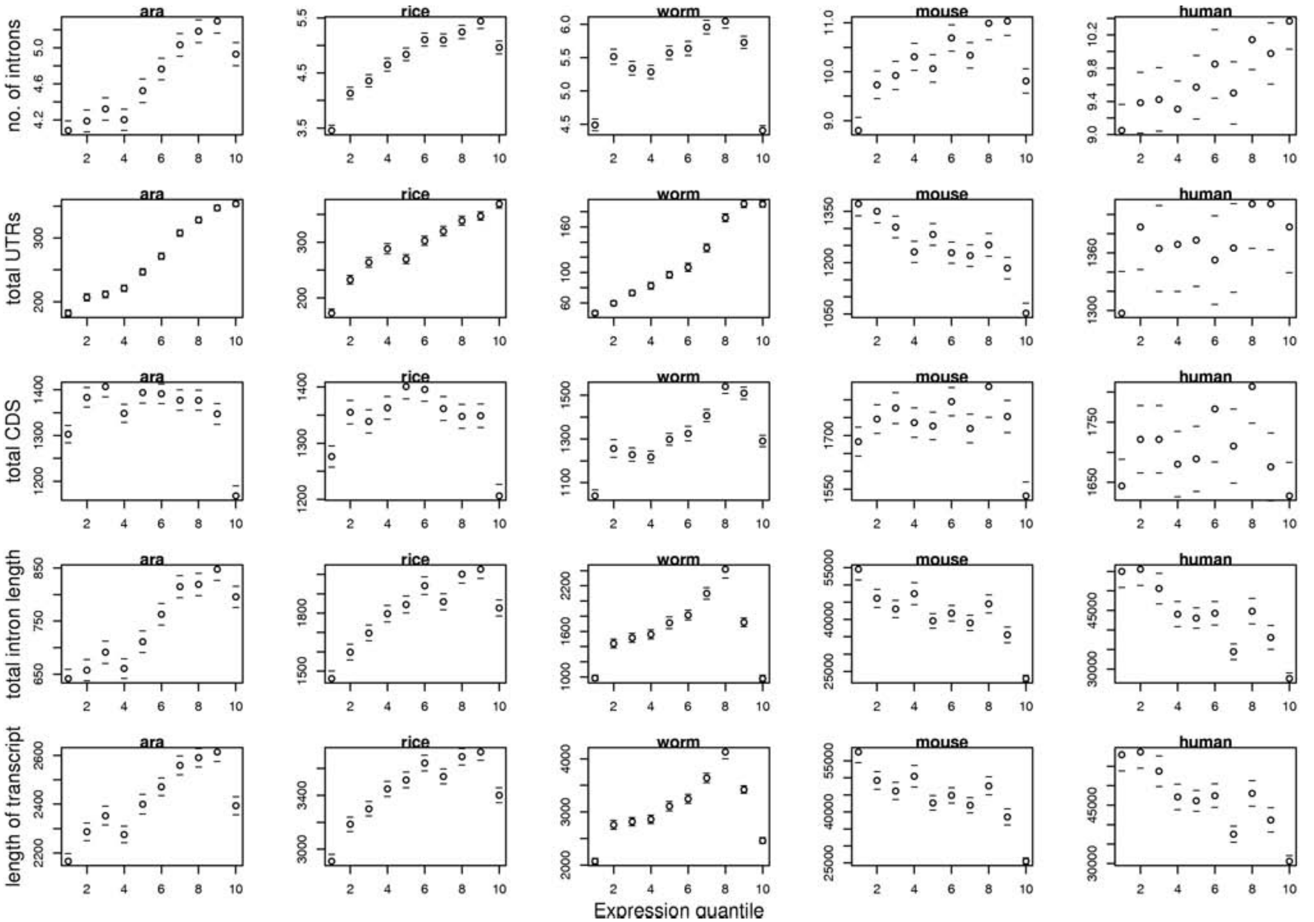


Figure 6.3 The relationship between gene expression level and gene structure relationship. The averages of the five structural parameters (5 rows) in each expression quantile based on the geometric means of ranks excluding no expression (**geo_exrE**) were plotted against the 10 quantiles for each organism. The X-axis and the Y-axis are as in Figure 6.2. The error bars are the standard error of the mean (SE) of the various structural parameters.

Discussion

Using a comparative genomics approach with comparable annotation and expression data, we have analyzed the whole-genome relationships that exist between gene expression and five parameters for gene structure. Five genomes were included in the analyses: Arabidopsis, rice (plants), worm (invertebrate), mouse and human (vertebrates). This analysis that had previously demonstrated that in plants the higher expressing genes are the least compact (Ren et al., 2006), reveals a remarkable and hitherto unknown aspect of genome configuration: plant genomes differ dramatically from animal genomes in their relationships between gene structure and gene expression. For all genomes, the same approach and the same definitions of structural parameters were used in this study. Therefore, the differences observed cannot be due to methodological issues or statistical analyses or precise definitions. Plant genes show a largely positive correlation between expression and structure: they become larger and less compact with increasing expression. In contrast, animal genes become smaller and compacter with increasing expression levels (Figure 6.3). Largely similar trends are present when all zero values are taken into account (Figure 6.2). Worm takes an intermediate position in almost all of the parameters studied.

In addition, the relationship between gene structure and expression is far from straightforward when expression breadth is considered (Figure 6.1). House-keeping and housekeeping-like genes are only the most compact in the human genome, whereas in plant and worm they are among the more bulky genes (Figure 6.1). Both the mouse and the human genome have a remarkably complex pattern when also the middle expressing class of genes is considered (Figure 6.1). The gene coverage of the human genome based on public domain MPSS data is relatively small (27%; Table 6.1). Simulation studies with random subsets of the same size selected from the other genome data indicate that such a coverage is sufficient to reveal the genome-wide trends (data not shown), assuming that the available human MPSS data were not biased in any -yet unknown- methodological way.

The data presented here give a markedly different perspective on the relationships between gene structure and gene expression than put forward in previous literature which reported that in a wide variety of organisms, higher expressed genes have shorter introns and code for shorter proteins (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Comeron, 2004; Vinogradov, 2004, , 2005). For example, the earlier suggestion that highly expressed genes are under selection for lower intron density (Castillo-Davis et al., 2002) is not supported by the genome-wide data and analyses presented here. Some differences could be attributed to the much wider genome coverage presented here, or to differences in the definition of parameters chosen for analysis. For example, the parameter ‘average intron length per gene’ was used to show that highly expressed genes in animals are compact (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003), whereas we here show that the parameter 'total intron length per gene' gives a quite different impression of 'genomic truth'. In this particular case, total intron length per gene would seem a much better parameter to assess the time and energy costs involved in primary transcription and the compactness of genes. Even with a shorter 'average intron length per gene', genes can (and in case of high expressing genes, they do) contain more introns per gene. More discussions and more work are clearly required to define the more informative parameters for describing the structural characteristics of genes that should be correlated with expression characteristics.

If it is not true that higher expressing genes are more compact than lower expressing genes, all attempts to explain why higher expressing genes would be more compact than lower expressing genes loose value. The more compact nature of highly expressed genes was explained by either a selection for transcriptional efficiency to reduce time and energy (Castillo-Davis et al., 2002), a regional mutation bias that positions highly expressed genes in domains more prone to deletions (Urrutia and Hurst, 2003) or a genomic design into open chromatin (Vinogradov, 2004). Notably the transcriptional efficiency hypothesis is intuitively attractive: why would a genome invest so much energy in transcribing sequences that are essentially spliced out and degraded without apparent role or function. There is a growing body of literature that seems to contradict the efficiency hypothesis in a similar way as the data here presented do. In *C. elegans,* a positive correlation between gene expression level and CDS length was taken to suggest that longer proteins may lower the selective pressure on codon usage (Duret and Mouchiroud, 1999). Intron density in higher expressed genes may be higher to be able to generate more alternative splicing forms (Comeron, 2004) or actively promote RNA export (Le Hir et al., 2003). Longer introns in highly expressed genes may harbor essential regulatory elements (Shabalina and Spiridonov, 2004; Haddrill et al., 2005) or help to stabilize the pre-mRNA secondary structure (Kirby et al., 1995; Haddrill et al., 2005). It seems plausible that in certain conditions a longer, less compact gene size could have added value to reach an optimized (but not necessarily high) expression to the extent that selection favors the longer gene despite the higher energy costs of transcription.

This consideration does not address the main difference in genomic configuration between plants and animals: why is it, that plant genomes have less frequent and so dramatically much smaller introns than mammalian genomes and could this be the reason that the relationships between gene structure and gene expression are so different between plants and animals? Plants have vastly different genome sizes (Wendel et al., 2002). Among all known groups of organisms, they are the most diverse in terms of genome sizes. The data on rice and Arabidopsis indicate that there may be a relationship between genome size and intron size (Wendel et al., 2002). With the current state of knowledge, it should be considered highly unlikely that a plant with a considerably larger genome size (such as lily or Fritillaria) will have a gene structure (for example, in terms of number and length of introns) and a relationship between gene structure and gene expression that resembles the data here presented for the human genome. More data on more (and more varied) plant genomes would, however, be highly advantageous for future evolutionary considerations about the why and when of intron size and distribution in plants.

Although the analyses presented here were performed in an identical way for all five genomes, the results are the overall averages for all genes. A major issue in the interpretation of the relationships between gene structure and expression is the extent of selection during evolution. How much degree of freedom is allowed in intron number and intron size before there is going to be a positive or negative selection that forces gene structure or expression into a given direction? And if there is such a selection, is it the same for all genes in all circumstances? The latter seems not very plausible. Different parts of a genome, or different types of genes in a genome, may exhibit different characteristics, as is already the case for tissue specific genes compared to housekeeping genes. Also, different introns may have different roles. Regulatory sequences, for example, are thought to reside mainly in the proximal introns (Shabalina and Spiridonov, 2004; Seoighe et al., 2005). Although a first analysis ignoring the proximal introns did not change the relationships between gene structure and gene expression in plants (Ren et al., 2006), more detailed analyses of genomes and genes may reveal less obvious trends and relationships between structure and expression.

Notwithstanding the above reservations, let us assume that selective forces in evolution are directly reflected in the current genome configuration and the current relationships between gene structure and gene expression. In plants, higher expressing genes are the largest. Plant genes may be under positive selection forces to increase their gene sizes in order to reach high expression. In animal genes, higher expressing genes are less compact than middle expressing genes, which could suggest a negative selection forces to reduce gene sizes in reaching higher expression. In worm, the higher expressing genes have lower number of introns. Possibly, there has been selection for lower intron density, or there has been a higher loss of introns during evolution (Raible et al., 2005; Roy, 2006). The structural parameter that gives the cleanest differences among the species analyzed is the total length of introns per gene. If any selection forces have shaped the difference between the plant and the mammalian genome, it should have acted on total intron length. Therefore, there could have been selective reasons for introns to either grow in size or reduce in size. Possibly, there is a certain amount of either intron number and/or total sequence length of introns upon which selection can act. This amount, that could be different in different genomes, could be positively correlated with the size and number of introns. If intron size is large, but the gene is not often transcribed, this combination of factors may not trigger selection as to reduce intron size. On the contrary, if intron size is small, but the the frequency of transcription is high, it may trigger selection to reduce intron size. More studies are required to test the possible presence of such a threshold amount by analysing many more genomes for the relationships between intron characteristics and gene expression.

The situation with respect to intron size, intron number and the relationship(s) between gene structure and gene expression is obviously not known in the earliest common ancestor of plants and animals. Introns, irrespective from their precise time (early or late) of arrival in genomes, may have seen at least three different evolutionary scenarios: (1) the number and size of introns in the common ancestral cell was small and plants have kept the small numbers, whereas mammalian cells have undergone a dramatic increase; (2) the number and size of introns in the common ancestral cell was intermediate; plants have reduced the numbers, whereas mammalian cells increased it; (3) the number and size of introns in the common ancestral cell was vast and mammalian genomes have maintained this situation, whereas plant genomes have undergone a dramatic reduction. These scenarios all assume that introns have appeared only once in evolution. Another possibility is that the large differences in intron characteristics and relationships between gene structure and gene expression between plants and animals could be taken as suggestion that introns have been introduced more than once, notably after the split between animals and plants. A careful analysis of intron characteristics, sizes and positions should be carried out in order to investigate the possibility that (part of) current day plant introns are different from animal introns because some groups of plant introns have possibly been introduced considerably later in evolution.

Materials and methods

Genome data

Plant genomes were retrieved from the website of The Institute of Genomic Research (TIGR; www.tigr.org) The Arabidopsis genome (*Arabidopsis thaliana*; TIGR5, Jan. 2004) has 28,952 gene loci; the rice genome (*Oryza sativa*; TIGR v3, Dec. 2004/Jan. 2005) has 57,915 gene loci. The other genomes were downloaded from the genome website of the University of California Santa Cruz (UCSC; genome.ucsc.edu). Gene structure data of all available RefSeq genes were used. The worm genome (*Caenorhabditis elegans*; UCSC version ce2, Mar. 2004) has 23,254 genes, the mouse genome (*Mus musculus*; UCSC version mm7, Aug. 2005) has 19,774 genes and the human genome (*Homo sapiens*; UCSC version hg18, Mar. 2006) has 25,108 genes. All genes annotated as either (retro)transposon or pseudogene were excluded from the analyses. In case of documented alternative splicing, the longest variant was used. Overlapping gene loci were not included in the analysis in order to avoid the ambiguity of assigning expression data.

Expression data

The public domain Massively Parallel Signature Sequencing (MPSS) expression data were obtained from the Delaware Biotechnology Institute (DBI; mpss.udel.edu/) for Arabidopsis (mpss.udel.edu/at/; (Meyers et al., 2004)) and rice (mpss.udel.edu/rice/). For mouse, the MPSS mouse transcriptome data deposited in Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) with accession number GSE1581 were used. This is the only dataset used that is based on 20-b MPSS tags and the other three were based on 17-b tags. For human, GEO accession GSE1747, containing the transcriptome data of 32 normal (non-diseased) human tissues (Jongeneel et al., 2005) was used. Only the MPSS tags that mapped to the genome once were used in the analyses. Mapping of the unique MPSS tags to their respective genome resulted in a coverage of 18,394 (64% of the whole genome) Arabidopsis genes, 21,431 (37%) rice genes, 10,998 (56%) mouse genes and 6680 (27%) human genes expressed in at least one of the tissues available. Genes without expression data in all the tissues were not included in the analyses. Currently there is no public available MPSS expression dataset for worm. We used the *C. elegans* embryonic time course expression profile (Baugh et al., 2005) deposited in the GEO database with number GSE2180. Data from wild-type embryonic development (GSM39513 to GSM39546) were used. These profiling data cover ten successive developmental stages after embryo formation based on Affymerix whole genome microarrays. Combining microarray expression data with the latest genome annotation yielded 17,751 (76%) worm genes for further analyses. Data in libraries representing the same tissue or developmental stage were averaged, giving 5 tissues (from 14 libraries) for Arabidopsis, 9 tissues (from 18 libraries) for rice, 48 tissues (from 87 libraries) for mouse, 32 tissues for human and 10 stages for worm. The composition of the various tissue samples is given in the Supplementary Material.

Structural parameters of genes

The five structural parameters considered per gene are the number of introns and four length parameters: the total length of the untranslated regions (UTRs), the total intron length, the total length of the coding sequence (CDS) and the total length of primary transcript, which is the sum of the length of the UTRs, introns and CDS. Compactness of genes is intuitively defined as genes having less and shorter introns, shorter UTRs, shorter CDS and consequently a shorter primary transcript. The length of primary transcript can be considered to reflect the simplest measure of compactness. In this study, ‘transcript’ always refers to the primary transcript and every structural parameter is on a per gene basis, unless specified otherwise.

Analysis of expression data

The broadness or breadth of expression is defined as the number of tissues in which a gene is expressed. Expression was considered as having MPSS tag count >0 for the MPSS dataset and $\log_{10}$(expression)>1.6 for the worm microarray data. The latter value was chosen because microarray data do not give ‘zero’ expression and 80% of the whole-genome expression data is above this expression threshold value. The ranking method developed earlier (Ren et al., 2006) was used to define higher and lower expression. In short, the expression values of all the genes in each tissue or developmental stage were sorted, subsequently divided into 5 groups each containing 20% of the population and assigned a grouped rank from 1 (lower expressed) to 5 (higher expressed). Expression values smaller than and equal to a 20% division point were placed in the same rank group. For each gene, the grouped ranks over all tissues were averaged arithmetically. This averaged **r**ank of

expression (**rE**) indicates the relative expression level of a gene over all tissues (or developmental stages). The dataset was sorted by **rE** from low to high and was subsequently divided into 10 sequential quantiles (that is 0-10%, 10-20%, 20-30% of the population and so on) from low expression to high expression. The highest expressed genes are the genes in the 10th quantile and the lowest expressed genes are in the 1st quantile. To asses the relative influence of actual expression, an alternative ranking method was introduced. The expression values of all genes in each tissue (or developmental stage) was sorted and assigned a consecutive rank from 1 (lowest expressed) to the total number of genes (highest expressed) while dealing with tied groups. For each gene, the geometric mean of the ranks was calculated over the tissues in which this gene is expressed, using the same definition for expression as in the calculation of the breadth of expression: the number of MPSS counts > 0 was above zero, or $\log_{10}$(expression)>1.6 for the microarray data. This results in the 'geometric expressed rank of expression', indicated with **geo_exrE**. The datasets were sorted by geo_exrE and subsequently divided into 10 sequential quantiles as above. Additional weighting of the structural parameters on the basis of number of tissues or stages expressed was found not to affect the conclusions (data not shown).

Chapter 7

General discussion

General discussion

The research presented in this thesis has focused on two related aspects of gene structure and gene expression in genomes: the regulation of coexpression of physically neighboring genes and the relationship between gene structure and gene expression. Genomic position and gene structure are long considered to be of special importance in the regulation of gene expression (Wilson et al., 1990). Numerous studies have subsequently shown or deduced that higher order chromatin configurations play a decisive role in gene regulation (chapter 2). Yet, how chromatin decides on gene regulation is still largely unknown. The research as presented allowed us to deduce that chromatin organization is involved in the regulation of expression of neighboring genes in the genome of both rice and Arabidopsis.

A major topic in chromatin organization is the existence of domains, but various research groups use and approach the concept ‘domain’ in -sometimes subtly- different ways. The novel concept of ‘local coexpression domain’ was introduced and defined as the coexpression of physically neighboring genes. This strict definition of local expression domain contrasts with the much more loose definitions of chromatin domains used in previous studies (Spellman and Rubin, 2002; Williams and Bowles, 2004). The concept of local coexpression domain was motivated by prior results in a transgenic setup, showing that neighboring (trans)genes could exhibit correlated expression when embedded in an artificial chromatin domain created with the help of chromatin boundaries (Mlynarova et al., 2002). A small yet significant fraction of the genome of Arabidopsis and rice consists of local expression domains. Various explanations for the existence of local coexpression domains, such as shared promoter sequences, could be excluded. It was concluded that the coexpression in such local domains is regulated on the level of higher-order chromatin organization (chapter 3 and chapter 4). The gene pairs in local expression domains are for the major part not involved in the same functional category, so joint function was not a driving force for the existence or maintenance of local expression domains. The lack of microsynteny of genes in such domains between Arabidopsis and rice confirms that maintenance of coexpression has apparently not been an important driving force in evolution.

The criteria that establish a local coexpression domain as here defined are very stringent. Such domains may therefore coincide with so-called ‘strong’ domains. In such domains, the presence of distinct boundary elements, such as matrix-associated regions (MARs), is thought to isolate the domain from its surroundings (Dillon, 2006). Various software packages exist to predict such chromatin boundaries (Singh et al., 1997; van Drunen et al., 1999; Glazko et al., 2001; Frisch et al., 2002), but this software tends to identify too large numbers of apparently false positives in plant genomes (data not shown). It will therefore be worthwhile to examine in detail the DNA sequences surrounding these local coexpression domains. The surrounding sequences may consist of actual plant boundary elements that can have applications in transgenic approaches and can help to further define the characteristics of boundary elements in plants. The coexpression analyses have so far focused on a high positive correlation (R>0.7). The analysis of high negative or anti-correlation (R<-0.7) may help to identify genes that are separated by boundaries, insulators or enhancer blocks (Dillon, 2006). Future studies considering the anti-coexpression of neighboring genes are recommended to further advance the knowledge about these aspects of genome organization. In the analyses as presented, data from different expression libraries were combined to evaluate the occurrence of local coexpression domains under averaged conditions and cell types. Further studies on subsets of data and expression libraries are likely to reveal an appreciable dynamics of local coexpression of neighboring genes. It is feasible that there are time- and/or tissue specific local expression domains as result of expression differences between different organs, tissues or cells under different developmental stages. Such studies would contribute to the better understanding of the dynamics of chromatin structure in gene regulation (Dillon, 2006; Luger, 2006; Tremethick, 2006).

The second issue addressed by the research presented in this thesis is the relationship between the structure of a gene and its expression characteristics. This issue has also received a lot of attention in the scientific literature. In different biological systems, higher expressed genes were reported to be more compact (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Comeron, 2004; Vinogradov, 2004, , 2005). Various explanations have been put forward to explain this observation. One of these explanations involves a control on the level of higher chromatin organization (Vinogradov, 2004). However, the transcriptional efficiency hypothesis (Castillo-Davis et al., 2002) seems the most intuitive in a world that is centered on time, money and efficiency. Transcription apparently costs quite some time and a lot of (metabolic) energy. Why waste assumingly precious time and energy to transcribe DNA that is not necessary to make proteins any way? The research presented in this thesis (chapter 5) has shown that in the genomes of the plants *Arabidopsis thaliana* (thale cress) and *Oryza sativa* (rice) the relationship between expression and gene structure is not what is predicted by the transcriptional efficiency hypothesis. In these plants, the higher expressed genes are the least compact. In a more detailed comparative genomics approach (chapter 6), different genomes (human, mouse and worm in addition to rice and Arabidopsis) were compared for the relationships between gene expression and gene structure. We could exclude that the differences between plant and animal genomes were in some way due to methodological differences in analyses. These analyses confirmed and stressed the apparent differences between plant and animal genomes. If there has been any evolutionary selection on this particular aspect of genome configuration, the selection must have been different in animal genomes and plant genomes. Moreover, tissue specific genes, defined as genes that are expressed in only one (or a few) tissues or developmental stages, are among the more compact genes in all genomes evaluated.

Further studies are needed to pinpoint why the differences in gene structure and expression between different kingdoms occur. For this, plant and animal introns need more detailed analyses. Adding the dimension of gene expression could contribute to a better understanding of the role and origin of intron sequences. A major difference between the genomes of animals and plants is the absolute size of genes. This is notably different because of the larger number and considerably larger size of introns in mammalian genomes. Differences in gene structure between animal and plant genes are basically due to differences in the number and size of introns. As a result of this difference, the genes in plant genomes may not have seen selection for length, or the relative longer length was necessary for the regulation into high expression. In contrast, in animal genes the absolute longer length could have been sufficient for regulation and then absolute length became an issue in selection. Possibly, there is a threshold that decides on intron size or number upon which selection starts to act. If so, this threshold differs between genomes. What parameters detect and decide on such a threshold is not clear. Although there has been a lot of discussion and debate about when and why introns appeared in eukaryote genomes (Koonin, 2006), the quantitative aspect of the number of introns and the length of introns in different genomes has not received consideration. It would seem to be worthwhile to incorporate this quantitative aspect of intron and genome organization in thinking about introns, gene structure and the relationship between gene regulation and gene structure.

Plants (*Viridaeplantae*) differ in their genome size more than any other group of organisms (Wendel et al., 2002). Based on current data, it seems unlikely that plants with a much larger genome size than Arabidopsis or rice will have gene structures, notably intron sizes, that are comparable to mammalian genes. This, however, needs to be analyzed. More insight into the relationship between genome size and gene structure in different plant species would be necessary to establish the role of genome size in gene structure and expression relationships. To assess the evolutionary importance of the differences in the relationship between structure and expression here established between animals, plants (and possibly other genomes), more comparisons are required. It is recommended to analyze many more genes or genomes from a more diverse array of taxonomic groups. Genome and expression data of birds (*e.g.* chicken), insects (*e.g.* fruit fly), fish (*e.g.* fugu), amphibians (*e.g.* frog), oomycetes (e.g. *Phytophthora*) and fungi will be required to better define the parameters shaping gene structure and gene expression in evolution.

Although there has been a lot of attention for the relationships between gene organization and expression in the literature, we have concluded that the various results as reported are not easily comparable. The various research groups use different definitions of structural parameters for genes, different methods of analyses and different expression platforms. This situation is somewhat similar to the rather vague connotation of the concept of chromatin expression domain discussed above. Moreover, the studies tend to cover very different numbers of genes, so different representations of the genomes investigated. The structure of a gene is here defined on the basis of four length parameters, that is, the length of (1) the combined untranslated regions, (2) the introns, (3) the coding sequence and (4) the total primary transcript, although the latter obviously depends on the previous three. In addition, the total number of introns is taken into account. In future studies, more clarity about and consensus on the structural parameters used for analysis will be essential. Possibly, more parameters related to gene structure can be informative and will have to be taken into account. Likely candidates for such parameters are the GC content of introns, exons or surrounding regions as well as the regulatory signals in and around genes, such as MAR elements and transcription factor binding sites.

A similar need for clear definitions and more consensus is required for the method of analysis of expression data. The novel double ranking methods here developed seem appropriate and unbiased, but different variants are still conceivable. How to establish ‘the best’ method is currently not obvious, neither from a mathematical/statistical perspective, nor from a biological/functional perspective. We have shown that the relationship between gene structure and gene expression depends, among many other things, on how non-expressing genes are treated (chapter 6). It could be argued that absence of expression of a gene gives information that is of another (lesser) level than an expression level itself: expression is the result of many structural and regulatory factors coming together and the hypothesis is that only one needs missing to result in the absence of expression. Other ways of weighting expression values in the analyses are conceivable. More attention is necessary to investigate the influence of the method of analysis. It should be decided what is the best compromise between a statistical sound analysis, computational expenses and biological relevance.

The new results presented in this thesis should be considered to represent only the beginning of investigations into the relationships between gene structure and gene expression. The analyses are performed on whole-genome averages and combined expression data. Both data types could be subdivided and/or classified in groups. Given the variation between genes within a genome, regions in genomes may differ in their structure - expression relationships in a way that would detail genome configuration further. The existence of expression ridges (Caron et al., 2001) indicates indeed the existence of a higher order genome configuration that is ‘averaged away’ in our studies. Also, the parameters within a gene could be scrutinized. It is assumed that the most 5’ introns have more regulatory functions than the more distal introns, so analyses could distinguish intron positions. In plants, selection for shorter introns was only observed in genes higher expressed in haploid pollen when the proximal introns were discarded (Seoighe et al., 2005). Such intron skipping seems too arbitrary to be very informative. In addition, for simplicity reasons we have excluded overlapping genes and genes with alternative splicing. Notably the latter phenomenon deserves more attention in the type of analyses here presented. Moreover, if it is possible to distinguish ‘old’ genes or introns from ‘new’ genes and introns based on sequence conservation (Koonin, 2006), it would be very interesting indeed to compare the structural parameters such classes of genes within and between genomes.

Similar detailing is recommended for the treatment of the expression data. In the research here presented, expression data sets that cover a wide variety of tissues and conditions were combined. We have averaged expression data over presumably ‘similar’ tissues or cells, but also that is essentially arbitrary. Averaging over data from all tissues and stages available may give a good impression of the general, genome-wide trends and relationships, but may ignore equally interesting local phenomena. How to compare in a biologically meaningful way gene expression between systems as different as plant and animals that have different tissues and cell types in a biologically meaningful way is obviously a matter of debate. However, it should be pointed out that any expression data is generally already a combination of the expression of different cells. The relative contribution of a particular cell type in a given sample is rarely considered. In this way, expression data is inherently ‘averaging away’ potentially biologically informative differences between individual cells. On the other hand, biological systems seem to sustain quite a lot of -assumedly stochastic-variation in gene expression that should be considered ‘noise’ in the type of analyses presented here. Moreover, the expression analyses in this thesis have focused on expression on the RNA level, whereas the correlation between RNA level and protein (and subsequently phenotype) is not always very straightforward. Transcription rate and half-life of the mature RNA could be important parameters of regulation. Small regulatory RNAs (miRNAs) are not included in the expression data here considered. More work is clearly needed to separate the wheat from the chaff in expression data. More advanced statistical methods such as principal component analysis and its various derivatives could be helpful to identify the more decisive elements in either global or local structural or expression parameters that influence gene expression.

The bioinformatics research in this thesis is based on the current annotation of the genomes analyzed. Such annotations are incomplete, may contain errors and will change over time (Fiers, 2006). For example, genes that are now annotated as physically neighboring may not be so anymore when another gene is discovered in-between. We have made extensive use of the annotated starts and stops of genes to define the unit of transcription that was subsequently related to expression. In view of the latest data on genomes and transcription, this may not have been the best approach. Whole genome tiling arrays show that there is much more transcription of DNA than previously thought. This challenges the notion of a gene as a discrete unit of transcription (Pearson, 2006). The concept of a gene started in genetics as a locus defining a phenotype. Possibly whole genome expression arrays show us the way back to that more abstract notion of 'a gene'. If the genome is a continuum of transcripts, discrete genes may not exist (Pearson, 2006) and also 'non-coding DNA' may loose its descriptive value. As result of the now fuzzy concept of 'a gene', the starts and stops of genes that we have used in this study may be less important for gene expression than assumed. Widespread transcription could indicate that there is much more regulation on the RNA level. For example, the data accumulating on microRNAs is indicating that the role and importance of RNA in gene expression needs reassessment. The widespread transcription could also indicate that it is apparently advantageous to overtranscribe and then dispose of the non-coding and/or non-regulatory parts, than to invest in an organization that only produces what is necessary (Pearson, 2006). This seems an argument against the transcriptional efficiency hypothesis, in agreement with parts of the research presented in this thesis. However, such genomic overtranscription does not explain very well why different genes, or different parts of the genome, would result in different amounts of protein-coding RNA. Whereas in theory this could be accomplished by the regulated breakdown of an excess of genomic transcripts, convincing data to support such a model need yet to be presented in the literature. The model also needs to explain how the required parts of RNA are distinguished from the rubbish. If chromatin and higher order chromatin structures are not involved in the question to transcribe or not to transcribe, chromatin may be involved in deciding how much of the transcripts are processed and retained. In this, there may be an efficiency step after all. A detailed and more quantitative assessment of the expression data from whole genome tiling arrays would be necessary to resolve such issues.

At the end of the day, the understanding of genomic organization and use should contribute to the better understanding and possible use of the relationship between genome and organismal phenotype. The research presented in this thesis is part of the growing body of scientific evidence that the genetic material in eukaryotes that does not end up in protein is no junk. It has characteristics and functions that we are only beginning to appreciate.

Summary

The relationship between the structure of genes and their expression is a relatively new aspect of genome organization and regulation. With more genome sequences and expression data becoming available, bioinformatics approaches can help the further elucidation of the relationships between gene structure and gene expression. This will contribute to our understanding of a yet deeper level of gene regulation in higher eukaryotes. This thesis focuses on two issues of genome organization in relationship to expression. The genomic configuration involved in coexpression of neighboring genes is investigated (chapters 3 + 4) and the genome-wide relationships between structural parameters of a gene and its expression are analyzed (chapters 5 + 6). A short introduction (chapter 1) outlines the motivation and structure of this thesis. This is followed by an overview of issues that need to be considered in the study of gene and genome structure in relation to gene expression (chapter 2). DNA configuration in the nucleus is summarized and concepts as gene, chromatin and higher order domains are presented in the context of the measurement of gene expression and gene regulation. Special attention is given to the characteristics and functions of introns in the genomes of higher eukaryotes.

Expression of genes in eukaryotic genomes is known to cluster in domains, but domain size is generally loosely defined and highly variable. The concept of local coexpression domain is introduced and defined as set of physically adjacent genes that are highly coexpressed (chapter 3). The *Arabidopsis thaliana* genome was analyzed for the presence of such local coexpression domains and their functional characteristics were investigated. Public domain expression data from the Massively Parallel Signature Sequencing (MPSS) repository that cover a range of different experimental conditions, organs, tissues and cells and microarray data (Affymetrix) from a detailed analysis of gene expression in root were used. With these expression data, we identified 689 (MPSS) and 1481 (microarray) local coexpression domains consisting of 2 to 4 genes with a pair-wise Pearson's correlation coefficient larger than 0.7. This number is about 2 to 5-fold higher than the numbers expected by chance on the basis of genome randomizations. A small (5-10%) yet significant fraction of genes in the Arabidopsis genome is therefore organized into local coexpression domains. These local coexpression domains were apparently randomly distributed over the genome. Genes in such local domains were for the major part not categorized in the same functional category (GOslim). Neither tandemly duplicated genes, nor a shared promoter sequence, or gene distance fully explained the occurrence of coexpression of genes in such chromosomal

domains. This indicates that other parameters in genes or gene positions are important to establish coexpression of genes in local domains of Arabidopsis.

The analytical approach was extended to the analysis of the occurrence of local coexpression domains in the genome of rice (*Oryza sativa*), the monocotyledonous model plant (chapter 4). Also in the rice genome, there is a small, yet significant number of local coexpression domains that for the major part were not categorized in the same functional category (GOslim). The various configuration parameters studies could not fully explain the occurrence of local coexpression domains. The regulation of coexpression is therefore thought to be regulated at the level of chromatin structure. The characteristics of the local coexpression domains in rice are strikingly similar to such domains in the Arabidopsis genome. Yet, no microsynteny between local coexpresion domains in Arabidopsis and rice could be identified (chapter 4). Although the rice genome is not yet as extensively annotated as the Arabidopsis genome, the lack of conservation of local coexpression domains indicates that such domains have not played a major role in evolution.

In chapter 5, the relationships between the structure of a primary transcript and the expression level of the gene were investigated to identify the parameters and mechanisms that have helped shaping such relationships. In both monocotyledonous rice and dicotyledonous Arabidopsis, highly expressed genes were shown to have more and longer introns, as well as a larger primary transcript than lowly expressed genes. It is concluded that higher expressed genes tend to be less compact than lower expressed genes. In animal genomes, it is reported to be the other way round. Although the length differences in plant genes are much smaller than in animals, these findings indicate that plant genes are in this respect different from animal genes. Explanations for the relationship between gene configuration and gene expression in animals may be (or may have been) less important in plants. We speculate that selection, if any, on genome configuration has taken a different turn after the divergence of plants and animals.

To be able to exclude that the methodological differences were the reason for the reported differences between plant and animal gene structure and expression relationships, a comparative genomics study of five widely diverged genomes was undertaken (chapter 6).

The relationships between gene structure and gene expression were analyzed for five genomes (Arabidopsis, rice, worm, mouse, human), using public domain MPSS and affymetrix microarray (for worm) expression data sets that cover a wide variety of tissues and conditions. Five different parameters of gene structure were examined with the help of rank-based methods: the number of introns, as well as the total length of introns, combined untranslated regions, coding sequence and the combined total length of the primary transcript. In addition, the broadness or breadth of expression is evaluated. The methods of

analyses were identical for all genomes considered. It was found that tissue specific genes, defined as genes that are expressed in only one (or at most a few) tissues/conditions, are among the more compact genes in all genomes evaluated. Moreover, in plants the higher expressed genes tend to be longer and less compact than the lower expressed genes, whereas in the mammalian genomes analyzed the trend is the opposite. Worm takes an intermediate position. The different genomes differ markedly in the details of the relationship between expression and structure for the genes that are in the middle class of expression level. As the major difference in genome configuration is the absolute length of introns, possible explanations for the contrasting trends in plant and mammalian genomes question the role and evolutionary history of introns. Possibly there is a threshold amount of intron number and/or size upon which selection acts that differs between genomes. Alternatively, some groups of plant introns have been introduced in plant genomes well after the split between animals and plants.

The results of the research presented in this thesis are considered in the context and future prospects of the wider, more detailed and more comparative analyses of the relationships between gene structure and gene expression in the genomes of higher eukaryotes (chapter 7).

Samenvatting

De relatie tussen de structuur van genen en hun expressie is een relatief nieuw aspect in de studie van de organisatie van genomen en de regulatie van genexpressie. Nu meer genoomsequenties en expressiedata beschikbaar komen, kunnen benaderingen uit de bioinformatica helpen bij de verdere opheldering van de relatie tussen genstructuur en genexpressie. Dit zal bijdragen tot ons begrip van een weer dieper niveau van genregulatie in hogere eukaryoten. Dit proefschrift richt zich op twee aspecten van genoomorganisatie in relatie tot genexpressie. De genoomconfiguratie betrokken bij de coexpressie van naburige genen is onderzocht (hoofdstukken 3 + 4) en de verhouding tussen diverse structurele parameters van een gen en diens expressie is geanalyseerd op de schaal van hele genomen (hoofdstukken 5 + 6). Een korte inleiding (hoofdstuk 1) geeft de achtergrond en de structuur van dit proefschrift. Dit wordt gevolgd door een overzicht van onderwerpen die in de studie van gen- en genoomstructuur in relatie tot genexpressie een rol kunnen of moeten spelen (hoofdstuk 2). De configuratie van DNA in de kern wordt besproken en concepten als gen, chromatine en hogere ordedomeinen worden geplaatst in de context van het meten van genexpressie en genregulatie. Daarbij gaat de aandacht vooral uit naar de karakteristieken en de functies van introns in de genomen van hogere eukaryoten.

Eerder onderzoek heeft laten zien dat de expressie van genen in eukaryote genomen geclusterd is in domeinen. De omvang van die domeinen wordt over het algemeen nogal vaag gedefinieerd en blijkt dan zeer variabel. Het concept ‘lokaal coexpressiedomein’ wordt hier geïntroduceerd en gedefinieerd als een reeks fysiek naast elkaar gelegen genen die sterke coexpressie laten zien (hoofdstuk 3). Het genoom van de zandraket (*Arabidopsis thaliana*) is geanalyseerd op de aanwezigheid van dergelijke lokale coexpressiedomeinen en hun functionele karakteristieken zijn onderzocht. Voor deze analyses zijn publiek beschikbare gegevens over genexpressie gebruikt. Enerzijds zijn dit de Massively Parallel Signature Sequencing (MPSS) data, die een scala aan verschillende experimentele omstandigheden, organen, weefsels en cellen bieden. Anderzijds zijn microarray data (Affymetrix) gebruikt, die behoren bij een gedetailleerde analyse van genexpressie in wortel. Met deze expressiegegevens hebben wij 689 (MPSS) en 1481 (microarray) lokale coexpressiedomeinen geïdentificeerd; Deze domeinen bestaan uit 2 tot 4 genen waarvan de expressie voor iedere combinatie van twee genen een Pearson's correlatiecoëfficiënt heeft die groter is dan 0,7. Dit aantal domeinen is ongeveer 2-5 keer hoger dan het aantal dat door toeval kan worden verwacht op basis van willekeurige verdelingen van genvolgordes over het genoom. Een kleine (5-10%) maar significante fractie van de genen in het genoom van

Arabidopsis is dus georganiseerd in lokale coexpressiedomeinen. Deze lokale coexpressiedomeinen bleken willekeurig verdeeld over het genoom. De genen in dergelijke lokale domeinen komen voor het merendeel niet uit dezelfde functionele categorie (GOslim). Achterelkaar liggende gedupliceerde genen, een gedeelde promotorsequentie, of genafstand, konden het optreden van coexpressie van genen in lokale chromosomale domeinen niet volledig verklaren. Dit wijst erop dat andere parameters van genen of genposities belangrijk zijn om coexpressie van genen in lokale domeinen te bewerkstelligen. Eenzelfde analyse is uitgevoerd om lokale coexpressiedomeinen te identificeren en analyseren in het genoom van rijst (*Oryza sativa*), de eenzaadlobbige (monocotyle) modelplant (hoofdstuk 4). Ook in het rijstgenoom komt een klein, maar significant aantal lokale coexpressiedomeinen voor, waarvan de genen voor het merendeel niet uit dezelfde functionele categorie komen (GOslim). De diverse parameters voor genstructuur en gen-oriëntatie konden ook in rijst het bestaan van lokale coexpressiedomeinen niet volledig verklaren. We concluderen daarom dat coexpressie op het niveau van chromatinestructuren wordt gereguleerd. De globale karakteristieken van de lokale coexpressiedomeinen in rijst lijken opvallend veel op de karakteristieken van dergelijke domeinen in het genoom van Arabidopsis. Desondanks kon er geen microsyntenie tussen de lokale coexpressiedomeinen in Arabidopsis en rijst worden gedetecteerd (hoofdstuk 4). Hoewel het rijstgenoom nog niet zo uitgebreid is geannoteerd als het Arabidopsis genoom, wijst het gebrek aan conservatie van lokale coexpressiedomeinen erop dat dergelijke domeinen geen belangrijke rol in de evolutie hebben gespeeld.

In hoofdstuk 5 wordt de relatie tussen de structuur van een primair transcript en het expressieniveau van hetzelfde gen onderzocht om de parameters en de mechanismen te identificeren die hebben bijgedragen aan een dergelijke relatie. De resultaten laten zien dat voor genen in zowel eenzaadlobbige rijst als tweezaadlobbige Arabidopsis, de genen met hogere expressieniveaus meer en langere introns hebben, evenals een langer primair transcript, dan genen met lagere expressieniveaus. Dit betekent dat genen met hogere expressieniveaus minder compact neigen te zijn dan genen met een lager expressieniveau. Voor dierlijke genomen was het omgekeerde gerapporteerd. Hoewel de verschillen in lengte tussen plantengenen veel kleiner zijn dan de verschillen tussen dierlijke genen, betekenen deze resultaten dat plantengenen in dit opzicht verschillend zijn van dierlijke genen. De oorzaken die verantwoordelijk zijn voor de relatie tussen genconfiguratie en genexpressie in dieren, kunnen in planten minder belangrijk zijn, of minder belangrijk zijn geweest. We speculeren dat een selectie op genoomconfiguratie, als die er al is, een verschillende richting is ingeslagen na de evolutionaire splitsing tussen planten en dieren.

Om te kunnen uitsluiten dat methodologisch-analytische verschillen de oorzaak zijn voor de gevonden verschillen in de relatie tussen genstructuur en expressie van genen in planten en dieren, is een vergelijkende studie van vijf sterk gedivergeerde genomen uitgevoerd (hoofdstuk 6). De relatie tussen genstructuur en genexpressie is geanalyseerd voor de genomen van Arabidopsis, rijst, worm, mens, muis, gebruikmakend van publiek beschikbare MPSS en Affymetrix microarray (voor worm) expressiedata, die beide allerlei weefsels en experimentele condities omvatten. Vijf verschillende parameters voor de structuur van genen zijn onderzocht met behulp van kwantitatieve methoden die op rangordes zijn gebaseerd: het aantal introns, de totale lengte van introns, de gecombineerde niet getransleerde sequenties, de eiwitcoderende sequentie, als ook de gecombineerde totale lengte van het primaire transcript. Bovendien is de expressie in de breedte (over weefsels en condities) geëvalueerd. Eenzelfde analytische methode is gebruikt voor alle genomen die zijn bestudeerd. De analyses laten zien dat in alle onderzochte genomen de weefselspecifieke genen, die zijn gedefinieerd als genen die in slechts één (of hoogstens enkele) weefsels/condities tot expressie komen, tot de meer compacte genen behoren. Terwijl in planten genen met hogere expressieniveaus langer en minder compact neigen te zijn dan genen met een lager expressieniveau, is deze trend het tegenovergestelde in de geanalyseerde genomen van zoogdieren. Het genoom van worm neemt een tussenpositie in. De geanalyseerde genomen verschillen aanzienlijk in de details van de relatie tussen expressie en structuur voor de genen die tot de middenklasse van expressieniveaus behoren. Aangezien het belangrijkste verschil in genoomconfiguratie de absolute lengte van introns is, wijzen mogelijke verklaringen voor de contrasterende trends tussen de genomen van planten en zoogdieren naar de rol en de evolutionaire geschiedenis van introns. Misschien bestaat er een drempel voor het aantal introns en/of hun lengte, waarop selectie wordt gebaseerd; de drempel verschilt dan tussen genomen. Of wellicht zijn bepaalde groepen plant introns pas in plantengenomen geïntroduceerd na de evolutionaire splitsing tussen dieren en planten.

In hoofdstuk 7 worden de resultaten van het onderzoek dat in dit proefschrift is beschreven geplaatst in de context en de toekomstige perspectieven van bredere, meer gedetailleerde en meer vergelijkende analyses van de relaties tussen genstructuur en genexpressie in de genomen van hogere eukaryoten.

Supplementary Materials

Chapter 5 Supplementary Materials

Table of contents:

README

This is an extensive and detailed supplementary document. Please read this README section first, so you can find the right information easily. Each section and subsection is marked with a distinct name at the header of the page.

First part: Various analyses all confirmed the conclusion in the article
This part comprises 9 different analyses on 3 datasets in parallel.
For each analysis, 3 figures (legends are as in the article) and 3 tables (each for one dataset, 40% quantile) with detailed gene parameters were shown.

Three datasets are:

1. Dataset 1: Arabidopsis MPSS expression dataset
2. Dataset 2: Rice MPSS expression dataset
3. Dataset 3: Arabidopsis microarray (MA) expression dataset (not in the figure)

Nine different analyses are:

I. **Analysis I.** This analysis is described as in the article. The data is sorted by the average expression rank (rE) in an ascending order and highly and lowly expressed genes in equal quantiles are compared as described in the article.

II. **Analysis II.** Similar as Analysis I, with the only exception of the sorting method. Data is sorted based on ‘the peak expression rank (pE) over all the libraries’, in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared.

III. **Analysis III.** Similar as Analysis I, with the only exception of the sorting method. Data is sorted based on ‘the average expression value over all the libraries’, in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared.

IV. **Analysis IV.** Similar as Analysis I, with the only exception of the sorting method. Data is sorted based on ‘the peak expression value over all the libraries’, in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared.

V. **Analysis V.** Leave out intronless genes, repeat Analysis I.

VI. **Analysis VI.** Leave out 1st intron, only looking at genes with ≥ 2 introns, repeat Analysis I.

VII. **Analysis VII.** Leave out first 4 introns, only looking at genes with ≥ 5 introns, repeat Analysis I.

VIII. **Analysis VIII.** Leave out genes known to undergo alternative splicing, repeat Analysis I.

IX. **Analysis IV.** Only considering those genes that have both 5’ and 3’ UTR annotations, repeat Analysis I.

Second part: Information about expression libraries

Third part: Additional Analysis regard to gametophytic selection issue
There’s one additional analysis on the supplementary data from (Seoighe et al., 2005), using their sporophyte microarray expression data (in 5 libraries: root, leaf, stem, seedling green plant, hypocotyls) and our expression ranking method. We want to see whether in Arabidopsis (plant) sporophyte selection for short introns could also be observed as claimed in their pollen study. Our results show that selection for short introns could not be observed in sporophyte.

Extra remarks

- Explanation of the table content

- Title row: n in bracket is the number of genes in each group.
- "Difference" column is the average value in highly expressed column minus the average value in lowly expressed column, this value is by default very significant (p-value< 10^{-4}) according to the z value approximation of the non-parametric Mann-Whitney test for the comparison of two samples, unless otherwise mentioned in the bracket. (n.s.) means no significant difference (p-value>10^{-4}) between the average of two populations.
- Each cell in the main body of the table shows: average ± standard error of the mean and in bracket: median.
- CDS: coding sequence
- aa: amino acids
- Log-transformations of the length parameters are $\log_{10}$ based.

- Explanation of the figures

- Figure legends are similar as in the main text, with only combining Arabidopsis and rice MPSS data in one figure. Three figures in order in each analysis are: number of introns, total intron length and length of transcript relative to expression quantiles.
- Grey color is for Arabidopsis MPSS dataset and black color is for rice MPSS dataset. Triangles for highly expressed genes and squares for lowly expressed genes.
- In both tables and figures, all referrals to "transcript" means "primary transcript" in this study, if "primary" is omitted.

Chapter 5 Supplementary Materials
Analysis I. Sort data by average expression rank (rE)

First part: *Various analyses all confirmed the conclusion in the article*

Analysis I – Analyzed by average expression rank (rE)

This analysis is described as in the article, using the genes in the whole dataset, which is free of overlapping genes, retro(transposon), pseudo genes, using only the longest splicing variants for one gene locus.
Data is sorted by the **average expression rank** (**rE**, see article) in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared as described in the article.
Figures are as shown in the article, not repeated here anymore.
Tables presented here include more gene parameters and some log-transformed parameters, in 40% quantile.

Dataset 1: Arabidopsis MPSS, 18394 genes in total, 7358 genes in 40% quantile.
Dataset 2: Rice MPSS, 21431 genes in total, 8572 genes in 40% quantile
Dataset 3: Arabidopsis MA, 19046 genes in total, 7618 genes in 40% quantile

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.

Analysis I. Sort data by average expression rank (rE)

Dataset 1 Analysis I.	Arabidopsis MPSS			
	Highly expressed genes (n=7358)	Lowly expressed genes (n=7358)	Difference	All expressed Genes (n=18394)
Number of introns	5.5 ± 0.07 (4)	3.8 ± 0.06 (2)	1.7	4.7 ± 0.04 (3)
Average intron length per gene (bp)	164 ± 1.8 (133)	140 ± 2.1 (106)	24	152 ± 1.2 (120)
Average exon length per gene (bp)	372 ± 5.0 (212)	479 ± 5.7 (293)	-107	430 ± 3.5 (251)
Total intron length per gene (bp)	876 ± 11 (684)	603 ± 9.0 (367)	273	740 ± 6.3 (533)
Total CDS per gene (bp)	1396 ± 12 (1173)	1284 ± 9.9 (1113)	112	1350 ± 6.8 (1152)
Length of primary transcript (bp)	2692 ± 20 (2313)	2105 ± 17 (1822)	587	2411 ± 12 (2082)
Protein length (aa)	464 ± 3.9 (390)	427 ± 3.3 (370)	37	449 ± 2.3 (383)
Intron density per kb CDS	4.14 ± 0.038 (3.71)	2.95 ± 0.034 (2.21)	1.19	3.53 ± 0.023 (2.94)
Intron density per kb primary transcript	1.78 ± 0.014 (1.80)	1.50 ± 0.015 (1.35)	0.28	1.64 ± 0.009 (1.57)
Intergenic spacer (bp)	3679 ± 53 (2398)	4393 ± 97 (2766)	-714	4024 ± 47 (2582)
Log(primary transcript length)	3.36 ± 0.003 (3.36)	3.23 ± 0.003 (3.26)	0.13	3.30 ± 0.002 (3.32)
Log(Total intron length per gene)	2.44 ± 0.013 (2.84)	2.03 ± 0.015 (2.56)	0.41	2.24 ± 0.009 (2.73)
Total number of introns	40459	27940	12519	85590
Total length of introns (bp)	6444760	4434950	2009810	13616379
Average intron length per group	159	159	0	159

Dataset 2 Analysis I.	Rice MPSS			
	Highly expressed genes (n=8572)	Lowly expressed genes (n=8572)	Difference	All expressed Genes (n=21431)
Number of introns	5.9 ± 0.06 (4)	3.6 ± 0.05 (2)	2.3	4.7 ± 0.04 (3)
Average intron length per gene (bp)	416 ± 4.4 (333)	359 ± 4.6 (250)	57	387 ± 2.9 (298)
Average exon length per gene (bp)	329 ± 3.8 (203)	474 ± 5.4 (298)	-145	405 ± 3.0 (244)
Total intron length per gene (bp)	2204 ± 23 (1818)	1405 ± 20 (816)	799	1805 ± 14 (1368)
Total CDS per gene (bp)	1400 ± 11 (1164)	1251 ± 9.4 (1071)	149	1339 ± 6.6 (1128)
Length of primary transcript (bp)	3988 ± 30 (3420)	2842 ± 25 (2277)	1146	3435 ± 18 (2895)
Protein length (aa)	466 ± 3.7 (387)	416 ± 3.1 (356)	50	446 ± 2.2 (375)
Intron density per kb CDs	4.40 ± 0.036 (3.96)	3.05 ± 0.033 (2.20)	1.35	3.71 ± 0.022 (3.02)
Intron density per kb primary transcript	1.08 ± 0.010 (0.98)	1.31 ± 0.009 (1.28)	-0.23	1.19 ± 0.006 (1.13)
Intergenic spacer (bp)	7527 ± 64 (5975)	7476 ± 59 (6080)	51 (n.s.)	7533 ± 39 (6054)
Log(primary transcript length)	3.50 ± 0.003 (3.53)	3.32 ± 0.004 (3.36)	0.18	3.41 ± 0.002 (3.46)
Log(Total intron length per gene)	2.85 ± 0.012 (3.26)	2.32 ± 0.015 (2.91)	0.53	2.59 ± 0.009 (3.14)
Total number of introns	50229	30559	19670	101327
Total length of introns (bp)	18890507	12042037	6848470	38683787
Average intron length per group	376	394	-18	382

Dataset 3 Analysis I.	Arabidopsis MA			
	Highly expressed genes (n=7618)	Lowly expressed genes (n=7618)	Difference	All expressed Genes (n=19046)
Number of introns	5.5 ± 0.07 (4)	3.2 ± 0.05 (2)	2.3	4.4 ± 0.04 (3)
Average intron length per gene (bp)	165 ± 1.8 (136)	142 ± 2.2 (100)	23	152 ± 1.2 (119)
Average exon length per gene (bp)	360 ± 4.7 (203)	495 ± 5.7 (308)	-135	431 ± 3.4 (252)
Total intron length per gene (bp)	878 ± 9.3 (683)	521 ± 8.0 (295)	357	711 ± 6.1 (502)
Total CDS per gene (bp)	1346 ± 11 (1125)	1217 ± 9.3 (1062)	129	1310 ± 6.6 (1113)
Length of primary transcript (bp)	2631 ± 19 (2252)	1890 ± 15 (1632)	741	2301 ± 11 (1980)
Protein length (aa)	448 ± 3.7 (374)	405 ± 3.1 (353)	43	436 ± 2.2 (370)
Intron density per kb CDS	4.26 ± 0.038 (3.89)	2.65 ± 0.031 (2.01)	1.61	3.45 ± 0.022 (2.85)
Intron density per kb primary transcript	1.81 ± 0.014 (1.83)	1.43 ± 0.015 (1.25)	0.38	1.63 ± 0.009 (1.55)
Intergenic spacer (bp)	3697 ± 56 (2344)	5403 ± 119 (3108)	-1706	4462 ± 56 (2703)
Log(primary transcript length)	3.35 ± 0.003 (3.35)	3.18 ± 0.004 (3.21)	0.17	3.27 ± 0.002 (3.31)
Log(Total intron length per gene)	2.43 ± 0.013 (2.85)	1.93 ± 0.014 (2.47)	0.5	2.19 ± 0.009 (2.70)
Total number of introns	41605	24244	17361	84386
Total length of introns (bp)	6697209	3970487	2726722	13545532
Average intron length per group	161	164	-3	161

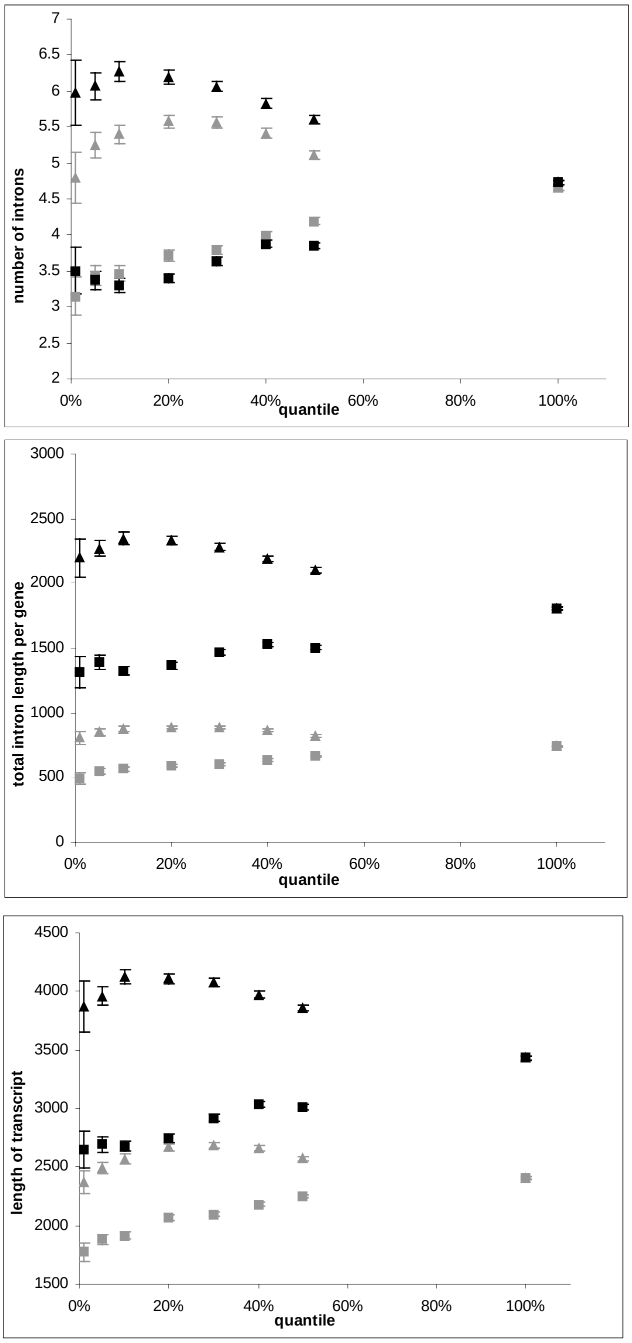
Analysis II – Analyzed by peak expression rank (pE)

Figure s5.2

This analysis is similar as **Analysis I**, with the only exception of the sorting method. Data is sorted based on 'the **peak expression rank** over all the libraries', in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared. Tables show the data of 40% quantile.

Dataset 1: Arabidopsis MPSS, 18394 genes in total, 7358 genes in 40% quantile.
Dataset 2: Rice MPSS, 21431 genes in total, 8572 genes in 40% quantile
Dataset 3: Arabidopsis MA, 19046 genes in total, 7618 genes in 40% quantile

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.



Dataset 1 **Analysis II.**	**Arabidopsis MPSS**			
	Highly expressed genes (n=7358)	**Lowly expressed genes (n=7358)**	**Difference**	**All expressed Genes (n=18394)**
Number of introns per gene	5.4 ± 0.07 (4)	4.0 ± 0.06 (2)	1.4	4.7 ± 0.04 (3)
Average intron length per gene (bp)	164 ± 1.8 (133)	142 ± 2.1 (108)	22	152 ± 1.2 (120)
Average exon length per gene (bp)	371 ± 4.9 (213)	483 ± 5.8 (291)	-112	430 ± 3.5 (251)
Total intron length per gene (bp)	861 ± 11 (677)	632 ± 9.3 (401)	229	740 ± 6.3 (533)
Total CDS per gene (bp)	1380 ± 12 (1161)	1319 ± 10 (1140)	61	1350 ± 6.8 (1152)
Length of primary transcript (bp)	2661 ± 20 (2291)	2179 ± 17 (1893)	482	2411 ± 12 (2082)
Protein length (aa)	459 ± 3.9 (386)	439 ± 3.4 (379)	20	449 ± 2.3 (383)
Intron density per kb CDS	4.11 ± 0.038 (3.67)	3.05 ± 0.035 (2.29)	1.06	3.53 ± 0.023 (2.94)
Intron density per kb primary transcript	1.77 ± 0.014 (1.78)	1.53 ± 0.015 (1.37)	0.24	1.64 ± 0.009 (1.57)
Intergenic spacer (bp)	3720 ± 53 (2434)	4330 ± 98 (2674)	-610	4024 ± 47 (2582)
Log(primary transcript length)	3.35 ± 0.003 (3.36)	3.25 ± 0.003 (3.28)	0.10	3.30 ± 0.002 (3.32)
Log(Total intron length per gene)	2.43 ± 0.013 (2.83)	2.06 ± 0.015 (2.60)	0.37	2.24 ± 0.009 (2.73)
Total number of introns	39810	29351	10459	85590
Total length of introns (bp)	6337272	4648538	1688734	13616379
Average intron length per group	159	158	1	159

Dataset 2 **Analysis II.**	**Rice MPSS**			
	Highly expressed genes (n=8572)	**Lowly expressed genes (n=8572)**	**Difference**	**All expressed Genes (n=21431)**
Number of introns per gene	5.8 ± 0.06 (4)	3.9 ± 0.05 (2)	1.9	4.7 ± 0.04 (3)
Average intron length per gene (bp)	416 ± 4.4 (332)	374 ± 4.6 (272)	42	387 ± 2.9 (298)
Average exon length per gene (bp)	330 ± 3.8 (203)	466 ± 5.4 (286)	-136	405 ± 3.0 (244)
Total intron length per gene (bp)	2193 ± 23 (1811)	1529 ± 20 (967)	664	1805 ± 14 (1368)
Total CDS per gene (bp)	1395 ± 11 (1161)	1301 ± 9.9 (1107)	94	1339 ± 6.6 (1128)
Length of primary transcript (bp)	3971 ± 30 (3410)	3036 ± 26 (2479)	935	3435 ± 18 (2895)
Protein length (aa)	464 ± 3.7 (386)	433 ± 3.3 (368)	31	446 ± 2.2 (375)
Intron density per kb CDs	4.40 ± 0.036 (3.94)	3.23 ± 0.034 (2.38)	1.17	3.71 ± 0.022 (3.02)
Intron density per kb primary transcript	1.31 ± 0.009 (1.27)	1.11 ± 0.010 (1.01)	0.2	1.19 ± 0.006 (1.13)
Intergenic spacer (bp)	7554 ± 64 (6012)	7412 ± 59 (6047)	142 (n.s.)	7533 ± 39 (6054)
Log(primary transcript length)	3.50 ± 0.003 (3.53)	3.35 ± 0.004 (3.39)	0.15	3.41 ± 0.002 (3.46)
Log(Total intron length per gene)	2.85 ± 0.012 (3.26)	2.41 ± 0.014 (3.00)	0.44	2.59 ± 0.009 (3.14)
Total number of introns	49958	33255	16703	101327
Total length of introns (bp)	18796160	13105103	5691057	38683787
Average intron length per group	376	394	-18	382

Dataset 3 **Analysis II.**	**Arabidopsis MA**			
	Highly expressed genes (n=7618)	**Lowly expressed genes (n=7618)**	**Difference**	**All expressed Genes (n=19046)**
Number of introns per gene	5.1 ± 0.06 (4)	3.3 ± 0.05 (2)	1.8	4.4 ± 0.04 (3)
Average intron length per gene (bp)	165 ± 1.8 (134)	141 ± 2.1 (102)	24	152 ± 1.2 (119)
Average exon length per gene (bp)	367 ± 4.7 (215)	492 ± 5.8 (302)	-125	431 ± 3.4 (252)
Total intron length per gene (bp)	827 ± 10 (660)	542 ± 8.5 (309)	285	711 ± 6.1 (502)
Total CDS per gene (bp)	1298 ± 11 (1088)	1243 ± 9.6 (1074)	55	1310 ± 6.6 (1113)
Length of primary transcript (bp)	2518 ± 19 (2169)	1942 ± 16 (1666)	576	2301 ± 11 (1980)
Protein length (aa)	432 ± 3.6 (361)	414 ± 3.2 (357)	18	436 ± 2.2 (370)
Intron density per kb CDS	4.07 ± 0.038 (3.57)	2.75 ± 0.032 (2.09)	1.32	3.45 ± 0.022 (2.85)
Intron density per kb primary transcript	1.74 ± 0.014 (1.72)	1.47 ± 0.015 (1.30)	0.27	1.63 ± 0.009 (1.55)
Intergenic spacer (bp)	3911 ± 58 (2515)	5313 ± 121 (2985)	-1402	4462 ± 56 (2703)
Log(primary transcript length)	3.33 ± 0.003 (3.34)	3.19 ± 0.004 (3.22)	0.14	3.27 ± 0.002 (3.31)
Log(Total intron length per gene)	2.37 ± 0.013 (2.82)	1.96 ± 0.014 (2.49)	0.41	2.19 ± 0.009 (2.70)
Total number of introns	38862	25735	13127	84386
Total length of introns (bp)	6299915	4176021	2123894	13545532
Average intron length per group	162	162	0	161

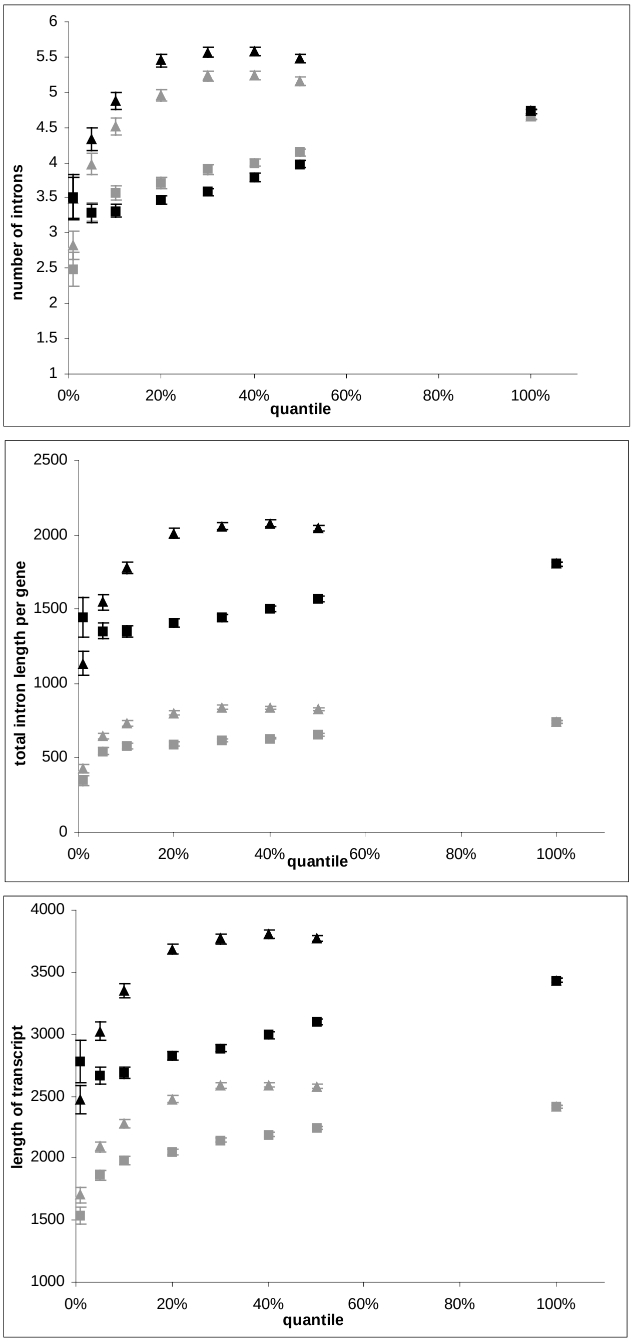
Analysis III – Analyzed by average expression value

This analysis is similar as **Analysis I**, with the only exception of the sorting method.
Data is sorted based on 'the **average expression value** over all the libraries', in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared.
Tables show the data of 40% quantile.

Dataset 1: Arabidopsis MPSS, 18394 genes in total, 7358 genes in 40% quantile.
Dataset 2: Rice MPSS, 21431 genes in total, 8572 genes in 40% quantile
Dataset 3: Arabidopsis MA, 19046 genes in total, 7618 genes in 40% quantile

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.

Figure s5.3



Dataset 1 **Analysis III.**	**Arabidopsis MPSS**			
	Highly expressed genes (n=7358)	**Lowly expressed genes (n=7358)**	**Difference**	**All expressed Genes (n=18394)**
Number of introns	5.2 ± 0.07 (4)	4.0 ± 0.06 (2)	1.2	4.7 ± 0.04 (3)
Average intron length per gene (bp)	164 ± 1.8 (132)	142 ± 2.1 (107)	22	152 ± 1.2 (120)
Average exon length per gene (bp)	369 ± 4.9 (215)	485 ± 5.9 (292)	-116	430 ± 3.5 (251)
Total intron length per gene (bp)	834 ± 10 (653)	632 ± 9.3 (398)	202	740 ± 6.3 (533)
Total CDS per gene (bp)	1341 ± 11 (1131)	1329 ± 10 (1149)	12 (n.s.)	1350 ± 6.8 (1152)
Length of primary transcript (bp)	2591 ± 19 (2221)	2189 ± 17 (1907)	402	2411 ± 12 (2082)
Protein length (aa)	446 ± 3.7 (376)	442 ± 3.4 (382)	4 (n.s.)	449 ± 2.3 (383)
Intron density per kb CDS	4.06 ± 0.038 (3.58)	3.03 ± 0.034 (2.28)	1.03	3.53 ± 0.023 (2.94)
Intron density per kb primary transcript	1.75 ± 0.014 (1.75)	1.53 ± 0.015 (1.37)	0.22	1.64 ± 0.009 (1.57)
Intergenic spacer (bp)	3860 ± 57 (2509)	4346 ± 97 (2679)	-486	4024 ± 47 (2582)
Log(primary transcript length)	3.34 ± 0.003 (3.35)	3.25 ± 0.003 (3.28)	0.09	3.30 ± 0.002 (3.32)
Log(Total intron length per gene)	2.41 ± 0.013 (2.81)	2.07 ± 0.014 (2.60)	0.34	2.24 ± 0.009 (2.73)
Total number of introns	38515	29428	9087	85590
Total length of introns (bp)	6165141	4649806	1515335	13616379
Average intron length per group	160	158	2	159

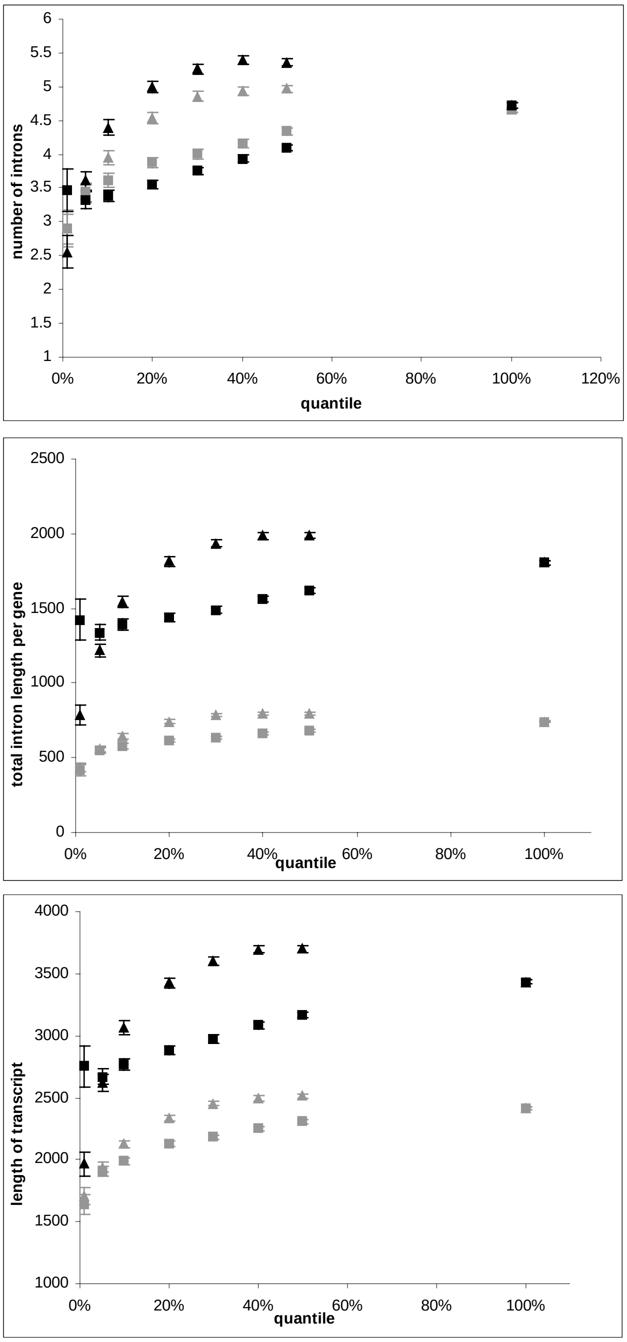
Dataset 2 **Analysis III.**	**Rice MPSS**			
	Highly expressed genes (n=8572)	**Lowly expressed genes (n=8572)**	**Difference**	**All expressed Genes (n=21431)**
Number of introns per gene	5.6 ± 0.06 (4)	3.8 ± 0.05 (2)	1.8	4.7 ± 0.04 (3)
Average intron length per gene (bp)	399 ± 4.3 (318)	374 ± 4.7 (271)	25	387 ± 2.9 (298)
Average exon length per gene (bp)	338 ± 3.8 (211)	469 ± 5.4 (288)	-131	405 ± 3.0 (244)
Total intron length per gene (bp)	2074 ± 23 (1687)	1502 ± 20 (938)	572	1805 ± 14 (1368)
Total CDS per gene (bp)	1359 ± 11 (1131)	1291 ± 9.8 (1104)	68	1339 ± 6.6 (1128)
Length of primary transcript (bp)	3806 ± 30 (3246)	2993 ± 26 (2444)	813	3435 ± 18 (2895)
Protein length (aa)	452 ± 3.6 (376)	429 ± 3.3 (367)	23	446 ± 2.2 (375)
Intron density per kb CDS	4.24 ± 0.036 (3.74)	3.19 ± 0.034 (2.35)	1.05	3.71 ± 0.022 (3.02)
Intron density per kb primary transcript	1.29 ± 0.010 (0.88)	1.10 ± 0.010 (1.01)	0.19	1.19 ± 0.006 (1.13)
Intergenic spacer (bp)	7630 ± 65 (6054)	7467 ± 59 (6100)	163 (n.s.)	7533 ± 39 (6054)
Log(primary transcript length)	3.47 ± 0.004 (3.51)	3.34 ± 0.004 (3.39)	0.13	3.41 ± 0.002 (3.46)
Log(Total intron length per gene)	2.77 ± 0.012 (3.23)	2.39 ± 0.014 (2.97)	0.38	2.59 ± 0.009 (3.14)
Total number of introns	47860	32501	15359	101327
Total length of introns (bp)	17780220	12874962	4905258	38683787
Average intron length per group	372	396	-24	382

Dataset 3 **Analysis III.**	**Arabidopsis MA**			
	Highly expressed genes (n=7618)	**Lowly expressed genes (n=7618)**	**Difference**	**All expressed Genes (n=19046)**
Number of introns per gene	5.3 ± 0.07 (4)	3.2 ± 0.05 (2)	2.1	4.4 ± 0.04 (3)
Average intron length per gene (bp)	167 ± 1.8 (136)	140 ± 2.2 (100)	27	152 ± 1.2 (119)
Average exon length per gene (bp)	361 ± 4.6 (209)	495 ± 5.8 (308)	-134	431 ± 3.4 (252)
Total intron length per gene (bp)	858 ± 10 (680)	527 ± 8.1 (294)	331	711 ± 6.1 (502)
Total CDS per gene (bp)	1320 ± 11 (1107)	1227 ± 9.5 (1065)	93	1310 ± 6.6 (1113)
Length of primary transcript (bp)	2578 ± 19 (2210)	1902 ± 15 (1638)	676	2301 ± 11 (1980)
Protein length (aa)	439 ± 3.7 (368)	408 ± 3.2 (354)	31	436 ± 2.2 (370)
Intron density per kb CDS	4.17 ± 0.038 (3.72)	2.67 ± 0.031 (2.01)	1.50	3.45 ± 0.022 (2.85)
Intron density per kb primary transcript	1.77 ± 0.014 (1.78)	1.44 ± 0.015 (1.26)	0.33	1.63 ± 0.009 (1.55)
Intergenic spacer (bp)	3791 ± 57 (2405)	5405 ± 120 (3066)	-1614	4462 ± 56 (2703)
Log(primary transcript length)	3.34 ± 0.003 (3.34)	3.18 ± 0.004 (3.21)	0.16	3.27 ± 0.002 (3.31)
Log(Total intron length per gene)	2.41 ± 0.013 (2.83)	1.93 ± 0.014 (2.47)	0.48	2.19 ± 0.009 (2.70)
Total number of introns	40314	24599	15715	84386
Total length of introns (bp)	6534428	4015094	2519334	13545532
Average intron length per group	162	163	-1	161

Chapter 5 Supplementary Materials
Analysis IV. Sort data by peak expression value

Analysis IV – Analyzed by peak expression value

Figure s5.4



This analysis is similar as **Analysis I**, with the only exception of the sorting method.
Data is sorted based on 'the **peak expression value** over all the libraries', in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared.
Tables show the data of 40% quantile comparison.

Dataset 1: Arabidopsis MPSS, 18394 genes in total, 7358 genes in 40% quantile.
Dataset 2: Rice MPSS, 21431 genes in total, 8572 genes in 40% quantile
Dataset 3: Arabidopsis MA, 19046 genes in total, 7618 genes in 40% quantile

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.

Dataset 1 **Analysis IV.**	**Arabidopsis MPSS**			
	Highly expressed genes (n=7358)	**Lowly expressed genes (n=7358)**	**Difference**	**All expressed Genes (n=18394)**
Number of introns per gene	4.9 ± 0.06 (3)	4.2 ± 0.06 (3)	0.7	4.7 ± 0.04 (3)
Average intron length per gene (bp)	162 ± 1.9 (130)	142 ± 2.1 (110)	20	152 ± 1.2 (120)
Average exon length per gene (bp)	379 ± 5.0 (223)	486 ± 5.9 (290)	-107	430 ± 3.5 (251)
Total intron length per gene (bp)	792 ± 10 (610)	658 ± 9.6 (426)	134	740 ± 6.3 (533)
Total CDS per gene (bp)	1302 ± 11 (1100)	1357 ± 10 (1176)	-55	1350 ± 6.8 (1152)
Length of primary transcript (bp)	2494 ± 19 (2132)	2254 ± 18 (1965)	240	2411 ± 12 (2082)
Protein length (aa)	433 ± 3.6 (366)	451 ± 3.5 (391)	-18	449 ± 2.3 (383)
Intron density per kb CDS	3.90 ± 0.038 (3.40)	3.09 ± 0.035 (2.34)	0.81	3.53 ± 0.023 (2.94)
Intron density per kb primary transcript	1.69 ± 0.014 (1.66)	1.55 ± 0.015 (1.41)	0.14	1.64 ± 0.009 (1.57)
Intergenic spacer (bp)	3964 ± 58 (2612)	4226 ± 95 (2596)	-262 (n.s.)	4024 ± 47 (2582)
Log(primary transcript length)	3.32 ± 0.003 (3.33)	3.27 ± 0.003 (3.29)	0.05	3.30 ± 0.002 (3.32)
Log(Total intron length per gene)	2.34 ± 0.013 (2.79)	2.09 ± 0.014 (2.63)	0.25	2.24 ± 0.009 (2.73)
Total number of introns	36296	30629	5667	85590
Total length of introns (bp)	5828279	4844707	983572	13616379
Average intron length per group	161	158	3	159

Dataset 2 **Analysis IV.**	**Rice MPSS**			
	Highly expressed genes (n=8572)	**Lowly expressed genes (n=8572)**	**Difference**	**All expressed Genes (n=21431)**
Number of introns per gene	5.4 ± 0.06 (4)	3.9 ± 0.05 (2)	1.5	4.7 ± 0.04 (3)
Average intron length per gene (bp)	386 ± 4.2 (308)	379 ± 4.7 (281)	7 (n.s.)	387 ± 2.9 (298)
Average exon length per gene (bp)	348 ± 3.9 (218)	463 ± 5.4 (282)	-115	405 ± 3.0 (244)
Total intron length per gene (bp)	1987 ± 23 (1584)	1564 ± 20 (1020)	423	1805 ± 14 (1368)
Total CDS per gene (bp)	1348 ± 11 (1121)	1310 ± 10 (1116)	38 (n.s.)	1339 ± 6.6 (1128)
Length of primary transcript (bp)	3695 ± 30 (3123)	3084 ± 26 (2535)	611	3435 ± 18 (2895)
Protein length (aa)	448 ± 3.6 (373)	436 ± 3.3 (371)	12 (n.s.)	446 ± 2.2 (375)
Intron density per kb CDS	4.09 ± 0.036 (3.54)	3.27 ± 0.034 (2.43)	0.82	3.71 ± 0.022 (3.02)
Intron density per kb primary transcript	1.27 ± 0.010 (1.22)	1.11 ± 0.010 (1.02)	0.16	1.19 ± 0.006 (1.13)
Intergenic spacer (bp)	7719 ± 65 (6173)	7470 ± 59 (6100)	249	7533 ± 39 (6054)
Log(primary transcript length)	3.45 ± 0.004 (3.50)	3.36 ± 0.004 (3.40)	0.09	3.41 ± 0.002 (3.46)
Log(Total intron length per gene)	2.71 ± 0.013 (3.20)	2.43 ± 0.014 (3.01)	0.28	2.59 ± 0.009 (3.14)
Total number of introns	46230	33735	12495	101327
Total length of introns (bp)	17028290	13405972	3622318	38683787
Average intron length per group	368	397	-29	382

Dataset 3 **Analysis IV.**	**Arabidopsis MA**			
	Highly expressed genes (n=7618)	**Lowly expressed genes (n=7618)**	**Difference**	**All expressed Genes (n=19046)**
Number of introns per gene	5.0 ± 0.06 (3)	3.4 ± 0.05 (2)	1.6	4.4 ± 0.04 (3)
Average intron length per gene (bp)	165 ± 1.9 (133)	141 ± 2.1 (102)	24	152 ± 1.2 (119)
Average exon length per gene (bp)	373 ± 4.8 (219)	491 ± 5.8 (302)	-118	431 ± 3.4 (252)
Total intron length per gene (bp)	808 ± 10 (640)	550 ± 8.5 (310)	258	711 ± 6.1 (502)
Total CDS per gene (bp)	1279 ± 11 (1080)	1246 ± 9.7 (1074)	33 (n.s.)	1310 ± 6.6 (1113)
Length of primary transcript (bp)	2475 ± 18 (2143)	1946 ± 16 (1665)	529	2301 ± 11 (1980)
Protein length (aa)	425 ± 3.5 (359)	414 ± 3.2 (357)	11 (n.s.)	436 ± 2.2 (370)
Intron density per kb CDS	4.01 ± 0.038 (3.49)	2.75 ± 0.032 (2.09)	1.26	3.45 ± 0.022 (2.85)
Intron density per kb primary transcript	1.72 ± 0.014 (1.69)	1.47 ± 0.015 (1.30)	0.25	1.63 ± 0.009 (1.55)
Intergenic spacer (bp)	3984 ± 61 (2559)	5316 ± 121 (2989)	-1332	4462 ± 56 (2703)
Log(primary transcript length)	3.32 ± 0.003 (3.33)	3.19 ± 0.004 (3.22)	0.13	3.27 ± 0.002 (3.31)
Log(Total intron length per gene)	2.35 ± 0.013 (2.81)	1.96 ± 0.014 (2.49)	0.39	2.19 ± 0.009 (2.70)
Total number of introns	37942	25758	12184	84386
Total length of introns (bp)	6153085	4188582	1964503	13545532
Average intron length per group	162	163	-1	161

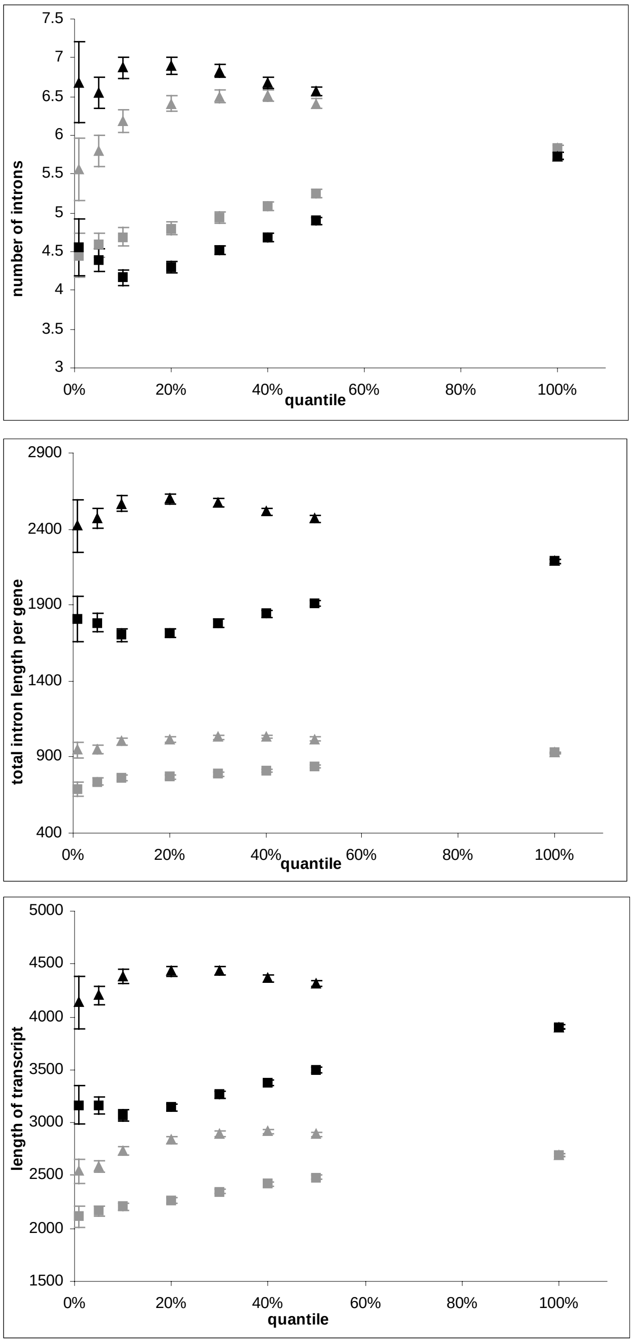
Analysis V – Remove intronless genes, repeat Analysis I.

Figure s5.5

This analysis is similar as **Analysis I**, with the only exception of removing intronless genes from the dataset.
Data is sorted by the **average expression rank** (**rE**, see article) in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared as described in the article.
Tables show the data of 40% quantile.

Dataset 1: Arabidopsis MPSS, 14686 genes in total, 5874 genes in 40% quantile.
Dataset 2: Rice MPSS, 17679 genes in total, 7072 genes in 40% quantile
Dataset 3: Arabidopsis MA, 15002 genes in total, 6001 genes in 40% quantile

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.



Analysis V. Remove intronless genes, repeat Analysis I.

Dataset 1	**Arabidopsis MPSS**			
Analysis V.	**Highly expressed genes (n=5874)**	**Lowly expressed genes (n=5874)**	**Difference**	**All expressed Genes (n=14686)**
Number of introns per gene	6.5 ± 0.08 (5)	5.1 ± 0.06 (4)	1.4	5.8 ± 0.04 (4)
Average intron length per gene (bp)	192 ± 1.9 (148)	188 ± 2.3 (134)	4 (n.s.)	190 ± 1.3 (141)
Average exon length per gene (bp)	253 ± 2.8 (181)	300 ± 3.0 (226)	-47	278 ± 1.9 (205)
Total intron length per gene (bp)	1035 ± 12 (808)	807 ± 10 (583)	228	927 ± 7.2 (709)
Total CDS per gene (bp)	1459 ± 14 (1218)	1383 ± 12 (1185)	76	1429 ± 8.0 (1215)
Length of primary transcript (bp)	2917 ± 23 (2499)	2424 ± 20 (2113)	493	2687 ± 13 (2323)
Protein length (aa)	485 ± 4.6 (405)	460 ± 3.9 (394)	25	475 ± 2.7 (403)
Intron density per kb CDS	4.89 ± 0.039 (4.49)	3.95 ± 0.035 (3.39)	0.94	4.42 ± 0.024 (3.87)
Intron density per kb primary transcript	2.10 ± 0.014 (2.06)	2.01 ± 0.014 (1.88)	0.09	2.05 ± 0.009 (1.97)
Intergenic spacer (bp)	3621 ± 62 (2285)	4129 ± 84 (2674)	-508	3866 ± 47 (2488)
Log(primary transcript length)	3.41 ± 0.003 (3.40)	3.32 ± 0.003 (3.32)	0.09	3.37 ± 0.002 (3.37)
Log(Total intron length per gene)	2.87 ± 0.005 (2.91)	2.72 ± 0.006 (2.77)	0.15	2.80 ± 0.003 (2.85)
Total number of introns	38214	29870	8344	85590
Total length of introns (bp)	6082096	4742897	1339199	13616379
Average intron length per group	159	159	0	159

Dataset 2	**Rice MPSS**			
Analysis V.	**Highly expressed genes (n=7072)**	**Lowly expressed genes (n=7072)**	**Difference**	**All expressed Genes (n=17679)**
Number of introns per gene	6.7 ± 0.07 (5)	4.7 ± 0.06 (3)	2	5.7 ± 0.04 (4)
Average intron length per gene (bp)	473 ± 4.7 (366)	469 ± 5.2 (347)	4 (n.s.)	469 ± 3.1 (357)
Average exon length per gene (bp)	247 ± 2.4 (179)	316 ± 3.1 (228)	-69	282 ± 1.8 (201)
Total intron length per gene (bp)	2514 ± 25 (2058)	1844 ± 22 (1874)	670	2188 ± 15 (1742)
Total CDS per gene (bp)	1461 ± 13 (1212)	1336 ± 11 (1143)	125	1415 ± 7.5 (1188)
Length of primary transcript (bp)	4368 ± 33 (3757)	3377 ± 29 (2875)	991	3905 ± 20 (3353)
Protein length (aa)	486 ± 4.2 (403)	444 ± 3.6 (380)	42	471 ± 2.5 (395)
Intron density per kb CDS	5.03 ± 0.037 (4.57)	3.98 ± 0.035 (3.22)	1.05	4.50 ± 0.023 (3.88)
Intron density per kb primary transcript	1.49 ± 0.009 (1.41)	1.40 ± 0.009 (1.27)	0.09	1.45 ± 0.006 (1.28)
Intergenic spacer (bp)	7379 ± 70 (5850)	7373 ± 64 (5985)	6 (n.s.)	7402 ± 42 (5911)
Log(primary transcript length)	3.56 ± 0.003 (3.58)	3.43 ± 0.004 (3.46)	0.13	3.50 ± 0.002 (3.53)
Log(Total intron length per gene)	3.24 ± 0.005 (3.31)	3.03 ± 0.006 (3.14)	0.21	3.14 ± 0.004 (3.24)
Total number of introns	47188	33150	14038	101327
Total length of introns (bp)	17777404	13042369	4735035	38683787
Average intron length per group	377	393	-16	382

Dataset 3	**Arabidopsis MA**			
Analysis V.	**Highly expressed genes (n=6001)**	**Lowly expressed genes (n=6001)**	**Difference**	**All expressed Genes (n=15002)**
Number of introns per gene	6.5 ± 0.07 (5)	4.5 ± 0.06 (3)	2	5.6 ± 0.04 (4)
Average intron length per gene (bp)	197 ± 2.0 (153)	195 ± 2.5 (130)	2 (n.s.)	193 ± 1.4 (141)
Average exon length per gene (bp)	243 ± 2.8 (173)	306 ± 3.0 (235)	-63	275 ± 1.8 (204)
Total intron length per gene (bp)	1046 ± 12 (822)	729 ± 9.4 (512)	317	903 ± 6.9 (512)
Total CDS per gene (bp)	1407 ± 13 (1176)	1320 ± 11 (1134)	87	1392 ± 7.8 (1179)
Length of primary transcript (bp)	2861 ± 22 (2447)	2215 ± 18 (1927)	646	2584 ± 13 (2234)
Protein length (aa)	468 ± 4.5 (391)	439 ± 3.7 (377)	29	463 ± 2.6 (392)
Intron density per kb CDS	5.07 ± 0.039 (4.69)	3.69 ± 0.033 (3.13)	1.38	4.38 ± 0.023 (3.82)
Intron density per kb primary transcript	2.14 ± 0.013 (2.11)	1.98 ± 0.014 (1.82)	0.16	2.07 ± 0.009 (1.98)
Intergenic spacer (bp)	3579 ± 65 (2203)	5525 ± 143 (3098)	-1946	4398 ± 67 (2605)
Log(primary transcript length)	3.40 ± 0.003 (3.39)	3.27 ± 0.004 (3.28)	0.13	3.34 ± 0.002 (3.35)
Log(Total intron length per gene)	2.88 ± 0.005 (2.91)	2.67 ± 0.006 (2.71)	0.21	2.79 ± 0.003 (2.84)
Total number of introns	38894	26875	12019	84386
Total length of introns (bp)	6279574	4377095	1902479	13545532
Average intron length per group	161	163	-2	161

Analysis VI – Leave out 1st introns, repeat Analysis I.

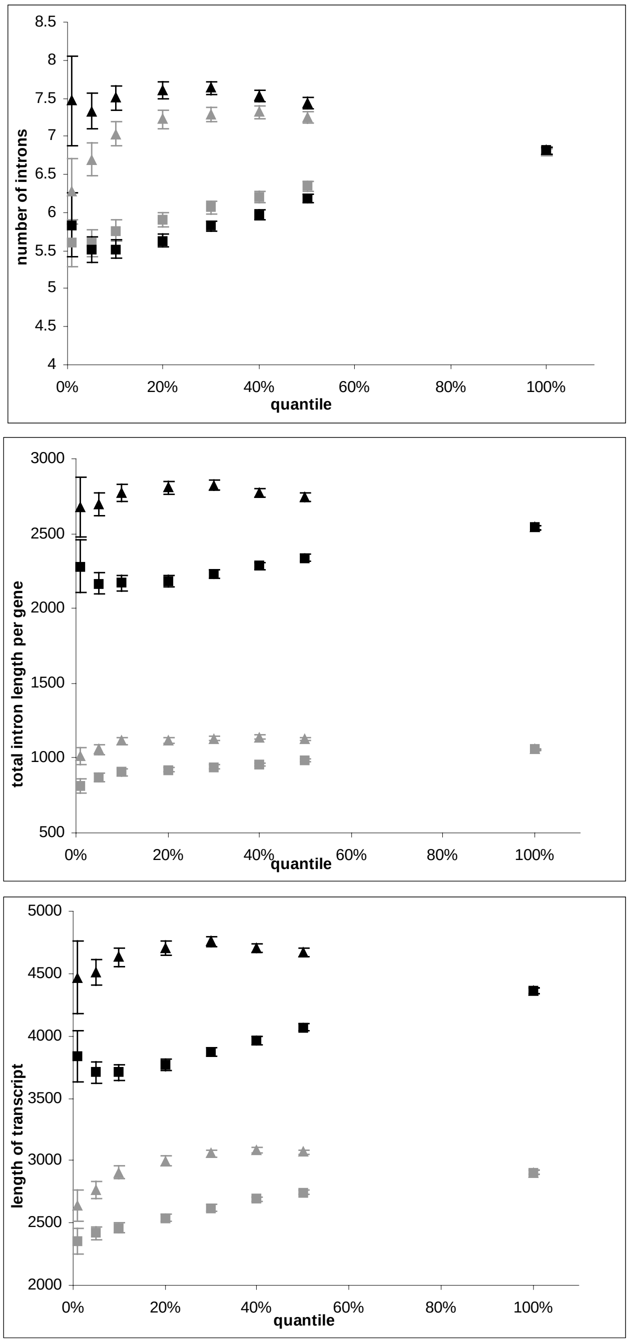
On the basis of the data of **Analysis IV** (eg. intron-containing genes), leave out the 5-prime 1st intron of all genes, only considering those that are still intron-containing genes afterwards (eg. genes with ≥ 2 introns).
Data is sorted by the **average expression rank** (**rE**, see article) in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared as described in the article.

Both 'Total intron length per gene' (eg. the sum of 1st intron till the end intron) and 'Total intron length per gene after leaving out 1st intron per gene' (eg. the sum of 2nd intron till the end intron) were compared between highly and lowly expressed quantiles, as well as the log-transformed these two parameters per gene.
Tables show the data of 40% quantile.

Dataset 1: Arabidopsis MPSS, 12229 genes in total, 4892 genes in 40% quantile.
Dataset 2: Rice MPSS, 14392 genes in total, 5757 genes in 40% quantile
Dataset 3: Arabidopsis MA, 12305 genes in total, 4922 genes in 40% quantile

This analysis confirmed that the positive correlation between expression characteristics and intron length is not restricted 5' proximal introns only.

Figure s5.6



Analysis VI. Leave out 1st introns, repeat Analysis I.

Dataset 1 **Analysis VI.**	**Arabidopsis MPSS**			
	Highly expressed genes (n=4892)	**Lowly expressed genes (n=4892)**	**Difference**	**All expressed Genes (n=12229)**
Number of introns per gene	7.3 ± 0.08 (6)	6.2 ± 0.07 (5)	1.1	6.8 ± 0.07 (5)
Average intron length per gene (bp)	173 ± 1.4 (146)	170 ± 1.7 (134)	3 (n.s.)	172 ± 1.0 (139)
Average exon length per gene (bp)	213 ± 2.1 (167)	247 ± 2.4 (202)	-34	230 ± 1.4 (184)
Total intron length per gene (bp)	1141 ± 13 (895)	957 ± 12 (726)	184	1057 ± 8.0 (821)
Total CDS per gene (bp)	1516 ± 15 (1266)	1488 ± 13 (1275)	28 (n.s.)	1509 ± 9.0 (1281)
Length of primary transcript (bp)	3084 ± 25 (2651)	2690 ± 22 (2340)	394	2906 ± 15 (2516)
Protein length (aa)	504 ± 5.1 (421)	495 ± 4.5 (424)	9 (n.s.)	502 ± 3.0 (425)
Intron density per kb CDS	5.40 ± 0.041 (5.03)	4.62 ± 0.038 (4.10)	0.78	5.00 ± 0.025 (4.51)
Intron density per kb primary transcript	2.31 ± 0.013 (2.26)	2.30 ± 0.014 (2.20)	0.01 (n.s.)	2.30 ± 0.009 (2.22)
Intergenic spacer (bp)	3613 ± 70 (2262)	4054 ± 87 (2600)	-441	3822 ± 49 (2445)
Log(primary transcript length)	3.43 ± 0.003 (3.42)	3.37 ± 0.003 (3.37)	0.06	3.41 ± 0.002 (3.40)
Log(Total intron length per gene)	2.95 ± 0.004 (2.95)	2.86 ± 0.005 (2.86)	0.09	2.91 ± 0.003 (2.91)
Total intron length per gene leave out 1st intron (bp)	911 ± 13 (655)	743 ± 11 (512)	168	834 ± 7.8 (590)
Average intron length leave out 1st intron (bp)	158 ± 1.69 (123)	157 ± 1.97 (115)	1 (n.s.)	158 ± 1.18 (119)
Log(Total intron length per gene leave out 1st intron)	2.78 ± 0.006 (2.82)	2.68 ± 0.006 (2.71)	0.1	2.73 ± 0.004 (2.77)

Dataset 2 **Analysis VI.**	**Rice MPSS**			
	Highly expressed genes (n=5757)	**Lowly expressed genes (n=5757)**	**Difference**	**All expressed Genes (n=14392)**
Number of introns per gene	7.5 ± 0.08 (6)	6.0 ± 0.06 (4)	1.5	6.8 ± 0.05 (5)
Average intron length per gene (bp)	433 ± 3.6 (364)	433 ± 4.1 (353)	0 (n.s.)	429 ± 2.4 (356)
Average exon length per gene (bp)	209 ± 1.9 (164)	253 ± 2.5 (194)	-44	231 ± 1.4 (178)
Total intron length per gene (bp)	2771 ± 28 (2270)	2285 ± 26 (1799)	486	2541 ± 17 (2073)
Total CDS per gene (bp)	1529 ± 15 (1263)	1458 ± 12 (1245)	71	1506 ± 8.6 (1260)
Length of primary transcript (bp)	4703 ± 37 (4042)	3964 ± 33 (3406)	739	4366 ± 22 (3772)
Protein length (aa)	509 ± 4.9 (420)	485 ± 4.2 (414)	24	501 ± 2.9 (419)
Intron density per kb CDS	5.56 ± 0.039 (5.16)	4.78 ± 0.039 (4.19)	0.78	5.17 ± 0.025 (4.70)
Intron density per kb primary transcript	1.61 ± 0.009 (1.52)	1.57 ± 0.010 (1.46)	0.04 (n.s.)	1.59 ± 0.006 (1.49)
Intergenic spacer (bp)	7340 ± 78 (5804)	7256 ± 71 (5834)	84 (n.s.)	7322 ± 47 (5826)
Log(primary transcript length)	3.61 ± 0.003 (3.61)	3.53 ± 0.003 (3.53)	0.08	3.57 ± 0.002 (3.58)
Log(Total intron length per gene)	3.33 ± 0.004 (3.36)	3.21 ± 0.005 (3.26)	0.12	3.28 ± 0.003 (3.32)
Total intron length per gene leave out 1st intron (bp)	2192 ± 27 (1701)	1754 ± 25 (1220)	438	1985 ± 16 (1480)
Average intron length leave out 1st intron (bp)	405 ± 4.7 (314)	410 ± 4.9 (303)	-5 (n.s.)	403 ± 3.0 (308)
Log(Total intron length per gene leave out 1st intron)	3.16 ± 0.006 (3.23)	3.01 ± 0.007 (3.01)	0.15	3.09 ± 0.004 (3.17)

Dataset 3 **Analysis VI.**	**Arabidopsis MA**			
	Highly expressed genes (n=4922)	**Lowly expressed genes (n=4922)**	**Difference**	**All expressed Genes (n=12305)**
Number of introns per gene	7.3 ± 0.08 (6)	5.7 ± 0.06 (4)	1.6	6.6 ± 0.05 (5)
Average intron length per gene (bp)	177 ± 1.5 (151)	178 ± 2.0 (131)	-1 (n.s.)	175 ± 1.1 (140)
Average exon length per gene (bp)	203 ± 2.0 (158)	256 ± 2.3 (212)	-53	229 ± 1.4 (184)
Total intron length per gene (bp)	1150 ± 13 (908)	901 ± 11 (677)	249	1041 ± 7.9 (811)
Total CDS per gene (bp)	1453 ± 15 (1218)	1462 ± 13 (1251)	-9 (n.s.)	1484 ± 8.8 (1260)
Length of primary transcript (bp)	3014 ± 25 (2591)	2545 ± 20 (2230)	469	2827 ± 15 (2446)
Protein length (aa)	483 ± 5.0 (405)	486 ± 4.2 (416)	-3 (n.s.)	494 ± 3.0 (419)
Intron density per kb CDS	5.62 ± 0.041 (5.34)	4.34 ± 0.036 (3.81)	1.28	4.98 ± 0.025 (4.50)
Intron density per kb primary transcript	2.35 ± 0.013 (2.31)	2.27 ± 0.015 (2.15)	0.08	2.32 ± 0.009 (2.24)
Intergenic spacer (bp)	3493 ± 69 (2138)	5416 ± 157 (3078)	-1923	4303 ± 72 (2557)
Log(primary transcript length)	3.41 ± 0.003 (3.41)	3.35 ± 0.003 (3.35)	0.06	3.39 ± 0.002 (3.39)
Log(Total intron length per gene)	2.96 ± 0.004 (2.96)	2.83 ± 0.005 (2.83)	0.13	2.90 ± 0.003 (2.91)
Total intron length per gene leave out 1st intron (bp)	913 ± 13 (663)	683 ± 10 (458)	230	814 ± 7.6 (578)
Average intron length leave out 1st intron (bp)	161 ± 1.6 (128)	166 ± 2.2 (112)	-5 (n.s.)	161 ± 1.2 (120)
Log(Total intron length per gene leave out 1st intron)	2.79 ± 0.006 (2.82)	2.64 ± 0.006 (2.66)	0.15	2.72 ± 0.004 (2.76)

Analysis VII – Leave out first 4 introns, repeat Analysis I.

On the basis of the data of **Analysis V** (eg. genes with ≥ 2 introns), leave out the 5-prime first 4 introns of all genes, only considering those that are still intron-containing genes afterwards (eg. genes with ≥ 5 introns).

Data is sorted by the **average expression rank** (**rE**, see article) in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared as described in the article.

Both 'Total intron length per gene' (eg. the sum of 1st intron till the end intron) and 'Total intron length per gene after leaving out first 4 introns per gene' (eg. the sum of 4th intron till the end intron) were compared between highly and lowly expressed quantiles, as well as the log-transformed these two parameters per gene.

Tables show the data of 40% quantile.

Dataset 1: Arabidopsis MPSS, 6988 genes in total, 2795 genes in 40% quantile.

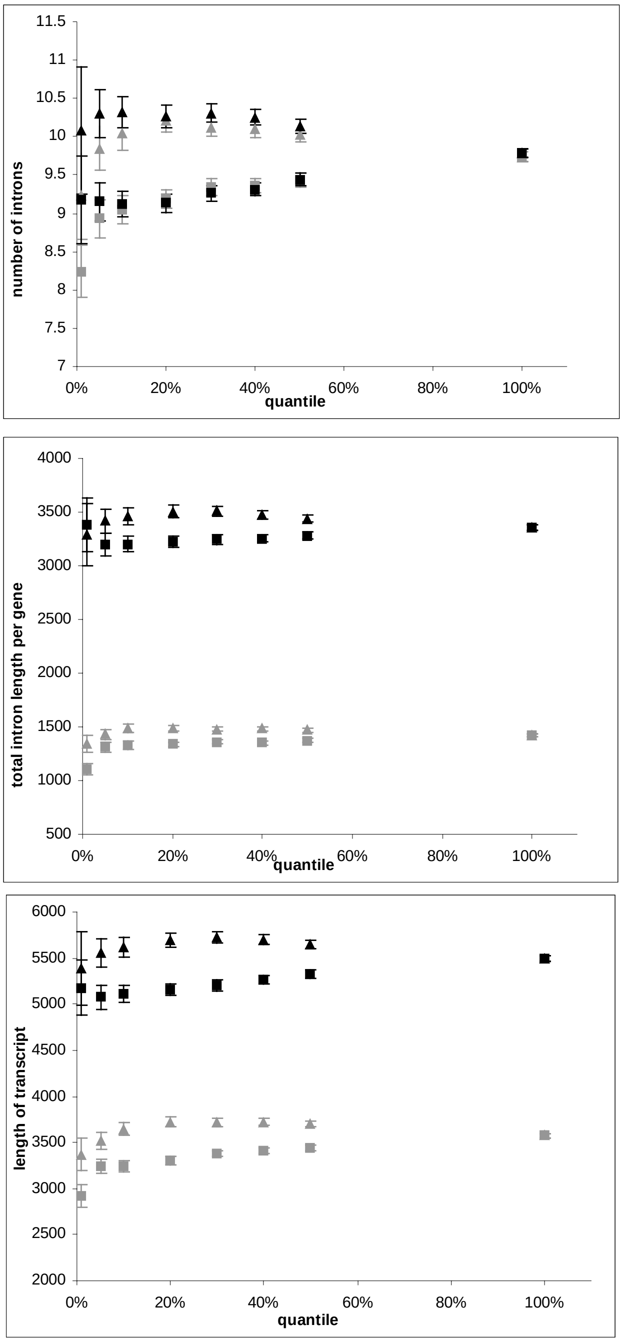
Dataset 2: Rice MPSS, 8212 genes in total, 3285 genes in 40% quantile

Dataset 3: Arabidopsis MA, 6805 genes in total, 2722 genes in 40% quantile

This analysis confirmed that the positive correlation between expression characteristics and intron length is not restricted 5' proximal introns only.

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.

Figure s5.7



Analysis VII. Leave out first 4 introns, repeat Analysis I.

Dataset 1 **Analysis VII.**	**Arabidopsis MPSS**			
	Highly expressed genes (n=2795)	**Lowly expressed genes (n=2795)**	**Difference**	**All expressed Genes (n=6988)**
Number of introns per gene	10.1 ± 0.11 (8)	9.4 ± 0.10 (8)	0.7	9.7 ± 0.07 (8)
Average intron length per gene (bp)	149 ± 1.1 (136)	147 ± 1.4 (128)	2 (n.s.)	148 ± 0.76 (133)
Average exon length per gene (bp)	170 ± 1.7 (145)	184 ± 1.8 (158)	-14	177 ± 1.1 (152)
Total intron length per gene (bp)	1484 ± 19 (1204)	1351 ± 17 (1087)	133	1422 ± 12 (1146)
Total CDS per gene (bp)	1798 ± 22 (1509)	1785 ± 19 (1539)	13 (n.s.)	1796 ± 13 (1521)
Length of primary transcript (bp)	3721 ± 36 (3231)	3409 ± 32 (2986)	312	3573 ± 22 (3103)
Protein length (aa)	598 ± 7.3 (502)	594 ± 6.4 (512)	4 (n.s.)	597 ± 4.4 (506)
Intron density per kb CDS	6.38 ± 0.049 (6.20)	5.89 ± 0.047 (5.62)	0.49	6.14 ± 0.031 (5.90)
Intron density per kb primary transcript	2.77 ± 0.015 (2.76)	2.84 ± 0.016 (2.79)	-0.07 (n.s.)	2.80 ± 0.010 (2.77)
Intergenic spacer (bp)	3566 ± 93 (2213)	3763 ± 104 (2416)	-197 (n.s.)	3625 ± 60 (2325)
Log(primary transcript length)	3.53 ± 0.004 (3.51)	3.49 ± 0.003 (3.48)	0.04	3.51 ± 0.002 (3.49)
Log(Total intron length per gene)	3.10 ± 0.004 (3.08)	3.06 ± 0.005 (3.04)	0.04	3.08 ± 0.003 (3.06)
Total intron length per gene leave out first 4 introns (bp)	825 ± 18 (535)	711 ± 15 (451)	114	771 ± 11 (493)
Average intron length leave out first 4 introns (bp)	132 ± 1.5 (110)	131 ± 1.6 (105)	1 (n.s.)	132 ± 1.0 (108)
Log(Total intron length per gene leave out first 4 introns)	2.71 ± 0.008 (2.73)	2.64 ± 0.008 (2.65)	0.07	2.68 ± 0.005 (2.69)

Dataset 2 **Analysis VII.**	**Rice MPSS**			
	Highly expressed genes (n=3285)	**Lowly expressed genes (n=3285)**	**Difference**	**All expressed Genes (n=8212)**
Number of introns per gene	10.3 ± 0.10 (9)	9.3 ± 0.09 (8)	1	9.8 ± 0.06 (8)
Average intron length per gene (bp)	357 ± 3.0 (326)	366 ± 3.4 (323)	-9 (n.s.)	360 ± 2.0 (323)
Average exon length per gene (bp)	168 ± 1.6 (142)	180 ± 1.7 (153)	-12	174 ± 1.1 (148)
Total intron length per gene (bp)	3477 ± 40 (2915)	3256 ± 38 (2706)	221	3360 ± 25 (2803)
Total CDS per gene (bp)	1805 ± 21 (1494)	1751 ± 18 (1515)	54 (n.s.)	1783 ± 13 (1503)
Length of primary transcript (bp)	5698 ± 53 (4967)	5267 ± 48 (4621)	431	5488 ± 32 (4804)
Protein length (aa)	601 ± 7.0 (497)	583 ± 6.2 (504)	18 (n.s.)	593 ± 4.2 (500)
Intron density per kb CDS	6.56 ± 0.049 (6.36)	6.16 ± 0.048 (5.81)	0.40	6.35 ± 0.031 (6.05)
Intron density per kb primary transcript	1.89 ± 0.011 (1.82)	1.90 ± 0.012 (1.81)	-0.01 (n.s.)	1.89 ± 0.007 (1.81)
Intergenic spacer (bp)	7245 ± 99 (5774)	7110 ± 93 (5682)	135 (n.s.)	7166 ± 61 (5712)
Log(primary transcript length)	3.71 ± 0.003 (3.70)	3.67 ± 0.004 (3.67)	0.04	3.69 ± 0.002 (3.68)
Log(Total intron length per gene)	3.47 ± 0.004 (3.47)	3.43 ± 0.005 (3.43)	0.04	3.45 ± 0.003 (3.45)
Total intron length per gene leave out first 4 introns (bp)	1852 ± 36 (1230)	1619 ± 32 (1036)	233	1740 ± 22 (1139)
Average intron length leave out first 4 introns (bp)	309 ± 4.0 (257)	323 ± 4.7 (252)	-14 (n.s.)	317 ± 2.8 (254)
Log(Total intron length per gene leave out first 4 introns)	3.04 ± 0.009 (3.09)	2.97 ± 0.009 (3.02)	0.07	3.00 ± 0.005 (3.06)

Dataset 3 **Analysis VII.**	**Arabidopsis MA**			
	Highly expressed genes (n=2722)	**Lowly expressed genes (n=2722)**	**Difference**	**All expressed Genes (n=6805)**
Number of introns per gene	10 ± 0.11 (8)	9.2 ± 0.09 (8)	0.8	9.7 ± 0.07 (8)
Average intron length per gene (bp)	150 ± 1.0 (138)	149 ± 1.6 (123)	1 (n.s.)	149 ± 0.8 (133)
Average exon length per gene (bp)	165 ± 1.7 (140)	190 ± 1.8 (168)	-25	178 ± 1.1 (153)
Total intron length per gene (bp)	1484 ± 19 (1205)	1323 ± 17 (1075)	161	1415 ± 12 (1148)
Total CDS per gene (bp)	1732 ± 22 (1455)	1818 ± 19 (1581)	-86 (n.s.)	1794 ± 13 (1530)
Length of primary transcript (bp)	3635 ± 36 (3160)	3358 ± 30 (2980)	277	3532 ± 22 (3089)
Protein length (aa)	576 ± 7.3 (484)	605 ± 6.2 (526)	-29 (n.s.)	597 ± 4.4 (509)
Intron density per kb CDS	6.57 ± 0.050 (6.41)	5.67 ± 0.047 (5.30)	0.90	6.11 ± 0.031 (5.85)
Intron density per kb primary transcript	2.81 ± 0.015 (2.80)	2.85 ± 0.017 (2.79)	-0.04 (n.s.)	2.83 ± 0.010 (2.80)
Intergenic spacer (bp)	3384 ± 90 (2105)	4778 ± 154 (2867)	-1394	3961 ± 77 (2408)
Log(primary transcript length)	3.52 ± 0.004 (3.50)	3.49 ± 0.004 (3.47)	0.03	3.51 ± 0.002 (3.49)
Log(Total intron length per gene)	3.10 ± 0.004 (3.08)	3.05 ± 0.005 (3.03)	0.05	3.08 ± 0.003 (3.06)
Total intron length per gene leave out first 4 introns (bp)	814 ± 18 (528)	676 ± 14 (439)	138	760 ± 11 (492)
Average intron length leave out first 4 introns (bp)	130 ± 1.2 (110)	134 ± 1.9 (102)	-4 (n.s.)	132 ± 1.0 (107)
Log(Total intron length per gene leave out first 4 introns)	2.70 ± 0.008 (2.72)	2.63 ± 0.008 (2.64)	0.07	2.67 ± 0.005 (2.69)

Analysis VIII – Remove genes undergo alternative splicing, repeat Analysis I.

Figure s5.8

This analysis is similar as **Analysis I**, with the only exception of leaving out genes known to undergo alternative splicing from the dataset.

Data is sorted by the **average expression rank** (**rE**, see article) in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared as described in the article.

Tables show the data of 40% quantile.

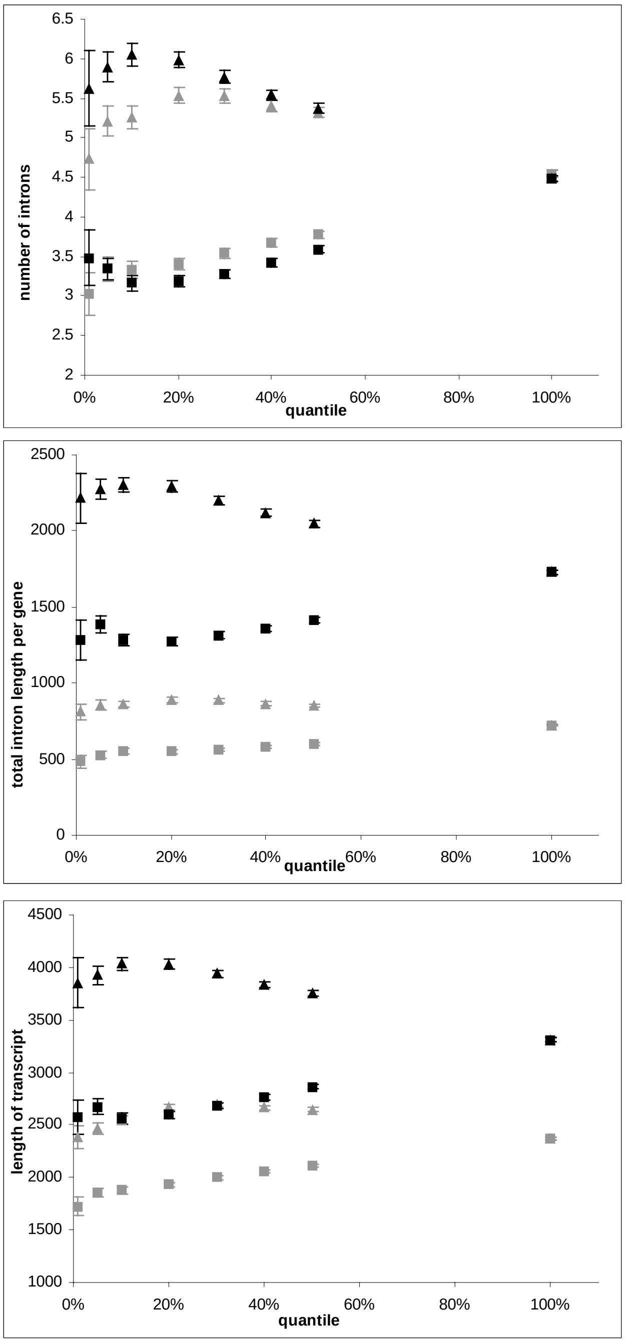
Dataset 1: Arabidopsis MPSS, 16461 genes in total, 6584 genes in 40% quantile.

Dataset 2: Rice MPSS, 19228 genes in total, 7691 genes in 40% quantile

Dataset 3: Arabidopsis MA, 17761 genes in total, 7104 genes in 40% quantile

This analysis cleared the doubts on the impact to the results by possible ambiguously combing expression values of alternatively spliced variants to the largest spliced variants.

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.



Analysis VIII. Only genes not known to undergo alternative splicing, repeat Analysis I.

Dataset 1	**Arabidopsis MPSS**			
Analysis VIII.	**Highly expressed genes (n=6584)**	**Lowly expressed genes (n=6584)**	**Difference**	**All expressed Genes (n=16461)**
Number of introns per gene	5.4 ± 0.07 (4)	3.7 ± 0.06 (2)	1.7	4.5 ± 0.04 (3)
Average intron length per gene (bp)	164 ± 2.0 (132)	139 ± 2.2 (103)	25	151 ± 1.3 (119)
Average exon length per gene (bp)	391 ± 5.5 (222)	491 ± 6.1 (307)	-100	446 ± 3.8 (261)
Total intron length per gene (bp)	866 ± 12 (672)	583 ± 9.5 (343)	283	725 ± 6.8 (510)
Total CDS per gene (bp)	1421 ± 13 (1191)	1284 ± 10 (1113)	137	1363 ± 7.4 (1161)
Length of primary transcript (bp)	2667 ± 22 (2263)	2059 ± 18 (1775)	608	2376 ± 13 (2030)
Protein length (aa)	473 ± 4.3 (396)	427 ± 3.5 (370)	46	453 ± 2.5 (386)
Intron density per kb CDS	3.97 ± 0.040 (3.50)	2.84 ± 0.035 (2.10)	1.13	3.39 ± 0.024 (2.76)
Intron density per kb primary transcript	1.75 ± 0.015 (1.75)	1.47 ± 0.016 (1.29)	0.28	1.61 ± 0.010 (1.53)
Intergenic spacer (bp)	3693 ± 56 (2408)	4430 ± 105 (2780)	-737	4050 ± 51 (2596)
Log(primary transcript length)	3.35 ± 0.003 (3.35)	3.22 ± 0.004 (3.25)	0.13	3.29 ± 0.002 (3.31)
Log(Total intron length per gene)	2.40 ± 0.014 (2.83)	1.99 ± 0.015 (2.53)	0.41	2.20 ± 0.009 (2.71)
Total number of introns	35586	24182	11404	74787
Total length of introns (bp)	5698612	3840996	1857616	11936582
Average intron length per group	160	159	1	160

Dataset 2	**Rice MPSS**			
Analysis VIII.	**Highly expressed genes (n=7691)**	**Lowly expressed genes (n=7691)**	**Difference**	**All expressed Genes (n=19228)**
Number of introns per gene	5.5 ± 0.07 (4)	3.4 ± 0.05 (2)	2.1	4.5 ± 0.04 (3)
Average intron length per gene (bp)	416 ± 4.8 (330)	355 ± 4.9 (243)	61	384 ± 3.1 (293)
Average exon length per gene (bp)	353 ± 4.3 (215)	486 ± 5.8 (306)	-133	424 ± 3.3 (258)
Total intron length per gene (bp)	2119 ± 24 (1727)	1359 ± 21 (748)	760	1730 ± 14 (1249)
Total CDS per gene (bp)	1403 ± 12 (1164)	1246 ± 9.9 (1062)	157	1339 ± 7.0 (1122)
Length of primary transcript (bp)	3835 ± 32 (3273)	2765 ± 26 (2189)	1070	3309 ± 19 (2730)
Protein length (aa)	467 ± 4.0 (387)	415 ± 3.3 (353)	52	445 ± 2.3 (373)
Intron density per kb CDS	4.11 ± 0.037 (3.60)	2.94 ± 0.034 (2.07)	1.17	3.49 ± 0.023 (2.78)
Intron density per kb primary transcript	1.27 ± 0.010 (1.23)	1.06 ± 0.010 (0.96)	0.21	1.16 ± 0.007 (1.09)
Intergenic spacer (bp)	7575 ± 67 (6013)	7435 ± 62 (6065)	140 (n.s.)	7540 ± 41 (6073)
Log(primary transcript length)	3.48 ± 0.004 (3.52)	3.31 ± 0.004 (3.34)	0.17	3.39 ± 0.003 (3.44)
Log(Total intron length per gene)	2.78 ± 0.014 (3.24)	2.28 ± 0.016 (2.87)	0.5	2.53 ± 0.009 (3.10)
Total number of introns	42611	26357	16254	86144
Total length of introns (bp)	16297572	10450438	5847134	33269772
Average intron length per group	382	396	-14	386

Dataset 3	**Arabidopsis MA**			
Analysis VIII.	**Highly expressed genes (n=7104)**	**Lowly expressed genes (n=7104)**	**Difference**	**All expressed Genes (n=17761)**
Number of introns per gene	5.4 ± 0.07 (4)	3.1 ± 0.05 (2)	2.3	4.3 ± 0.04 (3)
Average intron length per gene (bp)	165 ± 1.9 (135)	140 ± 2.3 (99)	25	151 ± 1.3 (118)
Average exon length per gene (bp)	369 ± 5.0 (209)	500 ± 6.0 (312)	-131	440 ± 3.6 (259)
Total intron length per gene (bp)	871 ± 11 (684)	509 ± 8.1 (280)	362	697 ± 6.3 (483)
Total CDS per gene (bp)	1358 ± 12 (1131)	1212 ± 9.7 (1053)	146	1312 ± 6.9 (1110)
Length of primary transcript (bp)	2612 ± 21 (2214)	1860 ± 15 (1596)	752	2269 ± 12 (1938)
Protein length (aa)	452 ± 4.0 (376)	403 ± 3.2 (350)	49	436 ± 2.3 (369)
Intron density per kb CDS	4.16 ± 0.039 (3.76)	2.59 ± 0.032 (1.94)	1.57	3.36 ± 0.023 (2.74)
Intron density per kb primary transcript	1.79 ± 0.015 (1.79)	1.41 ± 0.015 (1.22)	0.38	1.61 ± 0.010 (1.52)
Intergenic spacer (bp)	3707 ± 57 (2372)	5483 ± 126 (3117)	-1776	4515 ± 59 (2723)
Log(primary transcript length)	3.34 ± 0.003 (3.35)	3.17 ± 0.004 (3.20)	0.17	3.27 ± 0.002 (3.29)
Log(Total intron length per gene)	2.41 ± 0.013 (2.84)	1.90 ± 0.015 (2.45)	0.51	2.16 ± 0.009 (2.68)
Total number of introns	38325	22083	16242	76888
Total length of introns (bp)	6185627	3613269	2572358	12383251
Average intron length per group	161	163	-2	161

Analysis IX – Restrict to genes having both 5’ & 3’ UTRs, repeat Analysis I.

Figure s5.9

This analysis is similar as **Analysis I**, but only considering those genes that have both 5’ and 3’ UTR annotations.

Data is sorted by the **average expression rank** (**rE**, see article) in an ascending order. Highly and lowly expressed genes in equal quantiles from top and bottom list are compared as described in the article.

The parameters 5’ UTRs, 3’ UTRs and the sum of both 5’ and 3’ UTRs are compared between highly and lowly expressed quantiles and listed here.

Tables show the data of 40% quantile.

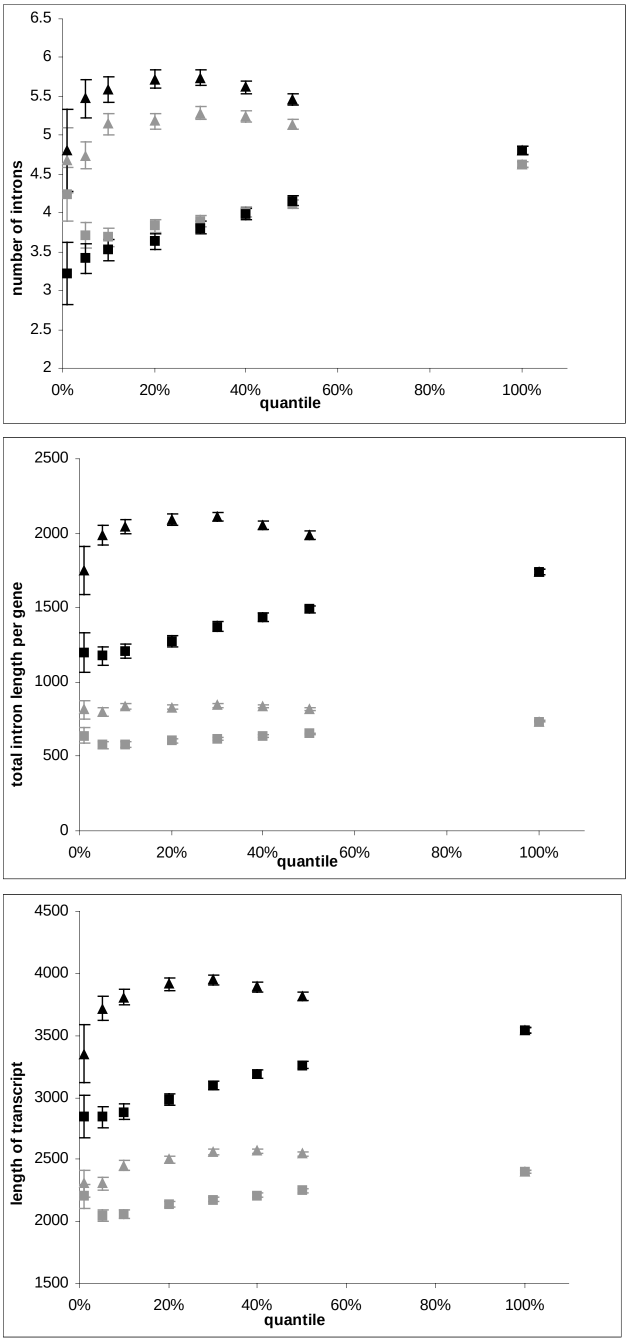
Dataset 1: Arabidopsis MPSS, 12687 genes in total, 5075 genes in 40% quantile.

Dataset 2: Rice MPSS, 9836 genes in total, 3934 genes in 40% quantile

Dataset 3: Arabidopsis MA, 11952 genes in total, 4781 genes in 40% quantile

This analysis looking at length of UTRs alone confirms that highly expressed genes always have longer UTRs than lowly expressed genes. UTRs were coordinately added into the parameter ‘Length of primary transcript (bp)’.

Highly expressed genes have significantly more, longer introns and larger transcripts than lowly expressed genes.



Analysis IX. Only genes with both 5’ and 3’ UTR annotations, repeat Analysis I.

Dataset 1. Analysis IX.	**Arabidopsis MPSS**			
	Highly expressed genes (n=5075)	**Lowly expressed genes (n=5075)**	**Difference**	**All expressed Genes (n=12687)**
Number of introns per gene	5.2 ± 0.07 (4)	4.0 ± 0.06 (3)	1.2	4.6 ± 0.04 (3)
Average intron length per gene (bp)	168 ± 2.2 (136)	147 ± 2.3 (114)	21	157 ± 1.4 (125)
Average exon length per gene (bp)	346 ± 5.5 (202)	420 ± 6.1 (256)	-74	386 ± 3.7 (229)
Total intron length per gene (bp)	836 ± 11 (683)	634 ± 10 (449)	202	735 ± 6.8 (580)
Total CDS per gene (bp)	1268 ± 11 (1104)	1177 ± 9.5 (1080)	91	1229 ± 6.5 (1101)
Length of primary transcript (bp)	2567 ± 20 (2259)	2211 ± 18 (1991)	356	2398 ± 12 (2140)
Protein length (aa)	422 ± 3.7 (367)	391 ± 3.2 (359)	31	409 ± 2.2 (366)
Intron density per kb CDS	4.30 ± 0.046 (3.87)	3.38 ± 0.044 (2.74)	0.92	3.82 ± 0.029 (3.31)
Intron density per kb primary transcript	1.79 ± 0.017 (1.81)	1.52 ± 0.017 (1.40)	0.27	1.65 ± 0.011 (1.61)
Intergenic spacer (bp)	3645 ± 63 (2328)	4219 ± 100 (2678)	-574	3889 ± 51 (2519)
Log(primary transcript length)	3.35 ± 0.003 (3.36)	3.28 ± 0.003 (3.30)	0.07	3.32 ± 0.002 (3.33)
Log(Total intron length per gene)	2.45 ± 0.015 (2.83)	2.13 ± 0.017 (2.65)	0.32	2.29 ± 0.010 (2.76)
5’ UTRs (bp)	144 ± 1.8 (107)	123 ± 1.7 (87)	21	136 ± 1.1 (98)
3’ UTRs (bp)	238 ± 1.6 (220)	212 ± 1.8 (190)	26	226 ± 1.1 (206)
Sum(UTRs) (bp)	382 ± 2.5 (344)	336 ± 2.6 (296)	46	361 ± 1.6 (323)

Dataset 2 Analysis IX.	**Rice MPSS**			
	Highly expressed genes (n=3934)	**Lowly expressed genes (n=3934)**	**Difference**	**All expressed Genes (n=9836)**
Number of introns per gene	5.6 ± 0.08 (4)	4.0 ± 0.07 (2)	1.6	4.8 ± 0.05 (3)
Average intron length per gene (bp)	423 ± 6.3 (335)	337 ± 6.3 (243)	86	379 ± 4.1 (293)
Average exon length per gene (bp)	303 ± 5.2 (186)	457 ± 7.7 (283)	-154	380 ± 4.2 (227)
Total intron length per gene (bp)	2057 ± 27 (1833)	1433 ± 26 (1007)	624	1738 ± 17 (1476)
Total CDS per gene (bp)	1250 ± 12 (1089)	1237 ± 12 (1107)	13 (n.s.)	1246 ± 7.5 (1103)
Length of primary transcript (bp)	3890 ± 35 (3520)	3188 ± 33 (2798)	702	3539 ± 22 (3193)
Protein length (aa)	416 ± 4.1 (362)	411 ± 3.9 (368)	5 (n.s.)	414 ± 2.5 (367)
Intron density per kb CDS	4.74 ± 0.056 (4.27)	3.40 ± 0.053 (2.43)	1.34	4.05 ± 0.035 (3.33)
Intron density per kb primary transcript	1.28 ± 0.013 (1.23)	1.03 ± 0.014 (0.92)	0.25	1.16 ± 0.009 (1.09)
Intergenic spacer (bp)	7334 ± 93 (5815)	7372 ± 92 (5738)	-38 (n.s.)	7368 ± 58 (5834)
Log(primary transcript length)	3.52 ± 0.004 (3.55)	3.41 ± 0.005 (3.45)	0.11	3.45 ± 0.003 (3.50)
Log(Total intron length per gene)	2.88 ± 0.017 (3.26)	2.35 ± 0.022 (3.00)	0.53	2.62 ± 0.013 (3.17)
5’ UTRs (bp)	189 ± 3.4 (130)	164 ± 3.3 (103)	25	178 ± 2.2 (116)
3’ UTRs (bp)	395 ± 4.4 (330)	354 ± 4.9 (285)	41	377 ± 3.0 (309)
Sum(UTRs) (bp)	583 ± 5.8 (489)	518 ± 6.2 (415)	65	555 ± 3.9 (455)
Log(sum(UTRs))	2.71 ± 0.003 (2.69)	2.64 ± 0.004 (2.62)	0.07	2.68 ± 0.002 (2.66)

Dataset 3 Analysis IX.	**Arabidopsis MA**			
	Highly expressed genes (n=4781)	**Lowly expressed genes (n=4781)**	**Difference**	**All expressed Genes (n=11952)**
Number of introns per gene	5.1 ± 0.07 (4)	3.7 ± 0.06 (2)	1.4	4.6 ± 0.04 (3)
Average intron length per gene (bp)	171 ± 2.2 (140)	144 ± 2.5 (109)	27	157 ± 1.5 (125)
Average exon length per gene (bp)	335 ± 5.6 (190)	442 ± 6.5 (279)	-107	387 ± 3.8 (231)
Total intron length per gene (bp)	830 ± 11 (696)	578 ± 9.8 (377)	252	726 ± 7.0 (573)
Total CDS per gene (bp)	1193 ± 11 (1038)	1182 ± 10 (1077)	11 (n.s.)	1218 ± 6.8 (1086)
Length of primary transcript (bp)	2472 ± 20 (2176)	2126 ± 18 (1882)	346	2357 ± 12 (2088)
Protein length (aa)	397 ± 3.7 (345)	393 ± 3.3 (358)	4 (n.s.)	405 ± 2.2 (361)
Intron density per kb CDS	4.47 ± 0.049 (4.09)	3.04 ± 0.042 (2.38)	1.43	3.80 ± 0.030 (3.25)
Intron density per kb primary transcript	1.79 ± 0.018 (1.82)	1.43 ± 0.018 (1.28)	0.36	1.64 ± 0.011 (1.59)
Intergenic spacer (bp)	3591 ± 69 (2244)	4284 ± 75 (2921)	-693	3876 ± 44 (2538)
Log(primary transcript length)	3.33 ± 0.003 (3.34)	3.26 ± 0.004 (3.27)	0.07	3.31 ± 0.002 (3.32)
Log(Total intron length per gene)	2.43 ± 0.016 (2.84)	2.05 ± 0.018 (2.58)	0.38	2.27 ± 0.011 (2.76)
5’ UTRs (bp)	136 ± 1.7 (103)	116 ± 1.7 (80)	20	129 ± 1.1 (94)
3’ UTRs (bp)	236 ± 1.5 (220)	201 ± 1.6 (185)	35	219 ± 1.0 (203)
Sum(UTRs) (bp)	373 ± 2.4 (338)	317 ± 2.5 (284)	56	349 ± 1.6 (315)
Log(sum(UTRs))	2.54 ± 0.003 (2.53)	2.45 ± 0.003 (2.45)	0.09	2.50 ± 0.002 (2.50)

Second part: *Information about expression libraries*

Library information for Arabidopsis MPSS:

Code	Title
CAF	Callus - actively growing, classic MPSS
INF	Infloresence - mixed stage, immature buds, classic MPSS
LEF	Leaves - 21 day, untreated, classic MPSS
ROF	Root – 21 day, untreated, classic MPSS
SIF	Silique - 24 to 48 hr post-fertilization, classic MPSS
AP1	ap1-10 infloresence - mixed stage, immature buds
AP3	ap3-6 infloresence - mixed stage, immature buds
AGM	Agamous infloresence - mixed stage, immature buds
INS	Infloresence - mixed stage, immature buds
ROS	Root – 21 day, untreated
SAP	sup/ap1 infloresence - mixed stage, immature buds
S04	Leaves, 4 hr after salicylic acid treatment
S52	Leaves, 52 hr after salicylic acid treatment
LES	Leaves - 21 day, untreated

Library information for Rice MPSS:

Code	Title
NCA	35 days - Callus
NCL	14 days - Young leaves stressed in 4C cold for 24h
NCR	14 days - Young roots stressed in 4C cold for 24h
NDL	14 days - Young leaves stressed in drought for 5 days
NDR	14 days - Young roots stressed in drought for 5 days
NGD	10 days - Germinating seedlings grown in dark
NGS	3 days - Germinating seed
NIP	90 days - Immature panicle
NL_avr	60 days - Mature Leaves – averaged over Replicate A,B,C,D
NME	60 days - Meristematic tissue
NOS	Ovary and mature stigma
NPO	Mature Pollen
NR_avr	60 days - Mature Roots – average over Replicate A,B
NSL	14 days - Young leaves stressed in 250 mM NaCl for 24h
NSR	14 days - Young roots stressed in 250 mM NaCl for 24h
NST	60 days - Stem
NYL	14 days - Young leaves
NYR	14 days - Young Roots

Library information for Arabidopsis MA expression data from 15 separate subzones of root (five cell types by three stages); expression data from (Birnbaum et al., 2003).

Code	Title
stele-stage1	Stele stage1
stele-stage2	Stele stage2
stele-stage3	Stele stage3
endo-stage1	endodermis stage1
endo-stage2	endodermis stage2
endo-stage3	endodermis stage3
cortex-endo-stage1	endodermis and cortex stage1
cortex-endo-stage2	endodermis and cortex stage2
cortex-endo-stage3	endodermis and cortex stage3
epidermis-stage1	epidermal atrichoblasts stage1
epidermis-stage2	epidermal atrichoblasts stage2
epidermis-stage3	epidermal atrichoblasts stage3
lat-root-cap-stage1	lateral root cap stage1
lat-root-cap-stage2	lateral root cap stage2
lat-root-cap-stage3	lateral root cap stage3
stage 1, where the root tip reached its full diameter (about 0.15 mm from the root tip); stage 2, where cells transition from being optically dense to a more transparent appearance as they begin longitudinal expansion (about 0.30 mm from the root tip); stage 3, where root hairs were fully elongated (about 0.45 to 2 mm from the root tip)	

Third part: *Additional analysis regard to gametophytic selection issue*

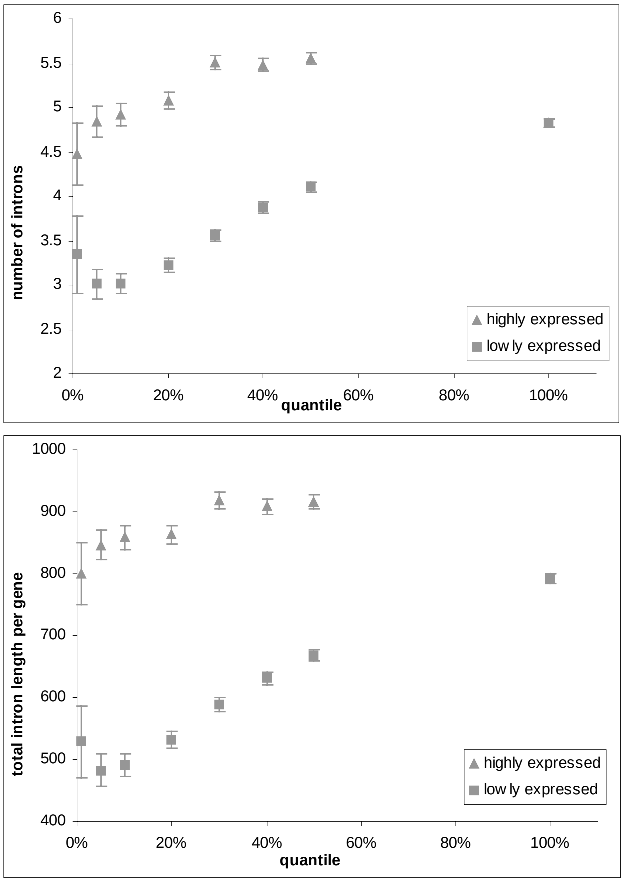
Additional analysis on the data from (Seoighe et al., 2005), using their sporophyte microarray (MA) expression (in 5 libraries: root, leaf, stem, seedling green plant, hypocotyls) and our expression ranking method (using average sporophyte expression revealed the same trend of results, not shown). We want to see whether in Arabidopsis (plant) sporophyte selection for short introns could also be observed as claimed in their pollen study. Our results show that selection for short introns could not be observed in sporophyte.

15390 genes in total are in sporophyte MA expression, 6156 genes in 40% quantile.
We've compared gene characteristics between highly and lowly expressed sporophyte active genes and we obtained the same trend as we described in the article that highly expressed genes have more and larger introns than lowly expressed genes.
Figure show the number of introns and Total intron length per gene between highly and lowly expressed quantiles.
Table shows data in 40% quantile. We analysed gene parameters given in their supplementary data.

This additional analysis confirms that in Arabidopsis (plant) sporophyte, previous proposed selection model in animals or in Arabidopsis pollens could not be observed. Either plant sporophyte does not undergo the same selection or selection gives different results.

Additional Analysis	**(Seoighe et al., 2005), sporophyte MA expression**			
	Highly expressed genes (n=6156)	**Lowly expressed genes (n=6156)**	**Difference**	**All expressed Genes (n=15390)**
Number of introns per gene	5.5 ± 0.07 (4)	3.9 ± 0.06 (2)	1.6	4.8 ± 0.04 (3)
Mean length of introns 5-10 (bp) 0bp for 1-4 introns	59 ± 1.1 (0)	40 ± 1.0 (0)	19	52 ± 0.66 (0)
Total intron length per gene (bp)	908 ± 12 (735)	632 ± 10 (380)	276	792 ± 7.2 (609)
Total CDS per gene (bp)	1642 ± 13 (1443)	1573 ± 12 (1399)	69	1643 ± 7.9 (1455)
Intergenic spacer (bp)	1822 ± 19 (1370)	1958 ± 20 (1511)	-136	1891 ± 12 (1443)

Figure s5.10


number of introns
6
5.5
5
4.5
4
3.5
3
2.5
2
0%
20%
40%
60%
80%
100%
quantile
▲ highly expressed
■ low ly expressed
total intron length per gene
1000
900
800
700
600
500
400
0%
20%
40%
60%
80%
100%
quantile
▲ highly expressed
■ low ly expressed

Chapter 6 Supplementary Materials

I. Geometric mean of the ranks of expressions (geo_rE)

In this method, the same 5 genomes as described in the main text were treated in parallel. In each dataset (genome), we first calculated for each gene the average expression per tissue, then sorted the expression values of all the genes in each tissue (or developmental stage for *C.elegans*) in an ascending order, and assigned consecutive ranks from 1 (lowest expressed) to the total number of genes (highest expressed) to this sorted list. A group of genes having equal values (due to integer MPSS tag counts) all obtain the average rank of this group. Then for each gene, we multiplied all the ranks in all x tissues (eg. rank product) and took the x-th root of this rank product, where x is the number of tissues (or developmental stages) in each dataset. With this treatment, we get a geometric mean of all the ranks of expressions (**geo_rE**) of a gene in all the tissues studied. Each dataset (Arabidopsis, rice, worm, mouse or human) was sorted by geo_rE in an ascending order and subsequently divided into 10 sequential quantiles from low expression to high expression (eg. 0-10%, 10-20%, 20-30% of the population etc.). We compared various gene structural parameters among these 10 quantiles within each genome and among 5 genomes.

The average of each parameter in each quantile was plotted against 10 expression quantiles for each organism in Figure s6.1. Applying this alternative method, we didn’t find any significant difference in the trends of results than in Chapter 6, only in this figure, the data points appeared to be more smoothly in one line than Figure 6.2 in Chapter 6, this might be due to the fine rankings in the first step of this method. Therefore, changing to fine rankings and taking geometric mean of ranks to define high and low expressions do not seem to influence the trends of results we have found and conclusions we have drawn in the main text.

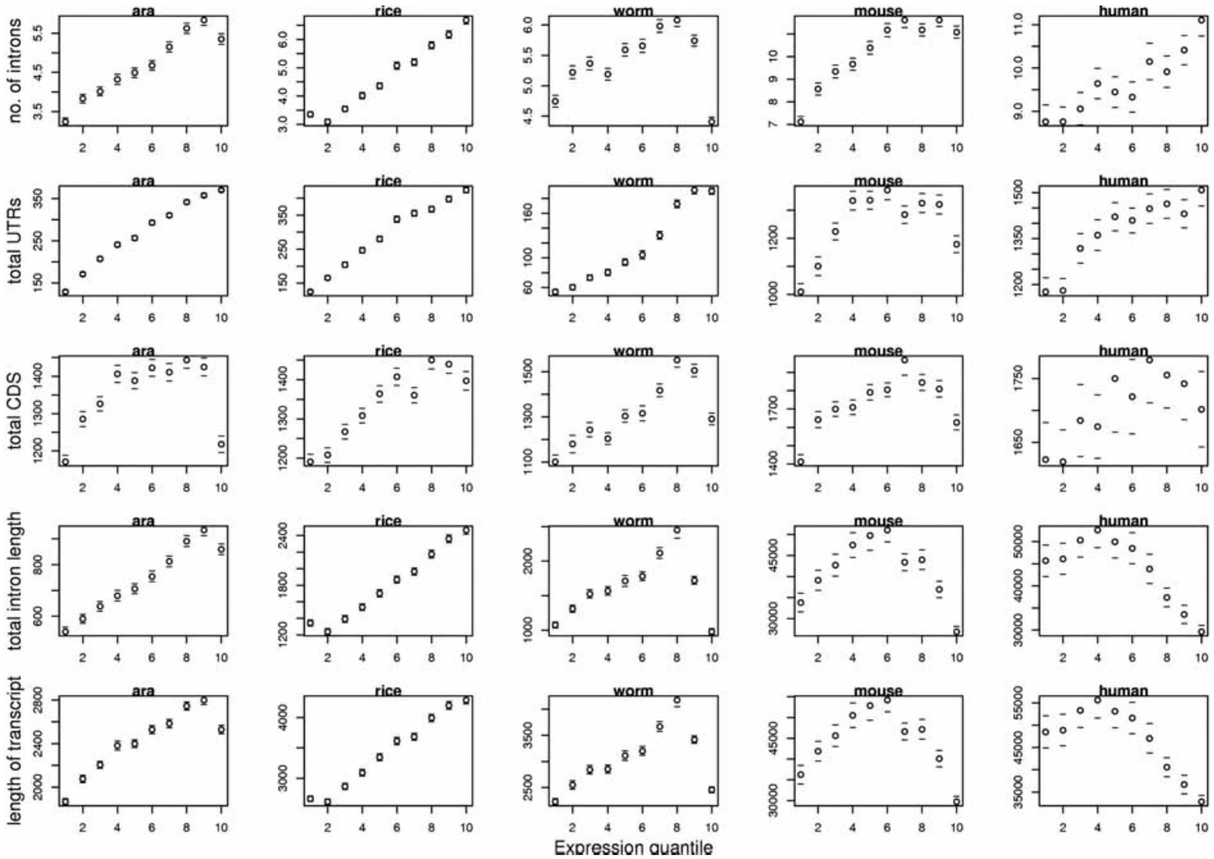


Figure s6.1 The relationship between gene expression level and gene structure. The averages of the five structural parameters of genes (5 rows) in each expression quantile based on geometric means of ranks (**geo_rE**) were plotted against the 10 expression quantiles for each organism. The X-axis depicts the expression from low (1: 1st quantile) to high (10: 10th quantile). The Y-axis depicts the number of introns (1st row), total UTRs (2nd row), total CDS length (3rd row), total intron length (4th row) and the length of transcript (last row). From left to right, first column subfigures are for Arabidopsis, 2nd column for rice, 3rd for worm, 4th for mouse and 5th column is for human. There are 1,839, 2,143, 1,775, 1,100 and 668 genes in each quantile in Arabidopsis, rice, worm, mouse and human respectively.

II. Figure Intron density per kb CDS

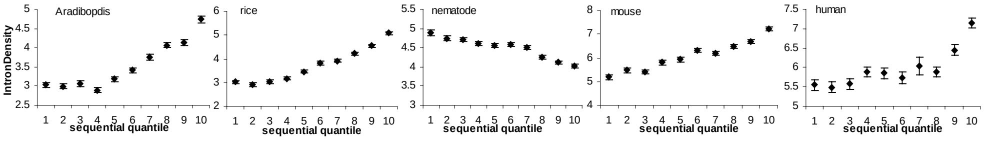


Figure s6.2 Gene expression – gene structure (Intron Density per kb CDS length) relationship in 5 diverse orgasims. The x-axis depicts the expression from low (1: 1st quantile) to high (10: 10th quantile). Y-axis depicts the intron density per kb CDS length.

III. Comparisons of the rE and geo_exrE methods

From the ways the final ranks as indication of expression levels are calculated, we know that the difference between rE and geo_exrE methods mainly lies in that rE method divides sum of ranks over all the tissues studied, while geo_exrE divides sum of ranks (in all expressed tissues) over tissues that are expressed. Due to this difference, we can foresee that genes in the 10 quantiles defined by rE method will not be the same as genes in the 10 quantiles defined by geo_exrE method. In order to illustrate these differences, we plotted the overlaps and outliers between each of the 10 quantiles defined by rE method and geo_exrE method in Figure s6.3 (only for 2 plants and 2 vertebrates). If rE method and geo_exrE method have high correlation with each other, the data points should centre on the diagonal. From Figure s6.3 we can see that only for the (most) high-expressing genes there are high correlations between the two methods. For both low and middle-expressing genes, there are a lot of disputes in grouping into certain expression quantiles by both methods. These disputes are further pinpointed to the different grouping of (mainly) the expression of TS and TS-like genes (Figure s6.4). From Figure s6.4 we can see that rE method groups TS and TS-like genes to low-expressing quantiles, while geo_exrE method groups them to high-expressing quantiles (Figure s6.4, the right column subfigures). Therefore, in Figure s6.4, there are high correlations between rE method and expression breadth, while there are almost no correlations between geo_exrE method and expression breadth. From Figure 6.1 in the main text we have learned that TS and TS-like genes are relatively rather compact in all 5 organisms. Grouping these genes (that are compact) into high-expressing quantiles like geo_exrE method did (Figure s6.4) would decrease the average length parameters of these quantiles, and removing these compact genes from low-expressing quantiles would increase the average length parameters in low-expressing quantiles. These caused the down-shifting (decrease in average length) of the data points in high-expressing quantiles and up-shifting (increase in average length) of the data points in low-expressing quantiles comparing the right-bottom 4 subfigures (total intron length and length of transcript for mouse and human) in Figure 6.2 and Figure 6.3 in Chapter 6. That is why we have seen only decreasing trends for the total intron length and length of transcript in mouse and human in Figure 6.3 comparing to Figure 6.2 in Chapter 6.

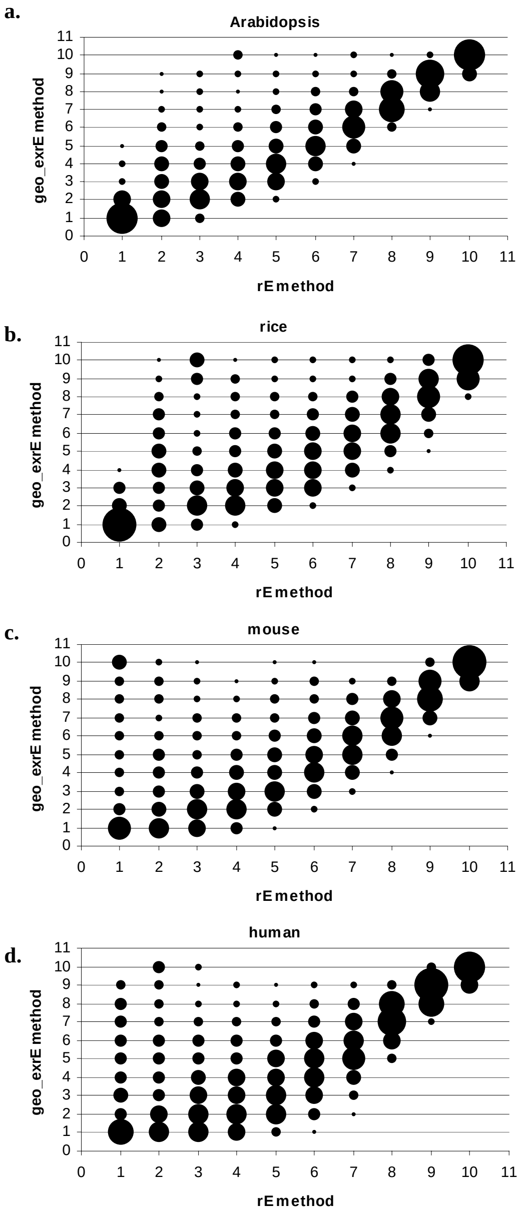


Figure s6.3 Comparisons of the overlaps and outliers of the genes in each of the 10 quantiles between rE method and geo_exrE method in Arabidopsis (**a**), rice (**b**), mouse (**c**) and human (**d**).

On the X-axis are the 10 expression quantiles defined by rE method and on the Y-axis are the 10 expression quantiles defined by geo_exrE method. The sizes of the bubbles denote the number of genes.

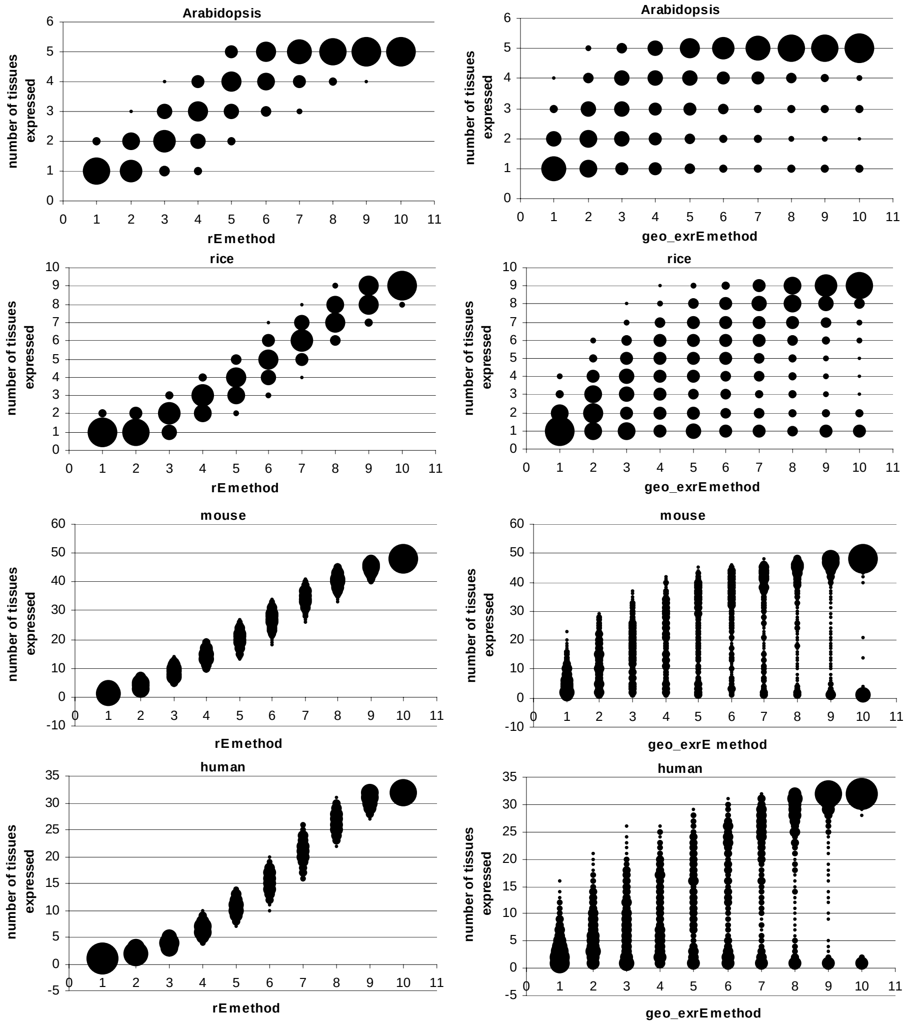


Figure s6.4 Comparisons of the correlations between expression breadth (number of tissues expressed) and expression level defined by rE method and geo_exrE method. On the X-axis are the 10 expression quantiles defined by either rE method (left) or geo_exrE method (right). On the Y-axis are the expression breadths, eg. number of tissues expressed (>0 MPSS tag count). Arabidopsis has 5 tissues, rice has 9 tissues, mouse has 48 tissues and human has 32 tissues. The sizes of the bubbles reprent the number of genes.

IV. Tissue sample information

- The 5 tissues of Arabidopsis are: callus, infloresence, leaves, root, silique.
- The 9 tissues of rice are: callus, panicle, leaves, root, germinating seed and seedling meristem, ovary and stigma, pollen, stem.
- The 48 tissues of mouse are (separated by semicolumn): adrenal; bladder; bone / femur; brain / amygdala; brain (multi parts) + caudate, putamen, medulla + pons; brain / cerebellum; brain / cortical mantle; brain / hippocampus; brain / hypothalamus, preoptic; brain / midbrain; brain / olfactory bulb; brain / olfactory tubercle, prefrontal; brain / thalamus; brown fat; cartilage; cervix / vagina; embryo / E18; ES cells / 129; ES cells / C57BL6; esophagus; eye; heart / aorta; heart / atria; heart / ventricles, septum; kidney / cortex; kidney / medulla; large intestine; liver / left lobe; liver / right lobe; lung; lymph nodes; mammary gland; ovary; pituitary; placenta / E18; prostate; skeletal muscle / thigh; skin / hairy from back; small intestine; spinal cord / entire; spleen; stomach; testis; thymus; thyroid + parathyroid; uterus; uterus (pregnant); white fat / abdomen;
- The 32 tissues of human are: adrenal gland; bladder; bone marrow; brain amygdala; brain caudate; brain cerebellum; brain corpus callosum; fetal brain; brain hypothalmus; brain thalamus; monocytes; peripheral blood lymphocytes; heart; kidney; lung; mammary gland; pancreas; pituitary gland; placenta; prostate; retina; salivary gland; small intestine; spinal cord; spleen; stomach; testis; thymus; thyroid; trachea; colon transversum; uterus.
- The 10 developmental stages for worm embryonic cell are: 0, 23, 41, 53, 66, 83, 101, 122, 143, 186 minutes after the 4-cell embryonic stage.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell. (New York: Garland Science).

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25,** 3389-3402.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell Biochem. **94,** 890-898.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. **33,** D562-566.

Baugh, L.R., Hill, A.A., Claggett, J.M., Hill-Harfe, K., Wen, J.C., Slonim, D.K., Brown, E.L., and Hunter, C.P. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. Development **132,** 1843-1854.

Beckingham, K., and Rubacha, A. (1984). Different chromatin states of the intron$^-$ and type 1 intron$^+$ rRNA genes of *Calliphora erythrocephala*. Chromosoma **90,** 311-316.

Bell, A.C., West, A.G., and Felsenfeld, G. (2001). Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. Science **291,** 447-450.

Belshaw, R., and Bensasson, D. (2006). The rise and falls of introns. Heredity **96,** 208-213.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. (2000). GenBank. Nucleic Acids Res. **28,** 15-18.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., and Rhee, S.Y. (2004). Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol. **135,** 745-755.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at 5' terminus of adenovirus 2 late messenger-RNA. Proc. Natl. Acad. Sci. USA **74,** 3171-3175.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science **306,** 2242-2246.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W., and Benfey, P.N. (2003). A gene expression map of the Arabidopsis root. Science **302,** 1956-1960.

Bode, J., Stengert-Iber, M., Kay, V., Schlake, T., and Dietz-Pfeilstetter, A. (1996). Scaffold/matrix-attached regions: topological switches with multiple regulatory functions. Crit. Rev. Eukaryot. Gene Expr. **6,** 115-138.

Boulikas, T. (1993). Nature of DNA sequences at the attachment regions of genes to the nuclear matrix. J. Cell Biochem. **52,** 14-22.

Boulikas, T. (1995). Chromatin domains and prediction of MAR sequences. Int. Rev. Cytol. **162A,** 279-388.

Brenner, S., Williams, S.R., Vermaas, E.H., Storck, T., Moon, K., McCollum, C., Mao, J.-I., Luo, S., Kirchner, J.J., Eletr, S., DuBridge, R.B., Burcham, T., and Albrecht, G. (2000a). In vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. Proc. Natl. Acad. Sci. USA **97,** 1665-1670.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000b). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. **18,** 630-634.

Brocchieri, L., and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. **33,** 3390-3400.

Burgess-Beusse, B., Farrell, C., Gaszner, M., Litt, M., Mutskov, V., Recillas-Targa, F., Simpson, M., West, A., and Felsenfeld, G. (2002). The insulation of genes from external enhancers and silencing chromatin. Proc. Natl. Acad. Sci. USA **99,** 16433-16437.

Burnette, J.M., Miyamoto-Sato, E., Schaub, M.A., Conklin, J., and Lopez, A.J. (2005). Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. Genetics **170,** 661-674.

Califano, A. (2001). Advances in sequence analysis. Curr. Opin. Struct. Biol. **11,** 330-333.

Carlini, D.B., Chen, Y., and Stephan, W. (2001). The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the Drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. Genetics **159,** 623 - 633.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heisterkamp, S., van Kampen, A., and Versteeg, R. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291,** 1289-1292.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. Nat. Genet. **31,** 415-418.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. Trends Genet. **7,** 145-148.

Cho, G., and Doolittle, R.F. (1997). Intron distribution in ancient paralogs supports random insertion and not random loss. J. Mol. Evol. **44,** 573-584.

Chodaparambil, J.V., Edayathumangalam, R.S., Bao, Y., Park, Y.J., and Luger, K. (2006). Nucleosome structure and function. Ernst Schering Res. Found Workshop **57,** 29-46.

Choi, T., Huang, M., Gorman, C., and Jaenisch, R. (1991). A generic intron increases gene expression in transgenic mice. Mol. Cell. Biol. **11,** 3070-3074.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. **26,** 183-186.

Comeron, J.M. (2004). Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. Genetics **167,** 1293-1304.

Comeron, J.M., and Kreitman, M. (2000). The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics **156,** 1175-1190.

Coughlan, S.J., Agrawal, V., and Meyers, B. (2004). A comparison of global gene expression measurement technologies in *Arabidopsis thaliana*. Comp. Func. Genomics **5,** 245-252.

Cremer, T., Kreth, G., Koester, H., Fink, R.H., Heintzmann, R., Cremer, M., Solovei, I., Zink, D., and Cremer, C. (2000). Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture. Crit. Rev. Eukaryot. Gene Expr. **10,** 179-212.

Darnell, J.E., Jr. (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science **202,** 1257-1260.

de Roos, A.D.G. (2005). Origins of introns based on the definition of exon modules and their conserved interfaces. Bioinformatics **21,** 2-9.

de Souza, S.J. (2003). The emergence of a synthetic theory of intron evolution. Genetica **118,** 117-121.

Devos, K.M., Beales, J., Nagamura, Y., and Sasaki, T. (1999). Arabidopsis-Rice: Will colinearity allow gene prediction across the eudicot-monocot divide? Genome Res. **9,** 825-829.

Dillon, N. (2006). Gene regulation and large-scale chromatin organization in the nucleus. Chromosome Res. **14,** 117-126.

Dillon, N., and Sabbattini, P. (2000). Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. Bioessays **22,** 657-665.

Doolittle, W.F. (1978). Genes in pieces: were they ever together? Nature **272,** 581-582.

Duester, G., Jornvall, H., and Hatfield, G.W. (1986). Intron-dependent evolution of the nucleotide-binding domains within alcohol dehydrogenase and related enzymes. Nucleic Acids Res. **14,** 1931-1941.

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA **96,** 4482-4487.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res **30,** 207-210.

Eisenberg, E., and Levanon, E.Y. (2003). Human housekeeping genes are compact. Trends Genet. **19,** 362-365.

Fiers, M.W.E.J. (2006). Bioinformatics for plant genome annotation. In Applied Bioinformatics, Plant Research International (Wageningen: Wageningen University and Research Center), pp. 129.

Frisch, M., Frech, K., Klingenhoff, A., Cartharius, K., Liebich, I., and Werner, T. (2002). *In silico* prediction of scaffold/matrix attachment regions in large genomic sequences. Genome Res. **12,** 349-354.

Fukuoka, Y., Inaoka, H., and Kohane, I.S. (2004). Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics **5,** 4.

Gaasterland, T., and Oprea, M. (2001). Whole-genome analysis: annotations and updates. Curr. Opin. Struct. Biol. **11,** 377-381.

Gilbert, W. (1978). Why genes in pieces? Nature **271,** 501.

Gilbert, W. (1987). The exon theory of genes. Cold Spring Harb. Symp. Quant. Biol. **52,** 901-905.

Glazko, G.V., Rogozin, I.B., and Glazkov, M.V. (2001). Comparative study and prediction of DNA fragments associated with various elements of the nuclear matrix. Biochim. Biophys. Acta **1517,** 351-364.

Glazko, G.V., Koonin, E.V., Rogozin, I.B., and Shabalina, S.A. (2003). A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet. **19,** 119-124.

Gollub, J., Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J.C., Schroeder, M., Brown, P.O., Botstein, D., and Sherlock, G. (2003). The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. **31,** 94-96.

Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., and Lewontin, R.C. (1999). The structure of genes and genomes. In Mordern genetic analysis, S. Tenney, ed (New York: W. H. Freeman and Company), pp. 23-50.

Haddrill, P., Charlesworth, B., Halligan, D., and Andolfatto, P. (2005). Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Genome Biol. **6,** R67.

Hatton, A.R., Subramaniam, V., and Lopez, A.J. (1998). Generation of alternative ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol. Cell **2,** 787-796.

Heng, H.H., Krawetz, S.A., Lu, W., Bremer, S., Liu, G., and Ye, C.J. (2001). Re-defining the chromatin loop domain. Cytogenet. Cell Genet. **93,** 155-161.

Hershberg, R., Yeger-Lotem, E., and Margalit, H. (2005). Chromosomal organization is shaped by the transcription regulatory network. Trends Genet. **21,** 138-142.

Holmguist, G.P. (1987). Role of replication time in the control of tissue-specific gene expression. Am. J. Hum. Genet. **40,** 151-173.

Hurst, L.D., Williams, E.J., and Pal, C. (2002). Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. **18,** 604-606.

Hurst, L.D., Pal, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5,** 299-310.

Jongeneel, C.V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C.D., Khrebtukova, I., Kuznetsov, D., Stevenson, B.J., Strausberg, R.L., Simpson, A.J.G., and Vasicek, T.J. (2005). An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Res. **15,** 1007-1014.

Kalari, K.R., Casavant, M., Bair, T.B., Keen, H.L., Comeron, J.M., Casavant, T.L., and Scheetz, T.E. (2006). First exons and introns - a survey of GC content and gene structure in the human genome. In Silico Biol. **6,** 0022.

Kazan, K. (2003). Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged. Trends Plant Sci. **8,** 468-471.

Khan, J., Bittner, M.L., Chen, Y., Meltzer, P.S., and Trent, J.M. (1999). DNA microarray technology: the anticipated impact on the study of human disease. Biochim. Biophys. Acta **1423,** 17-28.

Khavkin, E., and Coe, E. (1997). Mapped genomic locations for developmental functions and QTLs reflect concerted groups in maize (*Zea mays* L.). Theor. Appl.Genet. **95,** 343-352.

Kirby, D.A., Muse, S.V., and Stephan, W. (1995). Maintenance of pre-mRNA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92,** 9047 - 9051.

Koonin, E.V. (2006). The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Bio. Direct **1,** 22.

Korbel, J.O., Jensen, L.J., von Mering, C., and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat. Biotechnol. **22,** 911-917.

Kraakman, L.S., Mager, W.H., Maurer, K.T., Nieuwint, R.T., and Planta, R.J. (1989). The divergently transcribed genes encoding yeast ribosomal proteins L46 and S24 are activated by shared RPG-boxes. Nucleic Acids Res. **17,** 9693-9706.

Kruglyak, S., and Tang, H. (2000). Regulation of adjacent yeast genes. Trends Genet **16,** 109-111.

Labrador, M., and Corces, V.G. (2002). Setting the boundaries of chromatin domains and nuclear organization. Cell **111,** 151-154.

Laemmli, U.K., Kas, E., Poljak, L., and Adachi, Y. (1992). Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. Curr. Opin. Genet. Dev. **2,** 275-285.

Le Hir, H., Nott, A., and Moore, M.J. (2003). How introns influence and enhance eukaryotic gene expression. Trends Biochem. Sci. **28,** 215-220.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Res. **14,** 1085-1094.

Lercher, M.J., Blumenthal, T., and Hurst, L.D. (2003). Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. **13,** 238-243.

Lercher, M.J., Urrutia, A.O., Hurst, L.D., Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. **31,** 180-183.

Levitsky, V.G., Podkolodnaya, O.A., Kolchanov, N.A., and Podkolodny, N.L. (2001). Nucleosome formation potential of exons, introns and *Alu* repeats. Bioinformatics **17,** 1062 - 1064.

Li, L., Wang, X., Stolc, V., Li, X., Zhang, D., Su, N., Tongprasit, W., Li, S., Cheng, Z., Wang, J., and Deng, X.W. (2006a). Genome-wide transcription analyses in rice using tiling microarrays. Nat. Genet. **38,** 124-129.

Li, Q., Barkess, G., and Qian, H. (2006b). Chromatin looping and the probability of transcription. Trends Genet. **22,** 197-202.

Li, Y., Bor, Y.-c., Misawa, Y., Xue, Y., Rekosh, D., and Hammarskjold, M.-L. (2006c). An intron with a constitutive transport element is retained in a Tap messenger RNA. Nature **443,** 234-237.

Liebhaber, S.A., Cash, F., and Eshleman, S.S. (1992). Translation inhibition by an mRNA coding region secondary structure is determined by its proximity to the AUG initiation codon. J. Mol. Biol. **226,** 609 - 621.

Liu, J., and Rost, B. (2001). Comparing function and structure between entire proteomes. Protein Sci. **10,** 1970-1979.

Logsdon, J.J.M. (1998). The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8,** 637-648.

Luger, K. (2006). Dynamic nucleosomes. Chromosome Res. **14,** 5-16.

Luscombe, N.M., GreenBaum, D., and Gerstein, M. (2001). What is Bioinformatics? A proposed definition and overview of the field. Methods Inf. Med. **40,** 346-358.

Lynch, M., and Kewalramani, A. (2003). Messenger RNA surveillance and the evolutionary proliferation of introns. Mol. Biol. Evol. **20,** 563-571.

Majewski, J., and Ott, J. (2002). Distribution and characterization of regulatory elements in the human genome. Genome Res. **12,** 1827 - 1836.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature **416,** 499-506.

Marra, M.A., Hillier, L., and Waterston, R.H. (1998). Expressed sequence tags -- ESTablishing bridges between genomes. Trends Genet. **14,** 4-7.

Martelli, A.M., Cocco, L., Riederer, B.M., and Neri, L.M. (1996). The nuclear matrix: a critical appraisal. Histol. Histopathol. **11,** 1035-1048.

Mattick, J.S., and Gagen, M.J. (2001). The evolution of controlled multitasked gene network: the role of introns and other non-coding RNAs in the development of complex organisms. Mol. Biol. Evol. **18,** 1611-1630.

Meyers, B.C., Lee, D.K., Vu, T.H., Tej, S.S., Edberg, S.B., Matvienko, M., and Tindell, L.D. (2004). Arabidopsis MPSS. An online resource for quantitative expression analysis. Plant Physiol. **135,** 801-813.

Mlynarova, L., Jansen, R.C., Conner, A.J., Stiekema, W.J., and Nap, J.P. (1995). The MAR-mediated reduction in position effect can be uncoupled from copy number-dependent expression in transgenic plants. Plant Cell **7,** 599-609.

Mlynarova, L., Loonen, A., Mietkiewska, E., Jansen, R.C., and Nap, J.P. (2002). Assembly of two transgenes in an artificial chromatin domain gives highly coordinated expression in tobacco. Genetics **160,** 727-740.

Mlynarova, L., Loonen, A., Heldens, J., Jansen, R.C., Keizer, P., Stiekema, W.J., and Nap, J.P. (1994). Reduced position effect in mature transgenic plants conferred by the chicken lysozyme matrix-associated region. Plant Cell **6,** 417-426.

Mockler, T.C., and Ecker, J.R. (2005). Applications of DNA tiling arrays for whole-genome analysis. Genomics **85,** 1-15.

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S.S., Skogen, J.W., and Meyers, B.C. (2006). Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res. **34,** D731-735.

Nakao, J., Miyanohara, A., Toh-e, A., and Matsubara, K. (1986). *Saccharomyces cerevisiae* PHO5 promoter region: location and function of the upstream activation site. Mol. Cell Biol. **6,** 2613-2623.

O'Brien, K.P., Remm, M., and Sonnhammer, E.L.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. **33,** D476-480.

Obayashi, T., Okegawa, T., Sasaki-Sekimoto, Y., Shimada, H., Masuda, T., Asamizu, E., Nakamura, Y., Shibata, D., Tabata, S., Takamiya, K., and Ohta, H. (2004). Distinctive features of plant organs characterized by global analysis of gene expression in Arabidopsis. DNA Res. **11,** 11-25.

Osley, M.A., Gould, J., Kim, S., Kane, M.Y., and Hereford, L. (1986). Identification of sequences in a yeast histone promoter involved in periodic transcription. Cell **45,** 537-544.

Ott, R.L., and Longnecker, M. (2001). Categorical data. In An Introduction to Statistical Methods and Data Analysis, C. Crockett, L. Jackson, P. Rockwell, H. Walden, and L. Campobasso, eds (Pacific Grove, CA 93950 USA: Duxbury), pp. 482-485.

Palmer, J.D., and Logsdon, J.M., Jr. (1991). The recent origins of introns. Curr. Opin. Genet. Dev. **1,** 470-477.

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling -- a review. Gene **238,** 103-114.

Pearson, H. (2006). Genetics: What is a gene? Nature **441,** 398-401.

Pollock, J.D. (2002). Gene expression profiling: methodological challenges, results, and prospects for addiction research. Chem. Phys. Lipids **121,** 241-256.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., de Jong, P., and Weissenbach et, a. (2005). Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. Science **310,** 1325-1326.

Reinartz, J., Bruyns, E., Lin, J.-Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik, R. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Brief Funct. Genomic Proteomic **1,** 95-104.

Ren, X.Y., Fiers, M.W.E.J., Stiekema, W.J., and Nap, J.P. (2005). Local coexpression domains of two to four genes in the genome of Arabidopsis. Plant Physiol. **138,** 923-934.

Ren, X.Y., Vorst, O., Fiers, M.W.E.J., Stiekema, W.J., and Nap, J.P. (2006). In plants, highly expressed genes are the least compact. Trends Genet. **22,** 528-532.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. **31,** 224-228.

Richmond, T.J. (2006). Genomics: Predictable packaging. Nature **442,** 750-752.

Riddle, N., and Elgin, S. (2006). The dot chromosome of Drosophila: Insights into chromatin states and their change over evolutionary time. Chromosome Res. **14,** 405-416.

Roy, P.J., Stuart, J.M., Lund, J., and Kim, S.K. (2002). Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature **418,** 975-979.

Roy, S.W. (2006). Intron-rich ancestors. Trends Genet. **22,** 468-471.

Roy, S.W., and Gilbert, W. (2005). Complex early genes. Proc. Natl. Acad. Sci. USA **102,** 1986-1991.

Roy, S.W., and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. Nat. Rev. Genet. **7,** 211-221.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., nbsp, B, Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., nbsp, M, Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., Zhong, F., Zhong, W., Gibbs, R., Venter, J.C., Adams, M.D., and Lewis, S. (2000). Comparative genomics of the eukaryotes. Science **287,** 2204-2215.

Rudd, S., Frisch, M., Grote, K., Meyers, B.C., Mayer, K., and Werner, T. (2004). Genome-wide *in silico* mapping of scaffold/matrix attachment regions in Arabidopsis suggests correlation of intragenic scaffold/matrix attachment regions with gene expression. Plant Physiol. **135,** 715-722.

Sakharkar, M.K., Chow, V.T., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico* Biol. **4,** 387-393.

Salse, J., Piegu, B., Cooke, R., and Delseny, M. (2002). Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res. **30,** 2316-2328.

Saluz, H.P., Iqbal, J., Limmon, G.V., Ruryk, A., and Wu, Z. (2002). Fundamentals of DNA-chip/array technology for comparative gene-expression analysis. Curr. Sci. **83,** 829-833.

Sanderson, M.J., Thorne, J.L., Wikstrom, N., and Bremer, K. (2004). Molecular evidence on plant divergence times. Am. J. Bot. **91,** 1656-1665.

Sankoff, D., and Haque, L. (2006). The distribution of genomic distance between random genomes. J. Comput. Biol. **13,** 1005-1012.

Schalch, T., Duda, S., Sargent, D.F., and Richmond, T.J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. Nature **436,** 138-141.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H.W., and Mayer, K.F. (2002). MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. Nucleic Acids Res. **30,** 91-93.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature **442,** 772-778.

Semon, M., and Duret, L. (2006). Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol. Biol. Evol. **23,** 1715-1723.

Seoighe, C., Gehring, C., and Hurst, L.D. (2005). Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. PLoS Genet. **1,** e13.

Shabalina, S., and Spiridonov, N. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol. **5,** 105.

Shermoen, A.W., and O'Farrell, P.H. (1991). Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. Cell **67,** 303-310.

Singh, G.B., Kramer, J.A., and Krawetz, S.A. (1997). Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Res. **25,** 1419-1425.

Spellman, P.T., and Rubin, G.M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. J. Biol. **1,** 5.

Stryer, L. (1999). DNA and RNA: molecular of heredity. In Biochemistry (W.H. Freeman and Company), pp. 75-94.

Taft, R., and Mattick, J. (2003). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Genome Biol. **5,** P1.

Tetko, I.V., Haberer, G., Rudd, S., Meyers, B., Mewes, H.W., and Mayer, K.F. (2006). Spatiotemporal expression control correlates with intragenic scaffold matrix attachment regions (S/MARs) in *Arabidopsis thaliana*. PLoS Comput. Biol. **2,** e21.

Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., Grisar, T., Igout, A., and Heinen, E. (1999). Housekeeping genes as internal standards: use and limits. J. Biotechnol. **75,** 291-295.

Todd, R., and Wong, D.T.W. (2002). DNA hybridization arrays for gene expression analysis of human oral cancer. J. Dent. Res. **81,** 89-97.

Tremethick, D. (2006). Chromatin: the dynamic link between structure and function. Chromosome Res. **14,** 1-4.

Urrutia, A.O., and Hurst, L.D. (2001). Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics **159,** 1191-1199.

Urrutia, A.O., and Hurst, L.D. (2003). The signature of selection mediated by expression on human genes. Genome Res. **13,** 2260-2264.

van Drunen, C.M., Oosterling, R.W., Keultjes, G.M., Weisbeek, P.J., van Driel, R., and Smeekens, S.C. (1997). Analysis of the chromatin domain organisation around the plastocyanin gene reveals an MAR-specific sequence element in *Arabidopsis thaliana*. Nucleic Acids Res. **25,** 3904-3911.

van Drunen, C.M., Sewalt, R.G., Oosterling, R.W., Weisbeek, P.J., Smeekens, S.C., and van Driel, R. (1999). A bipartite sequence element associated with matrix/scaffold attachment regions. Nucleic Acids Res. **27,** 2924-2930.

Velculescu, V.E. (1999). Tantalizing transcriptomes--SAGE and its use in global gene expression analysis. Science **286,** 1491-1492.

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, J.D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. (1997). Characterization of the yeast transcriptome. Cell **88,** 243-251.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H. (2003). The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. **13,** 1998-2004.

Vinogradov, A.E. (2001). Intron length and codon usage. J. Mol. Evol. **52,** 2-5.

Vinogradov, A.E. (2004). Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. **20,** 248-253.

Vinogradov, A.E. (2005). Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucleic Acids Res. **33,** 559-563.

Watson, J.D., and Crick, F.H.C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature **171,** 737-738.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L., and Senchina, D.S. (2002). Intron size and genome size in plants. Mol. Biol. Evol. **19,** 2346-2352.

West, R.W., Jr., Yocum, R.R., and Ptashne, M. (1984). *Saccharomyces cerevisiae* GAL1-GAL10 divergent promoter region: location and function of the upstream activating sequence UASG. Mol. Cell Biol. **4,** 2467-2478.

Williams, E.J., and Bowles, D.J. (2004). Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. **14,** 1060-1067.

Wilson, C., Bellen, H.J., and Gehring, W.J. (1990). Position effects on eukaryotic gene expression. Annu. Rev. Cell Biol. **6,** 679-714.

Woodcock, C.L., Frado, L.L., and Rattner, J.B. (1984). The higher-order structure of chromatin: evidence for a helical ribbon arrangement. J. Cell Biol. **99,** 42-52.

Wu, C.H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z.Z., Ledley, R.S., Lewis, K.C., Mewes, H.W., Orcutt, B.C., Suzek, B.E., Tsugita, A., Vinayaka, C.R., Yeh, L.S., Zhang, J., and Barker, W.C. (2002). The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res. **30,** 35-37.

Zhan, S., Horrocks, J., and Lukens, L.N. (2006). Islands of co-expressed neighbouring genes in *Arabidopsis thaliana* suggest higher-order chromosome domains. Plant J. **45,** 347-357.

Zhu, T. (2003). Global analysis of gene expression using GeneChip microarrays. Curr. Opin. Plant Biol. **6,** 418-425.

Zhuang, Y., Ma, F., Li-Ling, J., Xu, X., and Li, Y. (2003). Comparative analysis of amino acid usage and protein length distribution between alternatively and non-alternatively spliced genes across six eukaryotic genomes. Mol. Biol. Evol. **20,** 1978-1985.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol. **136,** 2621-2632.

Zuckerkandl, E. (1997). Junk DNA and sectorial gene repression. Gene **205,** 323-343.

Acknowledgments

Yesterday, I still didn't believe that I would make it till today. But I survived and achieved this doctor's degree after all. Of course I could only achieve this with the indispensable support from all of those people to whom I owe much thanks. I am going to mention them in the text below. Please forgive me if I forgot one of you.

Four years of being a PhD student under the supervision of Jan-Peter Nap was not an easy life. I have had ever more tears, anger, complaints, midnight and weekend work. But I have also learned enormously and had ever more achievements. I want to thank you for that. Thanks to your input and drive, I've learned to be more sharp and more critical, to doubt about everything that I read and write, to think more scientifically and to writing succinctly. I have learned to make whatever complicated study into a 15' highlight presentation. I now know how to never lose my arguments in whatever I do. I could have learned a lot more from you. Above all, though, I think I should thank you for making my years so miserable that I am now ready to survive anybody and anything else in the future. In comparison, Willem Stiekema, my promoter, was so kind to agree with almost everything I asked for. Thanks Willem, for your input in my scientific work and supporting me with the Dutch language course, a three months' extension, extra cost of thesis printing and many more things. I also would like to thank Hans Sandbrink, the gentle giant from whom I learned so much during my MSc thesis. Without Hans, I would not have initiated my journey into bioinformatics. Unfortunately he passed away just before I started this PhD project. I regret a lot that he could not share my achievements and be proud on me today. Hans, you will always be remembered.

Mark, my best colleague and friend, I thank you for always being there and for supporting and listening to me whenever I needed you and also thanks for all the help in my research. Thanks to Roeland for always being the intermediary person solving any conflicts and for your input to our many discussions on evolution. Oscar, thank you for playing around with my data that gradually led to our TIG paper. Ate and Erwin, my office mates, now you finally will have some peace and quiet without me in the office. Thanks Bas, Joost and Jan for helping me with scripting and all kinds of computer problems. Thanks also to all the others of the Applied Bioinformatics cluster who gave me useful comments during working discussions and correct my faulty Dutch practices from time to time.

For the administration, I was part of the Laboratory of Molecular Biology, although I was hardly there, only once in a while I gave a presentation there. Then I always got a lot of valuable comments from Ton Bisseling and I want to thank him for his interest and input. I also like to thank Maria for always reminding me to use my holiday days.

To my bioinformatics colleagues on the top of the Wageningen mountain: thanks Jack for being one of my reading committee members and thanks Arnold, Jifeng, Anand for the nice chats from time to time. Also thanks Christine for helping me out a lot at the beginning of the “Phytoinformatics” project. Thanks Caroline for your help at the start of my project and Eric for your biological questions. I would like to thank Rudi and Ritsert from Groningen for the work we have done together. Hopefully we can publish our joint work in a highly cited journal.

Plant Research International (PRI) is especially “international” in the evenings and weekends. By then one can really experience the exotic mix of many nationalities: Israeli, Italian, Spanish, Indonesian, Chinese, Mexican, Lebanese, Slovakian, Bulgarian, Indian, Brazilian, American, Tunisian and an occasional Dutch. It is because of this large number of foreigners during these hours that we easily got closer to each other than to our direct (Dutch) colleagues. I thank all international friends for the time we have shared, now or before, for long or for short. The contacts with you all have enriched my life.

With some of you, who stayed at PRI a bit longer, I have built up particularly special memories. Thursday drinks (to break the week), dinners together (especially at Sa Lolla), midnight parties, Sunday movies, bowling Arnhem, city excursions, Barcelona reunion, chatty chatting from this to that… all in all, it’s just so much that we have shared that I simply cannot mention it all. Asaph Asaph (without your inspiration next door, I hardly work till 10pm anymore), Raffa, Raffa (step by step, we will be there!), Salva Salva (without you, we hardly kept the “tradition”, and thanks for those most true words which hurt the most), Tetty Tetty (thanks for the haircut), I miss you all! PRI without you is different. We still keep skyping and keep gathering somewhere sometime, ok? Nayelli & Stefan, Rumyana & Nicolay, Ruud and Rada, we got much closer after they left. Thanks for all the support I got from you especially during the writing stage. Mazen and Xena, welcome back to the club and thanks for fuelling me up with the “never give up” Lebanese spirit during the last period of writing my thesis. Xena and Ana, I like to thank you especially for joining me to the parties those Saturday nights till 5am. These hard-partying seemed indispensable for me during the hard-working period. To the other former or current PRI-ers or PRI-related friends: Cinzia, Teun, Andi, Rudi, Shital, Jochem & Agata, Radi, thanks for the time and talks we’ve shared here and there.

I would also like to thank my classmates and friends from my MSc period: Herman, Ruben, Susana, Xiaolian, Paul & Veronique, Yansen. Thanks for the happy occasional gatherings. I hope that together we can celebrate more 5 mei occasions in Holland. I thank my Chinese friends Sun, Chun-Ming, Sanwen & Wang Ming, Guo Hai, Ningwen, Yuling. I really enjoyed working for the Chinese Association of Students and Scholars in Wageningen (CASSW) and the activities we have initiated and organized, especially the Chinese New Year parties. Thanks for sharing the experience of living in Holland.

During the last year of my PhD, I especially owe much thanks as well as much sorry to Cedric, who supported me, tolerated me and loved me a lot. Success with your work and life! And Philippe, thanks for coming back to see me whenever you are in Europe. Your spirit supported me invisibly during these 4 years. Also thanks to Jean and Georgette for your constant care, for your interest in my research and for the end-of-the-year invitation. To my best friend Dong Lei: many thanks for accompanying me so long in my life and for being the one I can always turn to and count on. Your long-distance mental support is always just by my side.

My family is small, so it is easy to mention everybody, but I will never finish thanking all of you. I want to thank my brother and my sister-in-law for listening to and supporting me all the time. Then, last, but by far not least, deep thanks to my mum and dad. I am sorry for not coming home even once during these years to be with you, if only for a few days. Work is an unforgivable bad excuse. Curiosity always drove me to sacrifice going home for exploring new things (learning driving, skydiving, traveling around). I thank you very much for your understanding of my choices and ever supporting me with whatever I wanted. Thanks also for being the only ones who could listen to my phone calls in which usually already in a few sentences I threw out my “hot tempers” (as my father puts it)… Thanks for your tolerant love that never stops, that never restricts me and that is always by my side. I hope and trust that this PhD degree will make you proud of me and will compensate a bit for my absence during the past years.

To all of you my sincere thanks!

Xinying

Wageningen, November 2006

Curriculum vitae

Xin-Ying Ren (任新颖) was born on the 10th of January 1977 in Dongyang city in the southeastern part of China. She grew up in Urumqi, a city with many ethnic groups in the most northwestern part of China. In 1995, she started her academic study at Nanjing Agricultural University (NAU) and obtained a BSc in Horticulture in 1999. Thanks to the good relationship between NAU and Wageningen University, she came to Wageningen in August 2000 for an MSc in Biotechnology. With a score of 9 for her thesis on “In-depth genome annotation of White Spot Syndrome Virus of Shrimp”, she earned her MSc degree in January 2002. Subsequently she started her scientific journey into bioinformatics. After six months research on human immunology at Utrecht University, she came back to Wageningen. From September 2002 till December 2006 she carried out her PhD research on “Comparative genomics of the relationship between gene structure and expression” at the Applied Bioinformatics group of Plant Research International, within the Plant Sciences group of Wageningen University and Research Center. From 2007 on, she plans to move to Spain for post-doctoral research on the evolution and regulation of splicing in fungal genomes at the Research Unit on Biomedical Informatics, Pompeu Fabra University/Municipal Institute of Medical Research in Barcelona, Spain.

email: renxinying0110@yahoo.com

Wageningen
Urumqi
Barcelona
Nanjing
Dongyang

List of Publications

Ren, X.Y., Stiekema, W.J. and Nap, J.P. Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains. *Submitted*

Ren, X.Y., Vorst, O., Fiers, M.W.E.J., Stiekema, W.J. and Nap, J.P. (2006). In plants, highly expressed genes are the least compact. *Trends in Genetics* 22, 528-532
(Impact factor: 12.0)

Marks, H., **Ren, X.Y.**, Sandbrink, H., van Hulten, M.C.W. and Vlak, J. (2006). *In silico* identification of putative promoter motifs of White Spot Syndrome Virus. *BMC Bioinformatics* 7, 309
(Impact factor: 4.96)

Ren, X.Y., Fiers, M.W.E.J., Stiekema, W.J. and Nap, J.P. (2005). Local coexpression domains of two to four genes in the genome of Arabidopsis. *Plant Physiology* 138, 923-934
(Impact factor: 6.11)

MSc thesis:
Ren, X.Y. (2002). In-depth genome annotation of White Spot Syndrome Virus of shrimp. In the Laboratory of Viology and Genomics, Plant Research International (Wageningen: Wageningen University), pp. 97

Education Statement of the Graduate School
Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Xin-Ying Ren**
Date: **20-dec-06**
Group: **Applied Bioinformatics, Plant Research International & Bioinformatics,**
Wageningen University & Research Centre

1) Start-up phase	*date*
► **First presentation of your project**	
S/MAR element and its predictions by computer programs	Feb 6, 2003
► **Writing or rewriting a project proposal**	
The functional annotation and interpretation of the non-protein coding parts of the genome of *A.thaliana* (Ara-mining II)	Sep 2002-Feb 2003
► **Writing a review or book chapter**	
► **MSc courses**	
Advanced Statistics MAT-20304	Nov-Dec 2002
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*13.5 credits**

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD day, Utrecht University	Mar 27, 2003
EPS PhD day, Vrije Universiteit Amsterdam, with poster	Jun 3, 2004
► **EPS theme symposia**	
EPS theme 4 symposium 'Genome Plasticity', Radboud University Nijmegen	Dec 10, 2003
EPS theme 4 symposium 'Genome Plasticity", Wageningen University	Dec 9, 2004
► **NWO Lunteren days and other National Platforms**	
ALW/EPW meeting in Lunteren	Apr 7-8, 2003
ALW/EPW meeting in Lunteren, with poster	Apr 5-6, 2004
"Image of Life"- Netherlands Conference on BioInformatics, Groningen, with poster	Oct 7-8, 2004
ALW/EPW meeting in Lunteren, with poster	Apr 4-5, 2005
► **Seminars (series), workshops and symposia**	
Comparative genomics symposium Utrecht	Sep 17, 2002
1st GeNeYouS Symposium "Decisions in Genomics" in Utrecht	Jan 20, 2004
Workshop "Mathematics in Plant Biology", Paris	Jun 30-Jul 1, 2005
MATLAB workshop, Maarsen	Sep 1, 2005
Mini-symposium "*Microbial Genomics*", microbiology WUR	Nov 11, 2005
Phytoinformatics project meeting series (~8 times), every half year, each with oral presentation	2002-2006
Thematic meetings Business Unit Bioscience, with yearly presentations	2002-2006
Seminar series at Laboratory of Molecular Biology, with yearly presentations	2002-2006
► **Seminar plus**	
Flying seminar by Prof. Dr. Philip Benfey	Oct 24, 2005
► **International symposia and congresses**	
The Dutch-Chinese Life Science Forum, Wageningen	Oct 12, 2003
FEBS Adv. Workshop "Nuclear Architechture, Chromatin Structure and Gene Control: Plants vs. Animals vs. Yeast	Nov 14-17, 2003
"Genomics for our world" Genomics Momentum, Rotterdam	Aug 31-Sep 1, 2004
1st Benelux Bioinformatics Conference (BBC), Ghent University, Belgium, with oral presentation	Apr 14-15, 2005
Plant Genomics European Meeting (Plant GEMs) Amsterdam, with poster	Sep 20-23, 2005
2nd Benelux Bioinformatics Conference (BBC), Wageningen, with oral presentation	Oct 17-18, 2006
► **Presentations**	
SpringSchool: Chromosomal coexpression domains in Arabidopsis thaliana genome (oral)	Mar 31, 2004
poster presentation: Chromosomal coexpression domains in Arabidopsis thaliana genome	Oct 7-8, 2004
BBC 2005: Local coexpression domains in Arabidopsis genome (oral)	April 14, 2005
poster presentation: Local coexpression domains in Arabidopsis	Sep 20-23, 2005
BBC 2006: Comparative genomics of the relationship between gene struct. and expression in higher eukaryotes (oral)	Oct 17, 2006
► **IAB interview**	Jun 1, 2005
► **Excursions**	
Subtotal Scientific Exposure	*15.8 credits**

3) In-Depth Studies	*date*
► **EPS courses or other PhD courses**	
SpringSchool Bioinformatics, Data Triple I: Information, intergration, interpretation. Wageningen, NL	Mar 31-Apr2, 2004
PhD course "Computational Genomics", Cold Spring Harbor Laboratory, New York, USA	Nov 3-9, 2004
► **Journal club**	
literature discussions every two weeks, at least 4 times presentations	2003-2006
► **Individual research training**	
Subtotal In-Depth Studies	*4.5 credits**

4) Personal development	*date*
► **Skill training courses**	
Scientific writing course CENTA	Apr-Jun 2004
► **Organisation of PhD students day, course or conference**	
Organisation of the project "Phytoinformatics" meeting, at least 2 out of 8 times	2002-2006
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*3.3 credits**

TOTAL NUMBER OF CREDIT POINTS*	**37.1**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 credits

** A credit represents a normative study load of 28 hours of study*

Cover design by: Xin-Ying Ren
Printed by: PrintPartners Ipskamp

This research was supported by a program subsidy from the Dutch Organization for Scientific Research (NWO) and by the Center for BioSystems Genomics (CBSG).