

Ghent University
Faculty of Sciences
Department of Molecular Genetics
VIB-Department of Plant Systems Biology
Bioinformatics and Evolutionary Genomics Research Group

The use of
microarray technology
to analyse
the evolution
and functional divergence
of duplicated genes

Tineke Casneuf

Promotor: Prof. Dr. Yves Van de Peer

Dissertation submitted in fulfillment of the requirements for the degree
of Doctor (PhD) in Sciences, Biotechnology



Ghent University
Faculty of Sciences
Department of Molecular Genetics
VIB-Department of Plant Systems Biology
Bioinformatics and Evolutionary Genomics Research Group

**THE USE OF MICROARRAY TECHNOLOGY
TO ANALYSE
THE EVOLUTION AND FUNCTIONAL
DIVERGENCE OF DUPLICATED GENES**

by

Tineke CASNEUF

Promotor: Prof. Dr. Yves Van de Peer

Dissertation submitted in fulfillment of the requirements for the degree
OF DOCTOR (PHD) IN SCIENCES, BIOTECHNOLOGY

December 2007

Examination committee

Prof. Dr. Ann Depicker (Chairwoman)
Faculty of Sciences, Ghent University

Prof. Dr. Yves Van de Peer (Promotor)
Faculty of Sciences, Ghent University

Prof. Dr. Wolfgang Huber
European Bioinformatics Institute, EMBL, Cambridge, UK

Prof. Dr. Geert De Jaeger
Faculty of Sciences, Ghent University

Prof. Dr. Dirk Inzé
Faculty of Sciences, Ghent University

Dr. Martin Kuiper
Faculty of Sciences, Ghent University

Prof. Dr. Kathleen Marchal
Department of Microbial and Molecular Systems, K.U.Leuven

Prof. Dr. Paul Van Hummelen
VIB MicroArray Facility, K.U.Leuven

Acknowledgments

Yves, dankzij jouw enthousiasme raakte ik een viertal jaar geleden gefascineerd door de Moleculaire Evolutie en tuimelde ik de Bioinformatica wereld in. Alhoewel ik toen niet goed beseftte hoe bepalend mijn keuze was voor mijn verdere loopbaan, kan ik nu zeggen dat het uitpluizen van wetenschappelijke vraagstellingen met een computer mij als gegoten zit! Ik wil je bedanken voor de vele kansen die je me gaf, voor jouw onuitputtelijke inzet, vertrouwen, steun en... voor de ontspannende babbels tussendoor. Van jouw vele eigenschappen als goeie wetenschapper heb ik een beetje overgenomen, van sommige andere - zoals pragmatisch zijn - kan ik nog veel leren! *Yves*, bedankt!

Wolfgang, my partner in crime when it comes to enjoying the good life: chocolates, red wine, tapas.. Working with you was truly an honour to me. Your intelligence, enthusiasm and your ambition were highly motivating. Your commitment, curiosity and perfectionism were at times challenging, but always pushed me to do good science. I learned much from our discussions, as they improved my independent critical thinking and research skills. You were a great mentor and will remain a great friend!

Working in two different groups meant that I was given the opportunity to work with several different people. *Jeroen*, onze samenwerking was van korte duur maar ze heeft mijn doctoraatservaring sterk getekend. Jouw positieve en relativerende kijk op de uitdagingen waarvoor ik stond, heeft mij veel bijgebracht. Bedankt voor de leerrijke discussies, de oprechte belangstelling en om mij ten gepaste tijde met mijn twee voetjes op de grond te zetten! *Stefanie*, na Jeroens vertrek was het voor mij eventjes zoeken, maar gelukkig was jij daar. Bedankt voor het meedelen van jouw visie en creatieve oplossingen en voor de vele discussies waarbij ik mijn werk -en soms iets luchtere zaken- met jou kon bespreken. *Klaas*, ik heb veel bewondering voor de enorme hoeveelheid kennis die jij bezit en voor jouw doorzettingsvermogen. Het was een eer om samen aan een project te mogen werken! *Steven Rrrrrrobbens* en 'op uwe stoel moet ik u ni gaan zoeken' -*Lieven*, merci voor de vrolijke noten aan Eiland 1! *Dirk*, bedankt voor de afleidingen van de werk-sfeer. *Pierre*, thanks for the challenging discussions and motivating words. *Cindy*, *Tine2* en *Ann* merci voor de toffe babbels, het ga jullie goed,

meiskes! *Cedric* en *Jan*, bedankt voor alle hulp die ik kreeg als mijn laptop of mijn scripts het weer niet deden! *Steven M*, bedankt voor de hulp en samenwerking. IT-guys, *Eric*, *Yvan*, *Thomas*², *Kenny*, *Francis*, thanks for the tips! *Stephane*, bedankt om altijd klaar te staan. *Tom*, bedankt voor de looptoerkes die mijn hoofd verfristen. *Elisabeth*, *Vanessa*, *Anagha*, *Greg*, *Yao-Cheng*, *Jeffrey*, *Michiel*, *Sofie*, *Roeland*, *Pedro*, *Sebbe* and *Ben*: thanks for the good times! Thanks also to the people in the front building, who helped me out on several occasions: *Luc*, *Dany*, *Raf*, *Hendrik*, *Karel*, *Diane*, *Christine*, *Jacques*, *Martine*, *Hilde*, and *Linda*.

Jörn, where can I start? We did not share opinions on what is the best place to live in a town, what is the maximal number of pets one can have and which country has the best beer, but that we both like pub crawls, chocolates, Indian food and wine in a punt, that is a fact! Thanks for making days sunny in rainy England! *Oleg*, I would not have made it back and forth between Cambridge and Ghent so smoothly without your help. Thank you for always being willing to help, no matter what. *Richard*, early bus rides turned fun when you showed up. I was often impressed by your intelligence and cleverness. *Tommeke*... euh... *Tony*, your interest in the cute Belgian cyclist came as a surprise to me. Nevertheless, I was flattered and think of you each time a flashy cyclist overtakes me. Thanks for the late-night talks! *Greg*: xie xie! *Elin* and *Audrey*, thanks for the inspiring and pleasant chats.

I also would like to thank *Prof Ann Depicker* and the members of Examination Committee, *Prof Geert De Jaeger*, *Prof Dirk Inzé*, *Dr Martin Kuiper*, *Prof Kathleen Marchal* and *Prof Paul Van Hummelen* for reading my thesis and their helpful suggestions. A special thanks goes to *Dirk Inzé* for providing the stimulating environment for doing great science. It was a privilege to be a part of the Plant Systems Biology department!

Niet te vergeten zijn mijn vrienden, die voor een onvergetelijke tijd zorgden naast het werk: 'party girl' *Katheen*, 'kapiteinen van de korte omvaart' *Katelijne & Sarah*, 'kapiteinen van de iets langere omvaart' *Jeroen* en *Stijn* en vliindertje *Bram*. *Stesse*, *Boemel*, *Schoeli*, *der Schmutziger Kerstens*, *Koala Anneke* - zijn we onze wilde haren nu echt kwijt?!-. 'Apple-guard' *Q*, 'Badminton-ster' *Leen*, 'L^AT_EX rules!' *Jurgen*, 'your-Tunes' *David* en de andere WOZ'ers -of was't Brabo?of KDA?- en *Mr. Joinkes* -ferme merci voor de uitdagingen op verschillende niveaus. 10 mijl uitloppen met een bloedneus, Respect!.

Special thanks goes to: *Charlotte* voor het opvullen van creatieve leemte en het geduldig verwerken van details en opmerkingen -CMYK: 32,9,100,38, alstublieft. Bedankt *Lies*, voor het zorgvuldig nalezen van mijn thesis, het klaarstaan in tijden van zieke kippen en voor het luisterend en geïnteresseerd oor waaraan ik mijn passie voor de wetenschap kwijt kon, ook al verloor ik je wellicht ergens in het midden.

Veerle, bedankt voor de luisterende oren die je was en om de vele statistiekcursus-avonturen te doorspartelen met mij, ik heb er veel van geleerd. Veel succes met die laatste loodjes! Bedankt ook *Sam*, *Jeroen*, *Isabel* en *Marc* voor jullie steun en hulp, *Laurien*, *Joni*, *Hanne* en *Jef* voor jullie jeugdige onbezorgdheid. *Lutgard*, heel hartelijk dank voor de gezellige en culinair hoogstaande uitstapjes die het weekend heerlijk ontspannend maakten. Bedankt ook voor je steun en hulp en om altijd klaar te staan, als het eventjes moeilijker werd.

Mama, bedankt voor de kansen die je me gaf om mijn dromen te achtervolgen.
Bart, bedankt om met mij mijn dromen waar te maken!

Gent, November 2007
Tine

Malcolm Forbes once got lost floating in one of his famous balloons across miles and finally landed in the middle of a cornfield. He spotted a man coming toward him and asked: "Sir, can you tell me where I am?". The man said: "Certainly, you are in a basket in a field of corn." Forbes said: "You must be a statistician". The man said "That's amazing, how did you know that?" "Easy", said Forbes, "your information is concise, precise and absolutely useless!"

"Looking ahead: Cross-disciplinary opportunities for statistics",
by R. Gnanadesiken

Table of Contents

Acknowledgments	i
1 Introduction	1
1.1 Gene duplication: two's a company, three's a party!	2
1.1.1 Generation of duplicated genes	3
1.1.2 The fate of a duplicated gene	6
1.2 <i>Arabidopsis thaliana</i> : the weed that made it to model organism . .	8
1.3 Microarrays	9
1.3.1 GeneChip basics	11
1.3.2 GeneChip data pre-processing	12
1.3.3 Data pre-processing algorithms	13
1.4 Putting the pieces together	18
2 Non-Random Divergence of Gene Expression Following Gene and Genome Duplications in the Flowering Plant <i>Arabidopsis thaliana</i>	21
2.1 Background	22
2.2 Results and Discussion	24
2.2.1 Divergence of expression and mode of duplication	25
2.2.2 Divergence of expression and gene function	30
2.3 Conclusions	33
2.4 Methods	34
2.4.1 Duplicated genes	34
2.4.2 Gene Ontology functional classes	35
2.4.3 Microarray expression data	35
2.4.4 Correlation analysis	37
2.4.5 Regression analysis	37
3 Identification of Novel Regulatory Modules in Dicotyledonous Plants using Expression Data and Comparative Genomics	39
3.1 Background	40
3.2 Results and Discussion	42
3.2.1 General overview	42
3.2.2 Identification of individual TFBSs using co-expressed genes	44
3.2.3 Combining motif and expression data to identify additional TFBSs	49

3.2.4	Inferring functional regulatory modules	50
3.2.5	Properties of <i>cis</i> -regulatory modules	55
3.3	Conclusions	56
3.4	Methods	57
3.4.1	Expression data	57
3.4.2	Clustering of expression data	57
3.4.3	Detection of transcription factor binding sites	58
3.4.4	Clustering based on TFBS content	59
3.4.5	Network-level conservation score	59
3.4.6	Orthology determination	60
3.4.7	Functional annotation	61
3.4.8	Expression coherence	61
4	<i>In situ</i> Analysis of Cross-Hybridisation on Microarrays and the Inference of Expression Correlation	63
4.1	Background	64
4.2	Results and Discussion	67
4.2.1	Two definitions of probe set annotation	67
4.2.2	Off-target alignments	68
4.2.3	Correlation of microarray expression profiles	69
4.2.4	Probe set off-target sensitivity and expression correlation .	70
4.2.5	Reporter off-target sensitivity and expression correlation .	72
4.2.6	Examples	74
4.2.7	Effect of individual reporters on probe set summaries . . .	75
4.3	Conclusions	76
4.4	Methods	77
4.4.1	Two Chip Description Files	78
4.4.2	Reporter-to-transcript alignments	78
4.4.3	Microarray data	79
4.4.4	Identification of gene pairs with long stretches of sequence similarity	79
4.4.5	Metacorrelation	79
5	Application of the Microarray Technology to the Study of Evolution and Functional Divergence of Duplicated Genes	83
5.1	Background	84
5.2	Results and Discussion	85
5.2.1	Independent duplicated genes	85
5.2.2	Functional annotation and expression data	86
5.2.3	Expression divergence in different functional classes . . .	86
5.3	Conclusions	89
5.4	Methods	90
5.4.1	Identification of independent duplicated genes	90
5.4.2	Gene Ontology functional categories	91
5.4.3	Microarray data	91

5.4.4	Correlation analysis	92
5.4.5	Regression analysis	92
6	Concluding Remarks	93
7	Nederlandstalige Samenvatting	101
8	English Summary	105
A	Glossary	135
B	List of Acronyms	139
C	Publication List	141

List of Figures

1.1	Different mechanisms of gene duplication	4
1.2	Possible functional fates of a duplicate gene pair	7
1.3	Segmental duplications in the <i>Arabidopsis</i> genome	9
1.4	Overview of the design and use of a microarray	10
1.5	Basics of the Affymetrix GeneChip technology	12
1.6	Background correction in Affymetrix' MAS5.0	14
1.7	Effect of MAS5.0 and RMA normalisation	18
2.1	Six subclasses of duplicated genes in <i>Arabidopsis thaliana</i>	25
2.2	Expression correlation for anchor points and non-anchor points	26
2.3	Tissue co-expression of duplicated gene pairs	27
2.4	Possible scenarios for tissue-specific expression of a duplicated gene pair	28
2.5	Expression divergence in function of time for genes of different functional classes	32
3.1	Network-level conservation filter	43
3.2	Detection of TFBSs using two-way clustering	45
3.3	Motif synergy map for 139 modules with significant GO Biological Process annotation	52
3.4	Correlation between <i>cis</i> -regulatory modules and clusters of co-expressed genes	53
3.5	Motif distance distributions	56
4.1	Illustration of our approach	69
4.2	Custom-made versus Affymetrix CDF	71
4.3	ρ_{XY} by Q_{XY}^{75}	72
4.4	$\text{cor}(\rho_{X_iY}, a_i)$ by Q_{XY}^{75}	73
4.5	Three examples of cross-hybridisation	80
4.6	Effect of individual reporters on probe set summaries	81
5.1	Duplicates belonging to functional classes of slowly diverging genes	87
5.2	Duplicates belonging to functional classes of quickly diverging genes	88
5.3	Duplicates belonging to functional classes of moderately diverging genes	89

List of Tables

3.1	Overview of the TFBSs identified using co-expressed genes	46
4.1	Statistics of probe set definitions	68
4.2	Table with some of the highest Needleman-Wunsch scores	70

1

Introduction

Any good poet, in our age at least, must begin with the scientific view of the world; and any scientist worth listening to must be something of a poet, must possess the ability to communicate to the rest of us his sense of love and wonder at what his work discovers.

Edward Abbey, *The Journey Home*

The determination of the human [1] and other organisms' genomic sequence [2–9] enabled genetics in the mid-1990s to shift from gene-focused types of research approaches to large-scale, genome-wide analyses. Supplemented by technological advances in transcriptomics and proteomics this encouraged the emergence of functional genomics, comparative genomics and interdisciplinary fields like systems biology. Doors were thereby opened to the discovery of genes' functions, interactions between genes and the genetic architecture of biochemical and developmental pathways. Bioinformatics, at the interface between biological and computer sciences, thrived as the need for high-throughput data analysis, data mining and data integration increased.

Amongst other important discoveries, these projects revealed that gene duplication is rampant in eukaryotic genomes [10–18], an observation that had already been made through early research in comparative cytology (see [19] for an overview). The first part of this dissertation focuses on the fate of genes

after duplication and makes use of data generated using microarrays, a novel high throughput technology that had a substantial impact on life sciences. In a second part, these functional data are used to identify functionally related genes so as to enhance the identification of regulatory elements in promoter regions. A third part deals with the concern of cross-hybridisation of molecules with similar sequences on microarrays, giving rise to spurious correlations.

1.1 Gene duplication: two's a company, three's a party!

The few really big steps in evolution clearly required the acquisition of new information. But specialisation and diversification took place by using differently the same structural information.

François Jacob, The Possible and the Actual (1982)

Gene duplication and subsequent divergence, leading to the formation of families of evolutionary related genes, has given rise to an enormous variability in the number of genes among species. A high prevalence of gene duplicates is common among all eukaryotes: up to 30%, 38% and 65% of the genes of respectively yeast [20], human [21] and *Arabidopsis* [7] are estimated to be part of a gene family that arose through duplication and subsequent divergence of genes [7]. These duplicates have in turn played an important role in adaptive evolution of organisms and the origin of organismal complexity [22–28]. Understanding how gene copies have given rise to the genes present in extant organisms has fascinated evolutionary biologists for decades (see [19] for an overview). In 1970, well ahead of his time, geneticist Susumu Ohno highlighted the importance of gene duplication in a seminal work, *Evolution by Gene Duplication*:

Had evolution been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged. The creation of metazoans, vertebrates and finally mammals from unicellular organisms would have been quite impossible, for such big leaps in evolution required the creation of new gene loci with previously non-existent functions. Only the cistron that became redundant was able to escape from the relentless pressure of natural selection. By escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus.

Ohno thereby recognised that major advances in evolution such as the transition from single-celled organisms to complex multicellular animals and plants could

simply not have been brought about solely through natural selection based on existing allelic variation at particular genetic loci in populations. He postulated that large-scale gene duplication events are the principal forces by which extra raw genetic material is provided for increasing complexity during evolution. Furthermore, that novelty in evolution is most often based on genomic redundancy that is initially created by gene and entire genome duplications and which can act as substrate for subsequent divergent natural selection. He suggested that gene and genome duplications bring about evolutionary innovation by allowing for gene functions to diversify and for genes to take on novel functions. Natural selection's role in evolution was thereby reduced to fine-tuning the newly created material.

The importance of gene duplication in the evolution of genomes, however, could at that time not be fully appreciated without knowledge about the extent of gene duplication in a genome. The explosion of genomic sequence information and computer power, supplemented by theoretical advances in the design of algorithms for the detection of duplicated regions, at the dawn of the 21st century, put formal testing within the realms of possibility. Indications of genome-wide duplication events have been found at key evolutionary crossroads like the transition from invertebrates to vertebrates [15, 29–32], the explosive radiation of the teleost fish that resulted in 22,000 extant species [33–36] and the angiosperm radiation [37–39]. The debate on the extent of the different duplication events is ongoing [40, 41], but there is no doubt that gene duplication is a ubiquitous feature of genome evolution.

1.1.1 Generation of duplicated genes

Various events result in the creation of extra gene copies and outcomes range from the duplication of a few genes known as tandem or dispersed duplications, and duplication of subchromosomal-length regions known as segmental duplications, to the doubling of the entire genome [42–45]. These events include unequal crossing over, transposition, retrotransposition and polyploidisation and are schematically presented in Figure 1.1.

Unequal crossing over

Intrachromosomal small-scale duplication can be mediated by the presence of repetitive elements that cause a chromatid not to line up exactly with its corresponding region in the homologous chromosome (Figure 1.1A) or identical sister chromatid, resulting in the unequal exchange of DNA [46, 47]. The genetic information that is contained within this region is respectively gained in one chromosome or one chromatid and gets lost in the other. Unequal crossing over,

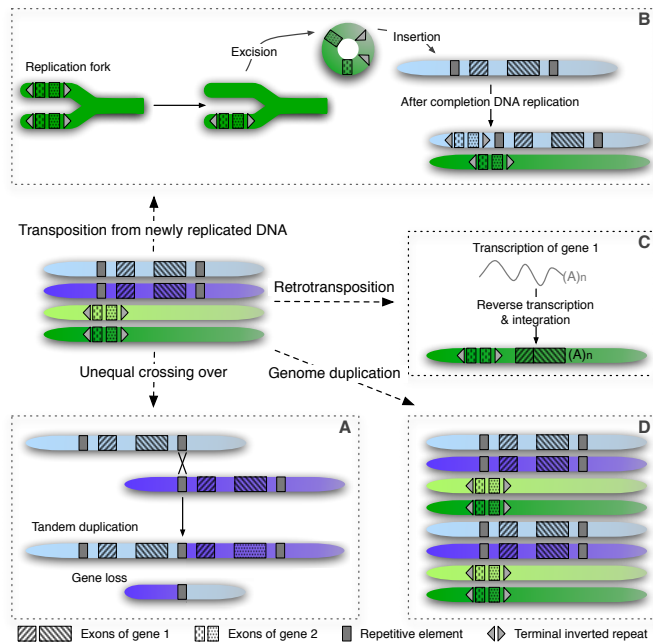


Figure 1.1: Different mechanisms of gene duplication. A hypothetical genome with two pairs of homologous chromosomes (light & dark blue and light & dark green; middle left), each having one gene that consists of two exons (shaded rectangles). The blue chromosome pair also contains repetitive elements (grey rectangles). The gene on the green chromosome pair is neighboured by inverted repeats. A) Local gene duplication can arise through unequal crossing over between homologous chromosomes. B) Dispersed duplicates can result from transposition, whereby a DNA transposon is moved from a location on one newly replicated DNA segment (dark green chromosome) into a region of the genome that has yet to be replicated (light blue chromosome), or C) from retrotransposition where a portion of the mRNA transcript of a gene is reverse transcribed back into cDNA and inserted into chromosomal DNA. D) Genome duplication or polyploidisation is a large-scale event in which the whole genomic content is doubled.

in case of homologous chromosomes, or unequal sister chromatid exchange is typically responsible for the generation of tandem duplicates and gene clusters (e.g. Hox [48] and Zinc finger clusters [49, 50]).

High concentrations of short (4 to 10 bp) repeats have been shown to trigger looping of the newly synthesised strand and mispairing to the template strand during DNA replication, resulting in an increase in the DNA content, a process called replication slippage. Typically this process involves short sequence stretches and generates short repeats, but, in theory if genes reside within

the looped region, gene duplication could occur [47, 51]. For instance, exon duplication linked to replication slippage has been described [46].

(Retro)Transposition

Gene transposition refers to the relocation of relatively small genomic segments from one chromosomal position to another. When accompanied by the duplication of a genomic segment, this process is responsible for the dispersion of related sequences and is referred to as duplicative transposition (Figure 1.1B). The process is denoted as retrotransposition if the transposition occurs by means of an RNA intermediate that is reverse transcribed into cDNA and inserted into the genome. A transposable retroelement is flanked by terminal inverted repeats containing binding sites for the enzyme transposase, that catalyses the transposition. The transposed duplicate can be recognised by the loss of the intronic region(s), as compared to the original gene and the presence of the poly-A tail (middle right of Figure 1.1). Because the process of retrotransposition is prone to various errors like point mutation owed to inaccurate reverse transcription, truncation or the absence of regulatory sequences in the novel genomic location, the copy of a (retro)transposon is bound to get inactivated and turned into a pseudogene. However, retrotransposition has been described to have generated new functional genes (retrogenes) in mammalian and invertebrate animal genomes, where DNA fragments, created by duplicative transposition contain repetitive DNA, portions of genes and complete genes [52–57] and in the genomes of rice, *Arabidopsis* and maize mediated by the Pack-MULE element [58]. The pericentromeric regions of chromosomes are known to be unstable and duplications followed by insertion into other chromosomes may be frequent [59].

A possible scenario in which transposition is linked to duplication is cut-and-paste transposition during DNA replication: a transposon is moved from one newly replicated DNA segment into a region that has yet to be replicated, resulting in one daughter copy that contains the transposable element only in its new location and one copy that has the element in both the original and the novel location (Figure 1.1C, only the latter daughter copy is shown) [60]. An alternative is that the transposon itself programs a replication, in which no excision occurs but where the elements duplicate using a semi-conservative DNA replication [61], a mechanism that has been widely described in bacteria but also in maize cells [62]. However, the chance that a transposed region includes and succeeds in duplicating a functional gene is slim.

Genome-wide duplication

Polyploidisation is the doubling of a complete single genome (autopolyploidy, Figure 1.1D) or the merging of two or more genetically different genomes (allopolyploidy). Genome duplication has shaped the genomes of most, if not all, eukaryotes [10–18, 63, 64] and is, especially for plants, a prominent force in genome evolution. In particular, it is a hallmark occurring at different frequencies among angiosperm families but it is not a common feature of gymnosperm genomes [65, 66]. Genomes of modern angiosperms contain remnants of multiple rounds of past polyploidization events, often followed by extensive genomic reorganisation, massive silencing and elimination of duplicated genes. Some plants exist as stable polyploids including, for example, a large portion of our most important crops such as wheat, maize, soybean, cabbage, oat, sugar cane, alfalfa, potato, coffee, cotton and tobacco [67–69], but it is believed that polyploidy is a transitory state and that polyploid genomes are bound to return to a functional diploid state through a process called diploidisation. Most, if not all, present-day diploid angiosperms thus are paleopolyploids.

An important benefit to becoming polyploid is heterosis, or hybrid vigour, manifested in increased size, fertility, biomass yield, growth rate or other parameters of the F1 organism over those of the diploid progenitors, resulting from the increase in heterozygosity [70]. Polyploid plants have, for example, increased potential to invade new niches and thereby enlarge their geographical and ecological range and reduce the risk of extinction (see [19] for more examples). Another key advantage is genetic redundancy that shields the polyploid from the deleterious effects of mutations [71–74]. Disadvantages of polyploidy include difficulties encountered during establishment [75], the propensity to produce aneuploid cells through abnormal division [76–78], disruption effects of nuclear and cell enlargement [79] and the epigenetic instability that results in non-additive gene regulation [80].

1.1.2 The fate of a duplicated gene

Different evolutionary models have been proposed to explain the functional divergence of duplicated genes (see Figure 1.2). Population genetics shows that the vast majority of duplicates is rapidly non-functionalised right after gene duplication [81–83]. One of the copies gets either physically lost from the genome due to genome rearrangements, or degrades into a pseudogene because of the accumulation of deleterious mutations (Figure 1.2, bottom row) or the absence of regulatory sequences.

A classical model put forward by Ohno holds that gene duplication creates

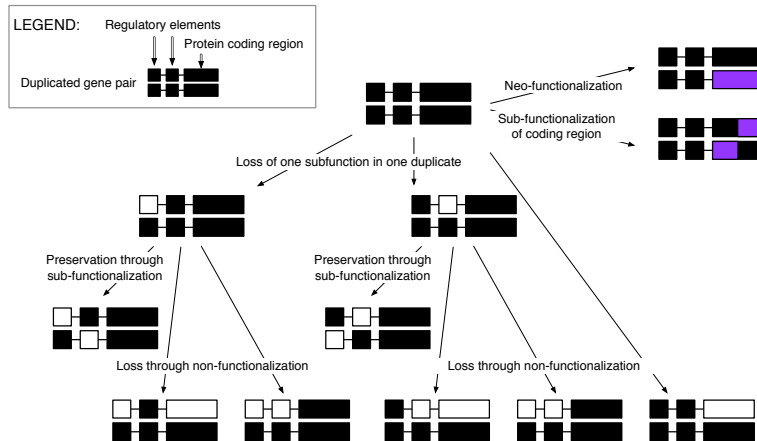


Figure 1.2: Possible functional fates of a duplicate gene pair. Following duplication, one of the copies is most likely to get non-functionalised (bottom right). Novel gene function can arise either through neo-functionalisation, where the novel duplicate acquires a new function through rare advantages mutations that are sustained by the virtue of the redundancy, or through sub-functionalisation of the coding region in case of genes coding for multi-domain proteins (top right). One subfunction in one duplicate can be lost by the inactivation of a regulatory element in the promoter region of a pleiotropic gene pair (left down). Subsequent mutations most likely bring about the non-functionalisation of one copy (bottom). However, subsequent complementary inactivation of regulatory elements may partition the functions of the ancestral gene between both duplicates, leaving both duplicates intact. Squares and rectangles denote respectively regulatory elements and coding regions. Functional and non-functional elements are respectively coloured black and white. An element with a modified function is depicted in purple.

functionally redundant loci where one is free to evolve under lack of functional constraint and to acquire a new function by random non-deleterious mutations, as long as the other remains to perform the ancestral task [24] (Figure 1.2, top right). Examples of neo-functionalisation of both the regulatory and the coding region have been found [84–86]. Given the little evidence that has been found for genes to have obtained novel functions in this manner, the large number of duplicated genes in most eukaryotic genomes, and the alterations that have been put forward to this model [87, 88], various alternative models were suggested [28, 73, 87, 89–91]. An obvious one being redundancy, where the two genes have divergent functions but maintain a partial or complete functional overlap [92–94], is the condition where both gene copies retain the ancestral gene’s function and are equally maintained such that the copies can substitute for each other. An important advantage of this condition seems genetic robustness against mutations [95, 96], but because redundancy is regarded as evolutionary unstable, most models for conserving

redundancy over the course of evolution require some degree of symmetry-breaking, one of the most important ones being the sub-functionalisation model by Lynch and Force [97] (Figure 1.2, down left from the initial state). These authors proposed that functional novelty acquired by one or both of the duplicates consists of a specialisation of its activity to particular developmental stages and tissues, due to complementary loss of regulatory elements in their promoter region. The pleiotropy of the ancestral gene is thereby lost and both copies are necessary in order to maintain the full arsenal of genetic functions.

In spite of the understanding that duplication of genes is of paramount importance in providing raw materials for the evolution of organisms and for genetic diversity, and that several hypotheses predicting the outcome of a duplication event have been proposed and reviewed, the relative prevalence of the different outcomes and factors playing in the process remain unknown.

1.2 *Arabidopsis thaliana*: the weed that made it to model organism

Arabidopsis thaliana is a small flowering plant with a broad natural distribution throughout Europe, Asia and North America. It is a member of the Brassicaceae or mustard family, which also includes cultivated members like cabbage, broccoli, cauliflower, Brussels sprouts, wasabi, horseradish and turnip. In 2000, *Arabidopsis* had the unique honour of being the first plant whose genome was fully sequenced [7]. Although not of major agronomic significance, different characteristics have made this dicotyledonous angiosperm the model system of choice for research in plant biology [98]: it has a small genome of 125 megabases and a short generation time of only six weeks, it is easily transformable, its small size allows plants to be grown in a greenhouse or laboratory, mature plants produce siliques with a large number of seeds and its natural pathogens include a variety of insects, bacteria, fungi and viruses. Knowledge gained from this organism has greatly increased our understanding in gene function and regulation, development, resistance to biotic and abiotic stress and metabolism of other plants, like economically important crops.

Notwithstanding a small genome size, detailed sequence similarity analyses revealed three genome duplication events in *Arabidopsis* [7,83,99–101]. The complex duplication structure is presented in Figure 1.3, where the five chromosomes of *Arabidopsis* are shown as horizontal lines and genes with a sequence identity of 80% or higher are connected by coloured lines. The widespread presence of duplicates in the *Arabidopsis* genome assumes the study of their dynamics of even



Figure 1.3: Segmental duplications in the *Arabidopsis* genome. The horizontal, grey bars represent its five chromosomes; the coloured lines connect genes with a sequence identity of 80% or higher. Image generated by GenomePixelizer, and reproduced with kind permission of Alexander Kozik.

greater significance when it is considered that most, if not all, of our important crop species also share this characteristic.

1.3 Microarrays

Microarrays constitute a prominent example of a technology that has emerged after the genome sequencing projects of the mid-1990s to facilitate the execution of experiments on a large number of genes simultaneously [102, 103]. Through gene expression profiling, the microarray technology provides biologists deep insights into the molecular functionality of genes. It aims to measure mRNA levels in cell or tissue samples, at a particular time or under a particular condition. *In situ* synthesised types of microarrays exploits to this end the recently obtained sequence resources, while arrays based on PCR-product libraries only require a short sequences of a particular cDNA clone, that contains the desired DNA fragment, for primer construction. Single strands of complementary DNA for the genes of interest ('reporters' [104]) are attached at fixed, known locations ('spots' or 'features') arranged in a grid ('array') on small solid supports, typically a glass slide or some other material, like a nylon membrane or a quartz wafer (left side of Figure 1.4). Features are either deposited on the support by a robot, synthesised by photo-lithography or printed by ink-jet printing. Current high density arrays contain up to 6.5 million features, each containing a huge number of identical molecules, of lengths ranging from twenty to hundreds of nucleotides. Ideally all genes of an organism of interest are represented on an array, but because of sequence similarity between gene family members, it is not always feasible to

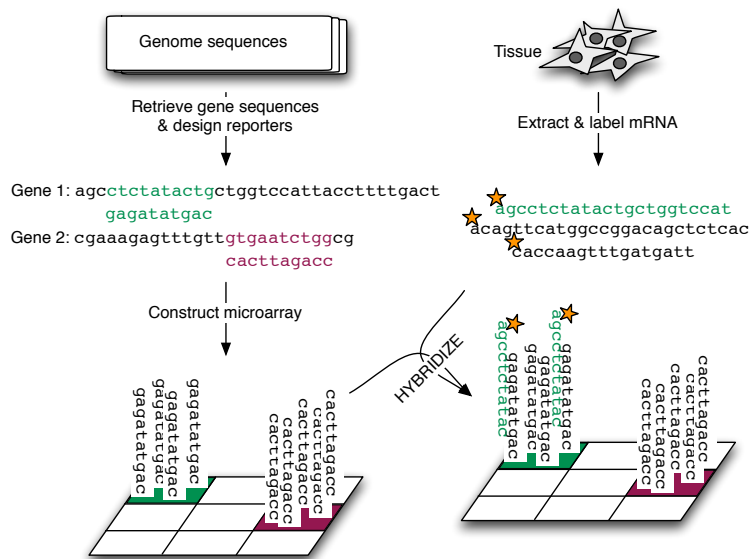


Figure 1.4: Overview of the design and use of a microarray. First, the genes in a sequenced genome are identified (left side). For every gene, a string of nucleotides unique in the genome is searched for and deposited in large amounts on a feature on a solid support. Expression of these genes is measured by extracting mRNA from a sample of interest (right side), labelling it and hybridising it to the array (bottom right). In this particular example, of the two genes shown on the array, only gene 1 is expressed.

identify gene-specific regions for each individual gene. A sample is prepared by extraction of the mRNA from cells of interest, removal of tRNA and rRNA, reverse transcription into cDNA, followed by transcription into RNA while attachment of a fluorescent or radioactive label and fragmentation. The sample is then hybridised to the array and incubated for 12 to 24 hours (right side of Figure 1.4). The microarray is then washed in order to remove RNA which has not hybridised to any reporter, or which has only hybridised weakly due to imperfect complementarity. Subsequent illumination with an appropriate light source allows the quantification of the label on each feature. Pre-processing of the data yields an intensity value for each gene that reflects the abundance of its corresponding mRNA in the sample.

With respect to the hybridised sample, two main types of microarrays exist: two-channel and single-channel arrays. The two-channel arrays are hybridised by a mix of two samples, for example the control labelled with fluorophore Cy3 (green) and the experiment sample with fluorophore Cy5 (red). Molecules of both samples will bind the few complementary reporters on the array in a

competitive manner. The array is then scanned for the two fluorophores separately and relative intensities of each are calculated. These arrays were important in the early developments of the microarray technology, as spots suffered from irregularities in shape, size and reporter density, which rendered the ability to contrast two measurements from the same spot as crucial. Single-channel arrays are hybridised by one sample only. Absolute intensity values are obtained from this type of arrays. Another way of distinguishing microarrays is by the type of reporter: cDNA microarrays are chips whose features contain complementary DNA (cDNA), typically generated via PCR amplification and attached on the slide by printing or through electrostatic attraction. High density oligonucleotide array contain short synthesised oligos, either *in situ* synthesised or deposited.

Numerous variations to the construction of the array and the protocol to generate the hybridisation liquid exist. In what follows, only the Affymetrix GeneChip technology will be introduced in more detail, since data generated with this type of array are the subject of research conducted for this dissertation.

1.3.1 GeneChip basics

The GeneChip technology was invented by a team of scientists in the late-1980s and was the basis for the founding of a new company, Affymetrix, in 1991. Thanks to the standardised procedures of array production, labelling, hybridisation and data analysis, these high-density oligonucleotide arrays are highly popular and are used world-wide by pharmaceutical, biotechnological, agrochemical, diagnostics and consumer product companies and academic, governmental and other non-profit research institutes to analyse gene expression. On their traditional Gene Arrays, a locus to be interrogated was represented on such an array by a probe set that consists of 11 to 20 perfect match and mismatch pairs (top of Figure 1.5) [105]. On their novel array types, like Exon Arrays, reporters are designed to interrogate the entire length of a gene. The reporters are small DNA fragments of 25 nucleotides that are synthesised during a photolithographic process at specific locations on a coated quartz surface. A perfect match reporter perfectly matches its target sequence, while the paired mismatch reporter contains a single mismatch located in the middle of its sequence (bottom of Figure 1.5). The mismatch reporters are designed by Affymetrix to quantify the amount of non-specific binding in their corresponding perfect match reporter, or to determine whether a gene is turned on or off in the investigated condition¹. A mismatch reporter is located one row below its corresponding perfect match and, to avoid bias introduced by spatial effects, the different probe pairs of a probe set are dispersed

¹Affymetrix offers a presence/absence algorithm which makes use of the difference between the observed intensity value of the perfect match and its corresponding mismatch

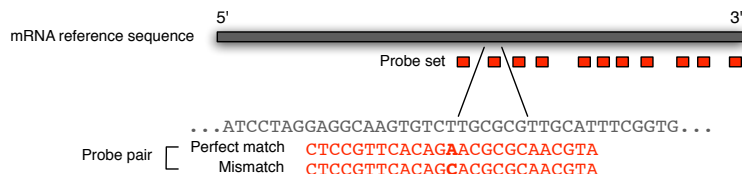


Figure 1.5: Basics of the Affymetrix GeneChip technology. A gene locus is represented on the GeneChip by a set of probe pairs that consist of a perfect match and a mismatch reporter. The sequence of the perfect match reporter is complementary to the locus' reference sequence, while the mismatch has one mismatch nucleotide in the middle.

across the array.

1.3.2 GeneChip data pre-processing

Substantial data pre-processing is required in order to obtain an accurate assessment of the expression level for a specific gene and involves different main steps: image analysis, background adjustment, normalisation, summarisation and quality assessment [106].

Image analysis

During image analysis, the raw one-channel pixel intensities produced by the scanner, contained in a '.DAT' file, are converted into numerical values. The output, the single-number intensity summary combining all pixels in a given feature, is saved to a '.CEL' file. The process involves gridding, to locate each feature on the slide, and segmentation, to divide a feature-containing region into foreground and background. Various sources describe this process in detail [107, 108]. Affymetrix GeneChip scanners are standardly equipped with this software, so this step is usually integrated in the early data generation.

Background adjustment

Background adjustment is conducted to calculate the background signal due to non-specific hybridisation and optical noise so as to obtain an accurate estimate of specific binding. For spotted arrays, the pixels of each spot region are divided in the spot itself and those in the background. There are a number of methods for doing this. For more detailed information, see [107, 109]. Since features on Affymetrix chips are densely packed, the intensities of the cells themselves, rather than the region adjacent to the features, as in cDNA arrays, are used to estimate the background intensity.

Normalisation: Getting the numbers comparable

The ultimate goal of working with microarrays is to make arbitrary comparisons between the gene expression levels of different samples. Various sources of variation, like chip processing, mRNA preparation, amplification, hybridisation, scanner settings, grid placement, segmentation and feature quantification obstruct this aim and add systemic variation to the data. Normalisation is the process of removing variation that might exist between arrays in a microarray experiment and that is caused by technology or by sample handling and preparation, rather than from biological differences. Many different algorithms have been developed to normalise microarrays, depending on the type and the aim of the study. Specific normalisation procedures include, for example, local print-tip normalisation for spotted microarrays, global normalisation to correct for incorporation efficiencies or scanning properties of the two dyes for cDNA arrays. Additional between-slide scaling can be conducted when large scale differences between different slides of one experiment exist. The reader is referred to specialised literature for detailed explanation regarding cDNA arrays [110–114]. Normalisation methods for Affymetrix GeneChips will be explained in detail below.

Summarisation: Obtaining an overall expression intensity

A gene can be represented on a microarray by one sole reporter, or alternatively, like on Affymetrix GeneChips, by a set of reporters, which offers the potential to calculate statistics and confidence about the measurements and results of the experiment. The summarisation step aggregates these multiple intensity values into one single expression value for the particular gene. Different outlier robust methods have been implemented in the various processing algorithms.

1.3.3 Data pre-processing algorithms

Various methods have been devised for each of the above steps for the pre-processing of raw Affymetrix GeneChip data [115–119]. Different software tools integrate the different steps into one algorithm. Only the tools used for this dissertation will be discussed below.

Single-array approach: MicroArray Suite 5.0

The Microarray Suite 5.0 (MAS5.0) is a method developed by manufacturer Affymetrix [118, 120]. Expression value calculation involves background correction, which is comprised of global background correction and perfect match intensity correction, a summarisation step and global scaling ².

²Currently, Affymetrix incorporated observations made by different research groups and moved to a more sophisticated approach, called PLIER. PLIER, or the Probe Logarithmic Intensity ERror,

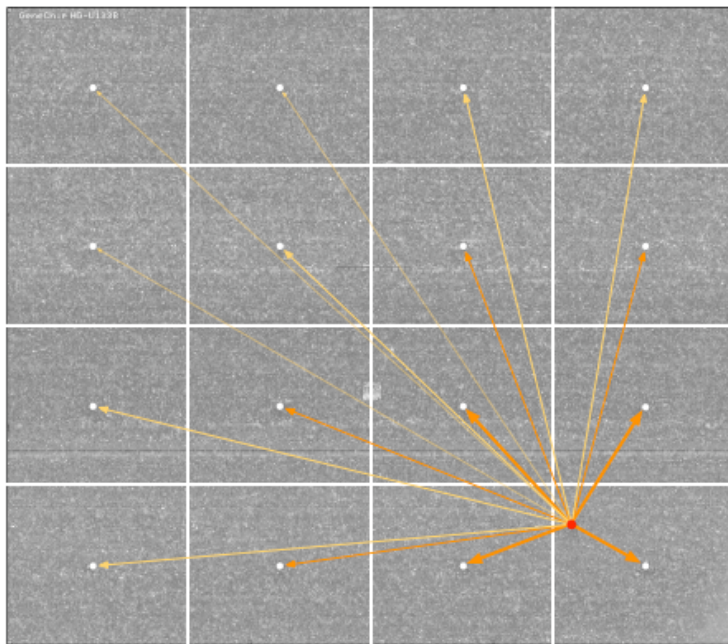


Figure 1.6: Background correction in Affymetrix' MAS5.0. First, the microarray is divided into 16 zones (white grid) and the mean of the second percentile of all intensity values in each zone is calculated (white dot). Every cell is then background corrected by subtracting the inverse-distance weighted trimmed average from every zone. The red dot is the cell for which the background correction is presented here: the width and intensity of orange of the arrows denote the weighting of each zone's mean on this particular cell in this process.

For global background correction, both the perfect match and mismatch reporter intensity values are used, as well as the location of the reporter on the array. First, the array is divided into different zones (default 16) for each of which a background and noise value is obtained by calculating the average of the second percentile of all intensities. Subsequently, each cell is background adjusted by computing a weighted sum of the background and noise value, where the weighting depends on the inverse distance to each of the 16 zone centres. This step is illustrated in Figure 1.6. The background is subtracted from the raw intensity. Unless this would lead to a value less than the noise value, in which case the

produces an improved signal by accounting for experimentally observed patterns for feature behaviour and handling error appropriately at low and high abundance. See PLIER for more details.

reporter intensity would be replaced by the noise value. Important adjustment with respect to background and non-specific hybridisation of the feature-level scores is made in the next step by calculating and subtracting an Ideal Mismatch value for each of the perfect match values. These Ideal Mismatch values are calculated to guarantee a positive value after perfect correction. A study on a typical GeneChip in 2001 revealed that as many as 30% of the mismatch reporters have an intensity value that is higher than its corresponding perfect match [121] and thus render negative values after subtraction. These negative values rule out the use of in microarray analysis widely-used logarithms. To solve this problem, Affymetrix came up with their Ideal Mismatch, where they propose to use the mismatch intensity when it is smaller than the perfect match or a truncated value in other cases. The data is then log transformed and per probe set, the reporters' intensity values are summarised with the one-step Tukey's bi-weight algorithm that provides an outlier robust estimator of the mean. This estimate is done for each probe set on each array separately. The signal output is the anti-log of the resulting value. Finally, global scaling is done by taking the trimmed mean (2%) of all intensity values on the array. For a more detailed description, see [118, 120].

Multi-array approach: RMA

The Robust Multi-Array algorithm developed by Irizarry and colleagues [115, 122, 123] is basically a three step procedure of background adjustment on each array, normalisation across arrays and summarisation per probe set.

Because the use of the mismatch reporters for Affymetrix' original purpose has been shown to be problematic ([121] and above) and their exclusion has shown to lead to expression values and fold change estimates with decreased variance [124], the authors of this algorithm chose to make use of perfect match intensities only and to ignore mismatch reporters. The background is estimated based on an additive background noise and the true signal model: $O = N + S$, where the background noise: $N \sim N(\mu, \sigma^2)$ and the true signal: $S \sim \exp(\alpha)$, with S and N independent of each other. Given O , the observed intensity values are adjusted by replacing them with the expected signal as follows:

$$E(S|O = o) = a + b \frac{\phi(\frac{a}{b}) - \phi(\frac{o-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{o-a}{b}) - 1}$$

where Φ and ϕ are the standard normal distribution function and density function, respectively, and $a = o - \mu - \sigma^2\alpha$ and $b = \sigma$. The parameters to obtain these values are estimated in an ad-hoc way using all perfect match intensities. The mode of the non-parametric density estimate of these intensities is used as an estimate of μ . The variability of the lower tail about μ is used for σ and an exponential is fitted to the upper tail to estimate α . For more detailed information the reader is referred

to [125].

The normalisation method implemented in RMA is quantile normalisation, which is aimed at giving the same empirical distribution of intensities to each array in the experiment. The basic assumption that is made is that the intensities of each array originate from the same underlying distribution. The normalisation distribution itself is obtained by averaging each quantile across the different arrays.

Summarisation is done by robustly fitting a linear model to the logarithm of the pre-processed perfect match intensities for each probe set over the different arrays:

$$\log_2(y_{ij}) = \mu + \beta_j + \alpha_i + \epsilon_{ij}$$

where y_{ij} is the normalised, background corrected intensity value of perfect match probe i on array j , α_i is the probe effect, β_j is the array effect and μ is the wanted expression value for the respective probe set on array j . The median polish method [126] is used to fit the model which proceeds by alternatively subtracting the row and column median of each of the values in the y_{ij} matrix, with the constraints that $\text{median}(\beta_j) = \text{median}(\alpha_i) = 0$ and $\text{median}_i(\epsilon_{ij}) = \text{median}_j(\epsilon_{ij}) = 0$. For more detailed information the reader is referred to [125] and [115, 122, 123].

The effect of normalisation on biological data

Investigating intensity distributions of a set of microarrays and evaluating the effectiveness of normalisation methods can be done with visualisation tools like boxplots and scatter plots. This is demonstrated for an experiment where triplicate samples were taken from 14 different plant tissues. The different rows in Figure 1.7 show the raw data (top), MAS5.0 normalised data (middle) and RMA normalised data (bottom). The first column contains boxplots of the expression values of all 42 slides. The data of every tissue is shown in a distinct colour with each of its three replicates all having the same colour. The second column contains for two replicates of sample 1 so-called MvA plots, in which the log ratio of the intensity of these two replicates is plotted in function of the average intensity. The red lines in these three plots denote the ideal situation where the intensities of a gene in both replicates are identical.

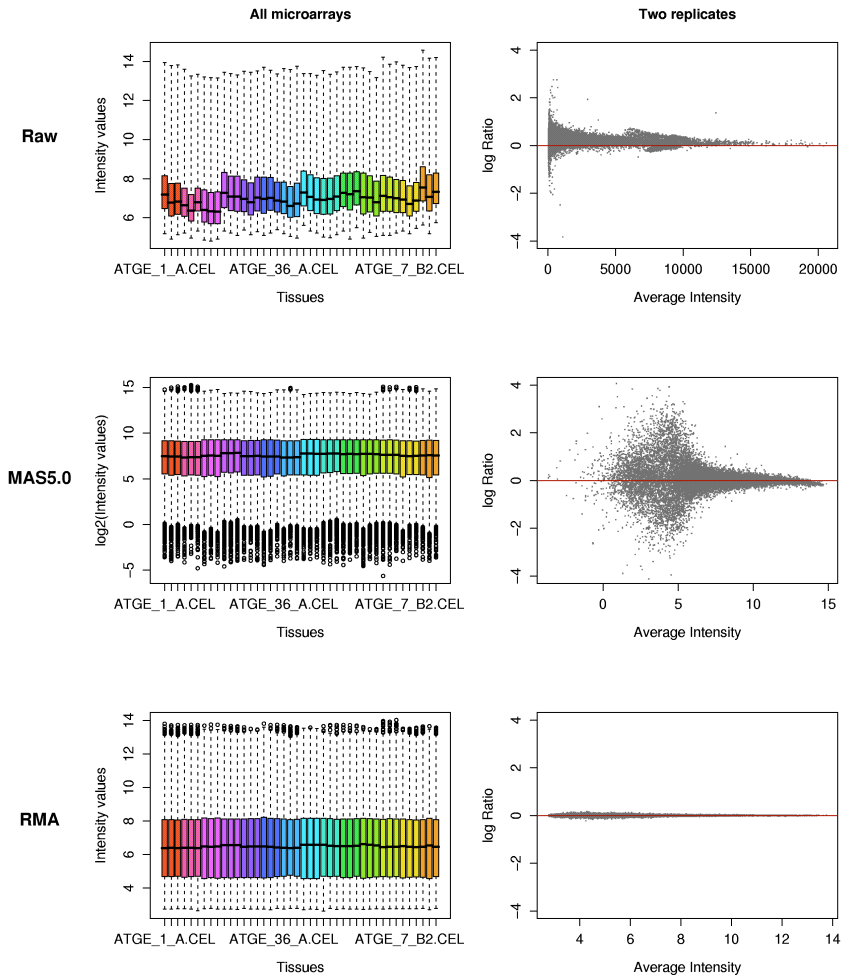


Figure 1.7: Effect of MAS5.0 and RMA normalisation. The expression data set is sampled from 14 plant tissues, each replicated three times. The three rows contain the raw data, MAS5.0 normalised and RMA normalised data. The first column contains boxplots, where the distribution of each of the 14 samples is plotted in a distinct colour and each of the triplicates in the same colour. The second column shows MvA plots of two replicates of tissue 1, where the log ratio of the data of these two replicates is plotted in function of their average intensity. The red line in all three plots depicts the ideal situation where a gene has the same intensity in both replicates.

The boxplot of the raw data reveals that the distributions and the median intensity value are different for the different arrays in this experiment. MAS5.0 normalisation succeeds in centering the distributions of intensity values of the different arrays. RMA is also effective in that respect, but is in addition effective in making the distributions of intensity values of the different arrays similar.

The MvA plot of the raw data for the two replicates reveal that many probes in one replicate have lower intensity values than in the other (since many of the dots are above the horizontal). The variability of the data is also not equal for different intensity values (vertical spread along the horizontal). The diagnostic MvA plots in row two and three, for MAS5.0 and RMA allow to compare the two normalisation methods in more detail. In the ideal case of sample replication, a probe set should have a highly similar intensity value on both arrays and hence have a log ratio of 0 subsequent to normalisation. The MvA plot of the MAS5.0 data reveals that the variability is strongly intensity-dependent: it is greater for lower than higher average intensity values. Intensity-dependent variability poses a serious concern for statistical methods that assume common variance. MAS5.0 does with respect to intensity-depend bias an adequate job: only on the very high intensity values, the dots lie below the horizontal. The bottom MvA plot reveals that RMA performs better at both criteria. The biggest concern that these data reveal in this figure is that the variance is considerably larger for MAS5.0 than for RMA: while RMA produces clean estimates for the equal intensity values in the two replicates, MAS5.0 produces a very noise estimate, with a slight bias at the high intensities. This noise can be largely attributed to the use of mismatch values, which is in fact the reason for ignoring them in RMA.

1.4 Putting the pieces together

Inferred from cytological experiments many decades ago and confirmed by recent genome sequencing projects, the high prevalence of redundant gene copies is a hallmark of eukaryotic genomes. Since decades, gene duplication has been granted acknowledgement as of paramount importance for evolutionary transitions and

increase in organismal complexity. Various models predicting the outcome of a gene following duplication have been put up and examples have been described, but their applicability on a genome-wide scale remains to be investigated. The microarray technology provides a genome-wide source of functional data and allows investigation of the relative prevalence of the different fates and factors proposed to be playing in these processes.

The first part of this dissertation, Chapter 2, focuses on the large set of duplicated genes in *Arabidopsis thaliana*. Different kinds of duplicates are identified, depending on whether the copy came about by large- or small-scale duplication, hence the mode of duplication, the time since the duplication event, and the functional class the gene pair belongs to. Expression correlation and tissue expression patterns measured with Affymetrix GeneChips are used to assess the extent to which these different factors influence the divergence rate and whether expression divergence is biased towards certain classes of duplicates.

The availability of genome sequences and gene expression information also creates the opportunity to unravel gene expression regulation, which is the result of the complex and tightly regulated interplay of many different constituents, like chemical or structural modification of the DNA, transcriptional and translational control, mRNA degradation and post-translational modification. The role of transcriptional regulation is played by transcription factors, who direct the timing and location of transcriptional activity by either inducing or repressing transcription. They do this by binding to specific parts of the DNA, so-called transcription factor binding sites, that are primarily located in the long non-coding sequences upstream of a gene. They function alone or in complex as different binding sites can organise into *cis*-regulatory modules that each integrate the input from a specific set of cooperating transcription factors. Identification of transcription factor binding sites and their organisation into modules is important in understanding the process of gene regulation. Genes that show co-expression are likely co-regulated and hence are likely to share regulatory elements and/or modules. Chapter 3 describes the identification of novel regulatory motifs in sets of co-expressed genes that are delineated by means of microarray data in combination with comparative genomics.

Microarrays are valuable instruments for obtaining gene co-expression relationships on a genome-wide scale. Inference tools, for instance of functional modules and regulatory networks, in systems biology are often based on such relationships. In Chapter 4, we investigate the basic assumption that correlated microarray signal profiles indicate biological co-expression of the target genes. In addition to genuine biological co-expression, signal correlation can also result

from cross-hybridisation. In this study, we investigate the nature and prevalence of this problem on a large scale, by studying the extent to which so-called gene-specific reporters bind off-targets and thereby lead to spurious down-stream correlations. We also describe a novel method for diagnosing individual probesets that are likely affected by off-target hybridisation.

2

Non-Random Divergence of Gene Expression Following Gene and Genome Duplications in the Flowering Plant *Arabidopsis thaliana*

Tineke Casneuf, Stefanie De Bodt, Jeroen Raes, Steven Maere and
Yves Van de Peer
Genome Biology (2006) 7, R13

Divide et impera

Philip II of Macedonia

Abstract

Genome analyses have revealed that gene duplication in plants is rampant. Furthermore, many of the duplicated genes seem to have been created through ancient genome-wide duplication events. Recently, we have shown that gene loss is strikingly different for large- and small-scale duplication events and highly biased towards the functional class to which a gene belongs. Here, we study the expression divergence of genes that were created during large- and small-scale gene duplication events by means of microarray data and investigate both the influence of the origin (mode of duplication) and the function of the duplicated genes on expression divergence. Duplicates that have been created by large-scale duplication events and that can still be found in duplicated segments have expression patterns that are more correlated than those that were created by small-scale duplications or those that no longer lie in duplicated segments. Moreover, the former tend to have highly redundant or overlapping expression patterns and are mostly expressed in the same tissues, while the latter show asymmetric divergence. In addition, a strong bias in divergence of gene expression was observed towards gene function and the biological process genes are involved in. By using microarray expression data for *Arabidopsis thaliana*, we show that the mode of duplication, the function of the genes involved, and the time since duplication play important roles in the divergence of gene expression and, therefore, in the functional divergence of genes after duplication.

2.1 Background

Recent studies have revealed a surprisingly large number of duplicated genes in eukaryotic genomes [82, 127]. Many of these duplicated genes seem to have been created in large-scale, or even genome-wide duplication events [128, 129]. Whole genome duplication is particularly prominent in plants and most of the angiosperms are believed to be ancient polyploids, including a large proportion of our most important crops such as wheat, maize, soybean, cabbage, oat, sugar cane, alfalfa, potato, coffee, cotton and tobacco [67–69, 130]. For over 100 years, gene and genome duplications have been linked to the origin of evolutionary novelties, because it provides a source of genetic material on which evolution can work ([19] and references therein). In general, four possible fates are usually acknowledged for duplicated genes. The most likely fate is gene loss or non-functionalisation [24, 82, 83, 97], while in rare cases one of the two duplicates acquires a new function (neo-functionalisation) [19]. Sub-functionalisation, in which both gene copies lose a complementary set of regulatory elements and thereby divide the ancestral gene's original functions, forms a third potential fate [89, 131–133]. Finally, retention is recognised for two gene copies that, instead

of diverging in function, remain largely redundant and provide the organism with increased genetic robustness against harmful mutations [95, 134, 135].

The functional divergence of duplicated genes has been extensively studied at the sequence level to investigate whether genes evolve at faster rates after duplication, or are under positive or purifying selection [136–141]. The recent availability of functional genomics data, such as expression data from whole-genome microarrays, opens up completely novel ways to investigate the divergence of duplicated genes. Several studies using such data have already provided intriguing new insights into gene fate after duplication. In yeast, for instance, Gu and co-workers [142] found a significant correlation between the rate of coding sequence evolution and divergence of expression and showed that most duplicated genes in this organism quickly diverge in their expression patterns. In addition, they showed that expression divergence increases with evolutionary time. Makova and Li [143] analysed spatial expression patterns of human duplicates and came to the same conclusions. They calculated the proportion of gene pairs with diverged expression in different tissues, and found evidence for an approximately linear relationship with sequence divergence. Wagner [144] showed that the functional divergence of duplicated genes is often asymmetrical because one duplicate frequently shows significantly more molecular or genetic interactions/functions than the other. Adams and co-workers [145] examined the expression of 40 gene pairs duplicated by polyploidy in natural and synthetic tetraploid cotton and showed that, although many pairs contributed equally to the transcriptome, a high percentage exhibited reciprocal silencing and biased expression and were developmentally regulated. In a few cases, genes duplicated through polyploidy events were reciprocally silenced in different organs, suggesting sub-functionalisation.

In *Arabidopsis*, Blanc and Wolfe [146] investigated the expression patterns of genes that arose through gene duplication and found that about 62% of the recent duplicates acquired divergent expression patterns, which is in agreement with previous observations in yeast and human. In addition, they identified several cases of so-called 'concerted divergence', where single members of different duplicated genes diverge in a correlated way, resulting in parallel networks that are expressed in different cell types, developmental stages or environmental conditions. Also in *Arabidopsis*, Haberer et al. [147] studied the divergence of genes that originated through tandem and segmental duplications by using massively parallel signature sequencing (MPSS) data and concluded that, besides a significant portion of segmentally and tandemly duplicated genes with similar expression, the expression of more than two-thirds of the duplicated genes diverged in expression. However, expression divergence and divergence time were not significantly correlated, as

opposed to findings in human and yeast (see above). In a small-scale study on regulatory genes in *Arabidopsis*, Duarte et al. [148] performed an analysis of variance (ANOVA) and showed that 85% of the 280 paralogs exhibit a significant gene by organ interaction effect, indicative of sub- and/or neo-functionalisation. Ancestral expression patterns inferred across a type II MADS box gene phylogeny indicated several cases of regulatory neo-functionalisation and organ-specific non-functionalisation.

In conclusion, recent findings demonstrate that a majority of duplicated genes acquire different expression patterns shortly after duplication. However, whether the fate of a duplicated gene also depends on its function is far less understood. The model plant *Arabidopsis* has a well-annotated genome and, in addition to many small-scale duplication events, there is compelling evidence for three genome duplications in its evolutionary past [99–101, 149], hereafter referred to as 1R, 2R, and 3R. Recently, a non-random process of gene loss subsequent to these different polyploidy events has been postulated [83, 146, 150]. Maere et al. [83] have shown that gene decay rates following duplication differ considerably between different functional classes of genes, indicating that the fate of a duplicated gene largely depends on its function. Here, we study the expression divergence of genes that were created during both large- and small-scale gene duplication events by means of two compiled microarray datasets. The influence of the origin (mode of duplication) and the function of the duplicated genes on expression divergence are investigated.

2.2 Results and Discussion

To examine general gene expression divergence patterns, we analysed two datasets containing genome-wide microarray data for *Arabidopsis* genes (see Methods). The first consisted of 153 Affymetrix ATH1 slides with expression data of various perturbation and knockout experiments (see Additional data file 1). The Spearman rank correlation coefficient was computed between the two expression patterns of every duplicated gene pair. To investigate whether divergence of gene expression varies for duplicates that were created by small-scale or large-scale (genome-wide) events, the complete set of duplicated genes was subdivided into different subgroups and their expression correlation was examined (see Methods; Figure 2.1).

We refer to anchor genes as duplicated genes that are still lying in recognisable duplicated segments. Such anchor-point genes, and consequently the segments in which they reside, are regarded as being created in large-scale duplication events. Six different sets of genes were distinguished: one set containing duplicates

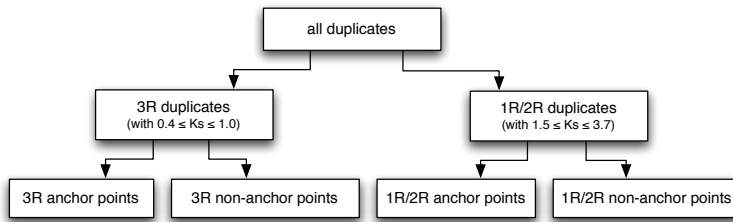


Figure 2.1: Six subclasses of duplicated genes in *Arabidopsis thaliana*. The duplicated genes of *Arabidopsis thaliana* were divided into six different subclasses according to the time and mode of duplication (see Methods for details)

with ages corresponding to 1R/2R ($1.5 \leq K_S \leq 3.7$), further subdivided into two sets of anchor and non-anchor points, and one set of younger duplicates with ages corresponding to 3R ($0.4 \leq K_S \leq 1.0$), again subdivided into two sets of anchor and non-anchor points (see Methods). Differences in expression divergence between anchor points and non-anchor points were evaluated by comparing their distributions of correlation coefficients using a Mann Whitney U test (see Methods). We further explored the difference between both classes of genes by means of a second dataset on tissue-specific expression (see Methods and Additional data file 2) [151]. Here, for each of the subgroups of duplicates described above we calculated present/absent calls in the 63 different tissues and computed both the absolute and relative amount of tissues in which the two genes of a duplicated gene pair are expressed.

In addition, the first dataset was used to identify possible biases toward gene function. The expression correlation of duplicated gene pairs, represented by the Spearman correlation coefficient, was studied in relation to the age of duplication, represented by K_S (amount of synonymous substitutions per synonymous site) for genes belonging to different functional categories (GO slim, see Methods).

2.2.1 Divergence of expression and mode of duplication

First, we investigated whether the mode of duplication that gives rise to the duplicate gene pairs affects expression divergence. Interestingly, for both younger (Figure 2.2A) and older (Figure 2.2B) duplicates, anchor points showed a significantly higher correlation in expression than non-anchor points (p values of $2.49e^{-07}$ and $1.67e^{-08}$ for young and old genes, respectively). Even for the younger duplicates the difference is striking (Figure 2.2A).

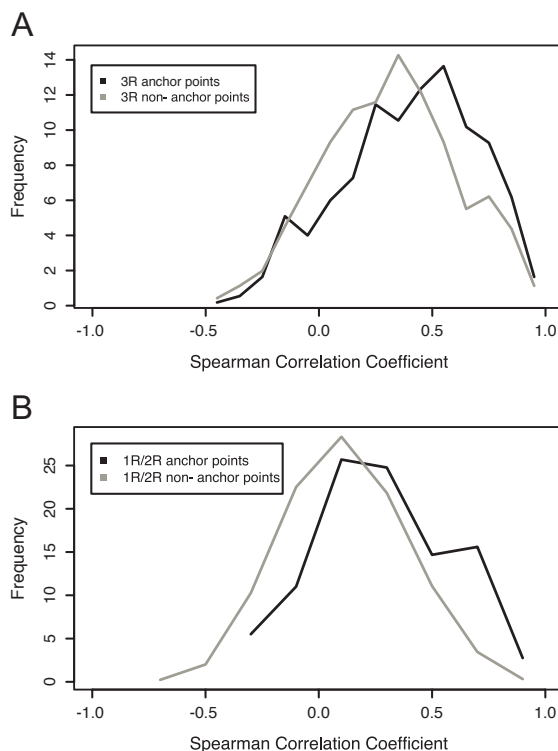


Figure 2.2: Expression correlation for anchor points and non-anchor points. Histograms of the Spearman correlation coefficients for anchor points (black) and non-anchor points (grey) for both (A) 3R genes and (B) 1R/2R genes. A Mann-Whitney U test was used to test whether both distributions are significantly different from each other. Mean correlation coefficients: 0.40 for 3R anchor points; 0.32 for 3R non-anchor points; 0.28 for 1R/2R anchor points; and 0.11 for 1R/2R non-anchor points.

We explored the second dataset on tissue-specific expression and first considered the absolute number of tissues in which genes are expressed, resembling the expression breadth (see Methods). Regarding anchor points, both genes are usually expressed in a high number of tissues (Figure 3A). This is only partly true for non-anchor points (or genes assumed to have been created in small-scale duplications), where many duplicates are expressed in a much smaller number of tissues (shown for young duplicates in Figure 3B). To further discriminate between redundancy, complementarity and asymmetric divergence, and thus to investigate if genes are expressed in the same tissues, we computed the relative number of tissues a gene is expressed in, which is the number of tissues in which a gene is expressed

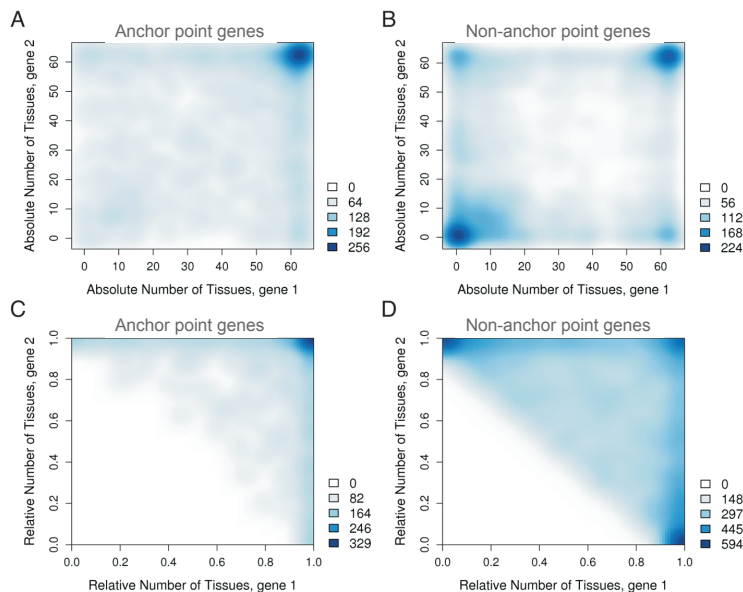


Figure 2.3: Tissue co-expression of duplicated gene pairs. Smoothed colour density representations of the scatterplots of the (A,B) absolute and (C,D) relative numbers of tissues in which the genes of a duplicated gene pair are expressed, for both (A,C) 3R anchor points and (B,D) non-anchor points. From (A,C) we can conclude that many anchor point genes are both expressed in a high number of tissues, and that many of these tissues are actually identical. On the other hand, (B,D) show that non-anchor point genes frequently show asymmetric divergence because many genes are expressed in a high number of tissues, while their duplicate is not. The plots were made using the 'smoothScatter' function, implemented in the R package 'prada' [152], by binning the data (in 100 bins) in both directions. The intensity of blue represents the amount of points in the bin, as depicted in the legend.

divided by the total number of tissues in which either one of the two duplicates is expressed. As schematically represented in Figure 2.4, two duplicated genes that remain co-expressed in the same tissues will both have a relative number equal to 1 (redundant genes; Figure 2.4A), whereas asymmetrically diverged genes, where one gene is expressed in a very small number of tissues as opposed to its duplicate that is expressed in a high number of tissues, can be identified by relative numbers close to 0 and close to 1, respectively (Figure 2.4B). The intermediate situation, where two duplicate genes are expressed in an equal number of different tissues, will result in both copies having a relative number equal to 0.5 (Figure 2.4C). When assuming that the ancestral gene was expressed in all tissues in which the two duplicate genes are expressed, the latter case hints at sub-functionalisation

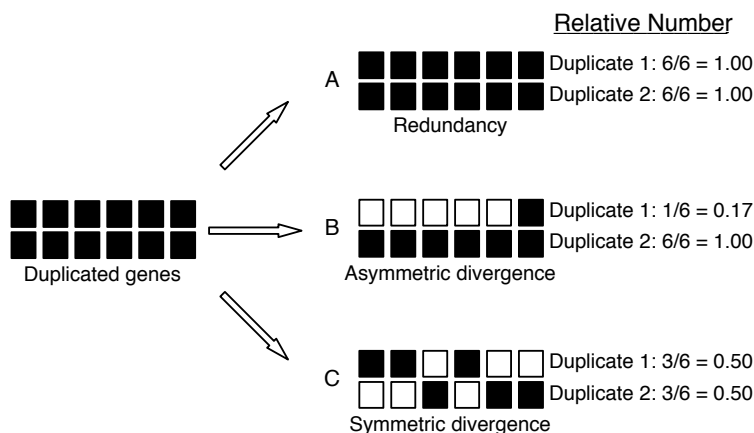


Figure 2.4: Possible scenarios for tissue-specific expression of a duplicated gene pair.

Hypothetical example showing possible scenarios for tissue-specific expression of two duplicates. A black box depicts expression in a particular tissue, whereas a white box represents no expression in that particular tissue. Following duplication of a gene that is expressed in six different tissues, the two copies can (A) both remain expressed in all six tissues (redundancy), (B) diverge asymmetrically, where one gene is expressed in only a small subset of the tissues, while its duplicate remains expressed in the original six tissues, or (C) diverge symmetrically, where tissue-specific expression is complementarily lost between both duplicates. The absolute number of tissues in which a gene is expressed is six for both duplicates in (A) and for the second duplicate in (B), one for the first duplicate in (B) and three for both duplicates in (C). The total number of tissues in which the pair is expressed is 6 in all three cases. The relative number is the fraction of the previous two, and is 1 for the two genes in (A) and for the second duplicate in (B), 0.17 for the first duplicate in (B) and 0.5 for both duplicates in (C).

after duplication. Figure 2.3C and D show these relative numbers for 3R anchor points and non-anchor points, respectively, and show that redundancy is much more common among anchor points (Figure 2.3C) than among nonanchor points (Figure 2.3D) of similar ages. Moreover, gene pairs resulting from small-scale duplications not only seem to have diverged more often than those created by segmental or genome duplications, but they also have diverged asymmetrically, where one gene is expressed in a high number of tissues, as opposed to its duplicate that is expressed in a small number of tissues (Figure 2.3D, top left and bottom right). Similar findings on tissue-specific expression were observed for the 1R/2R genes (results not shown).

The current study clearly shows that duplicated genes that are part of still recognisable duplicated segments (so-called anchor points) show higher correlation in gene expression than duplicates that do not lie in paralogs, despite their similar ages. In addition, the former have highly redundant or overlapping expression patterns, as they are mostly expressed in the same tissues. This is in contrast with what is observed for the non-anchor point genes, where asymmetric divergence is more widespread. There might be several explanations for these observations. The set of non-anchor point genes include genes created by tandem duplication, transpositional duplication, or genes translocated after segmental duplication events. One explanation might lie in different gene duplication mechanisms. Single-gene duplications, mostly caused by unequal crossing-over and duplicative transposition [153], are much more prone to promoter disruption than genes duplicated through polyploidy events, which might lead to the altered (or observed asymmetric) expression of genes after small-scale gene duplication events. Similarly, translocation of genes that originated from large-scale duplication events can also disrupt promoters, again contributing to the overall increase of expression divergence [88, 154].

Alternatively, the higher correlation of anchor points might result directly from co-expression of neighbouring genes, regardless of their involvement in the same pathway, as shown recently by Williams and Bowles [155]. It was also shown that genome organisation, and more in particular the chromatin structure, can affect gene expression [155–160]. Such additional structural and functional constraints might, therefore, reduce the freedom to diverge and, as a consequence, cause the expression patterns of genes in duplicated regions to remain similar, as observed here. Related to our observations, Rodin et al. ([161] and references therein) reported that position effects play an important role in the evolution of gene duplicates. Repositioning of a duplicate to an ectopic site is proposed to epigenetically modify its expression pattern, along with the rate and direction of mutations. This repositioning is believed to rescue redundant anchor point genes

from pseudogenisation and accelerate their evolution towards new developmental stage-, time-, and tissue-specific expression patterns [161].

As previously stated, non-anchor point genes not only appear to show higher expression divergence than anchor-point genes, they appear to diverge asymmetrically, where one gene is expressed in a high number of tissues, while its duplicate is expressed in a lower number of tissues. It should be noted that we cannot establish whether one duplicate is becoming highly specialised and dedicated to a very small number of tissues or whether it is losing much of its functionality (that is, turning into a pseudogene), nor can we distinguish between the gain of expression in new tissues for one gene versus the loss of expression for the other gene duplicate, as we would therefore need to know the expression pattern of the ancestral gene. In this respect, it is interesting to note that it is currently not known whether the ancient genome doublings in (the ancestor of) *Arabidopsis thaliana* resulted from auto- or allopolyploidisation. In the former case, the anchor point duplicates are in fact real paralogs, while in the latter case the expression of the two gene copies might have (slightly) differed from the start ([162, 163] and references therein). Nevertheless, our data clearly show that the duplicates that still lie in duplicated segments show high expression correlation and have highly overlapping expression patterns, as opposed to those that arose through small-scale duplication events or have been translocated afterwards.

In concordance with the results discussed above, Wagner [144] described asymmetric divergence of duplicated genes in the unicellular organism *Saccharomyces cerevisiae*. He reported that both the number of stressors to which two duplicates respond and the number of genes that are affected by the knockout of paralogous genes are asymmetric. He therefore proposed an evolutionary model in which the probability that a loss-of-function mutation has a deleterious effect is greatest if the two duplicates have diverged symmetrically. Asymmetric divergence of genes therefore leads to increased robustness against deleterious mutations. This seems to be confirmed by our results. Indeed, also in *Arabidopsis thaliana*, asymmetric divergence, rather than symmetric divergence, seems to be the fate for two duplicates, at least when they do not lie in duplicated segments.

2.2.2 Divergence of expression and gene function

Next, we studied how the expression correlation, measured as the Spearman correlation coefficient, changes over time for genes of ages up to a K_S of 3.7. Loess smoothers, which locally summarise the trend between two variables (see full black lines in Figure 2.5), clearly indicate that correlation of expression, in

general, is high for recently duplicated genes, declines as time increases, and saturates at a certain time point. Interestingly, considerable differences can be observed between genes belonging to different functional classes (Figure 2.5; Additional data file 3). For example, genes that are involved in signal transduction and response to external stimulus appear to have diverged very quickly after duplication (Figure 2.5A and B, respectively). Similar trends can be observed for genes involved in response to biotic stimuli and stress, cell communication, carbohydrate and lipid metabolism, and for genes with hydrolase activity (Additional data file 3). Interestingly, genes of many of these classes are involved in reactions against environmental changes or stress (signal transduction, cell communication, response to external and biotic stimuli and stress, lipid metabolism), which might suggest that *Arabidopsis* (or better its ancestors) quickly put these newborn genes into use by means of altered and diverged expression patterns, as compared to their ancestral copy, to survive and cope with environmental changes.

Slowly diverging expression patterns were found for proteins involved in, for example, macromolecule biosynthesis (Figure 2.5C) and structural molecule activity (Figure 2.5D) as reflected in the large number of young gene pairs with high correlation coefficients. Analogous trends can be observed for other functional classes containing genes involved in cell organisation and biogenesis, nucleic acid, macromolecule, protein and primary metabolism, biosynthesis and response to endogenous stimulus (Additional data file 3). Apparently, although duplicated genes within these classes are being retained, their fast diversification at the expression level is selected against, probably due to the essential nature and sensitive regulation of these highly conserved processes. Other classes of genes, like those having nucleotide binding capacity (Figure 2.5E) and those involved in regulation of biological processes (Figure 2.5F), show moderate divergence rates. The DNA binding, transcription, protein modification, and genes with catalytic, transcription factor and transporter activity (Additional data file 3) classes of genes show similar divergence patterns. We also tested whether the divergence patterns described above are significantly different from each other by interchanging the fitted models between functional classes (fit the locfit line of a particular class to the data of another class) and evaluating the model quality. Our results confirmed that there are indeed significant differences between slowly, moderately and quickly diverging genes (results not shown).

As opposed to Haberer et al. [147], but in agreement with Gu et al. [142] and Makova and Li [143], who described expression divergence of duplicated genes in yeast and human, respectively, we here show that in *Arabidopsis*, expression patterns of duplicates diverge as time increases. In addition, the rate of divergence seems to be highly dependent on the molecular function of the gene or the

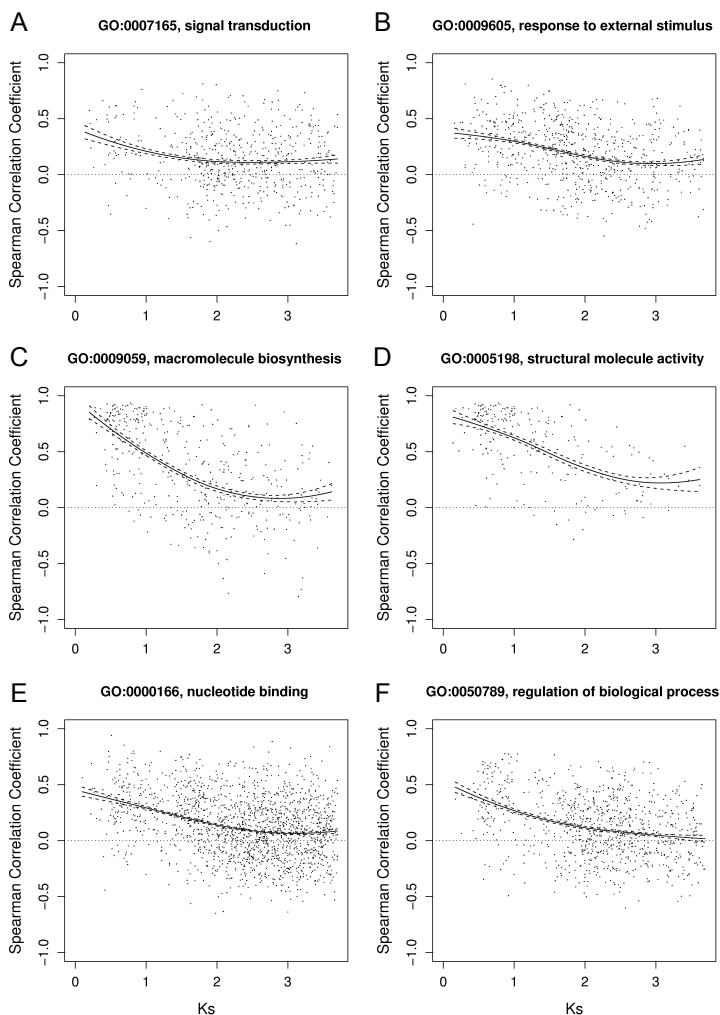


Figure 2.5: Expression divergence in function of time for genes of different functional classes. Scatter plots of the correlation coefficient in function of the K_S value of the gene pairs belonging to different functional classes. The full black line represents the local regression (lofit) line fitted to the data of that particular class, together with its 95% confidence interval (dashed line). (A-B) Gene pairs that have diverged quickly after birth have an intercept of the regression line with the y-axis close to zero; (C-D) whereas slow divergence is reflected by an intercept with the y-axis close to one and a steep slope. (E-F) A more average situation can be observed for most classes. Data of the following classes are displayed: (A) signal transduction; (B) response to external stimuli; (C) macromolecule biosynthesis; (D) structural molecule activity; (E) nucleotide binding; (F) regulation of biological process. Plots of other functional classes of genes can be found in Additional data file 3.

biological process in which it is involved. The rate of expression divergence ranges from very slow, for highly conserved proteins, such as ribosomal proteins, or genes involved in conserved processes, such as biosynthesis pathways or photosynthesis, to very quickly, for instance genes involved in adaptation to and reaction against changing environments.

Note that, because we removed expression data of genes without a unique probeset (see Methods), there are actually more young duplicates than the ones that were plotted in Figure 2.5. Although the current microarray technology does not allow measuring their expression, we can assume that their presence would increase the overall correlation, especially in the low value range of K_S . As the difficulty to design a gene-specific probeset is not related to the functional class, we assume that all functional classes suffer from this caveat to the same extent and that the differences we observe are reliable.

2.3 Conclusions

Investigating gene and genome duplication events as well as the subsequent functional divergence of genes is of fundamental importance in the understanding of evolution and adaptation of organisms. Previously, large-scale gene duplication events have been shown to be prominent in different plant species. Only recently, a pattern of gene retention after duplication has emerged that is biased towards function, time and mode of duplication [83, 130, 150]. For instance, genes involved in signal transduction and transcriptional regulation were shown to have been preferentially retained after large-scale duplication events, while genes of other important functional categories (such as DNA metabolism and cell cycle) were lost [83, 130, 150]. Still other categories of genes, such as those involved in secondary metabolism, are highly retained after small-scale gene duplication [83]. Here, we have studied the expression divergence of these retained duplicates by means of the genome-wide microarray expression data available for *Arabidopsis* genes. As clearly shown in the current study, there is not only a bias in the retention of genes after duplication events, but also in the rate of divergence of expression for different functional categories of genes. Surprisingly, this bias is much more outspoken for genes created by small-scale duplication events than for genes that have been created through large-scale segmental or entire genome duplication events. The latter genes, provided they are still found in duplicated segments, show much higher expression correlation and highly overlapping expression patterns compared to those duplicates that are created by small-scale duplication events or that no longer lie in duplicated segments.

2.4 Methods

2.4.1 Duplicated genes

To identify duplicated genes, an all-against-all protein sequence similarity search was performed using BLASTP (with an E value cut-off of e^{-10}) [164], followed by the application of a criterion based on length and sequence similarity, according to Li et al. [21]. To determine the time since duplication, the fraction of synonymous substitutions per synonymous site (K_S) was estimated. These substitutions do not result in amino acid replacements and are, in general, not under selection. Consequently, the rate of fixation of these substitutions is expected to be relatively constant in different protein coding genes and, therefore, to reflect the overall mutation rate. First, all pairwise alignments of the paralogous nucleotide sequences belonging to a gene family were made by using CLUSTALW [165], with the corresponding protein sequences as alignment guides. Gaps and adjacent divergent positions in the alignments were subsequently removed. K_S estimates were then obtained with the CODEML program [166] of the PAML package [167]. Codon frequencies were calculated from the average nucleotide frequencies at the three codon positions (F3 x 4), whereas a constant K_N/K_S (nonsynonymous substitutions per nonsynonymous site over synonymous substitutions per synonymous site, reflecting selection pressure) was assumed (codon model 0) for every pairwise comparison. Calculations were repeated five times to avoid incorrect K_S estimations because of suboptimal local maxima.

To compare expression patterns of duplicated genes that had arisen through genome duplication events with those created in small-scale duplication events, the complete set of duplicated genes was subdivided into six different subgroups (Figure 2.1), namely:

1. Set 1 containing all genes that are assumed to have been duplicated at a time coinciding with the most recent (3R) polyploidy event.
2. Set 2 containing all genes that are assumed to have been duplicated at a time coinciding with the two (1R/2R) older polyploidy events.
3. Set 3 is a subset of Set 1 and only contains the anchor points (pairs of duplicated genes that still lie on so-called paralogs [101], homologous duplicated segments that still show conserved gene order and content). These genes are thus assumed to have been created by 3R.
4. Set 4 containing the non-anchor point duplicates of Set 1.

5. Set 5 containing the anchor points of Set 2 assumed to have been created by 1R/2R.
6. Set 6 containing the non-anchor points of Set 2.

Previously, through modelling the age distribution of duplicated genes, we estimated that genes created during the youngest genome duplication have a K_S between 0.4 and 1.0, while genes that originated during the oldest two genome duplications were estimated to have a K_S between 1.5 and 3.7 [83]. The latter genes were grouped because it was difficult to unambiguously attribute them to 1R or 2R [83, 100]. The duplicated gene pairs that arose through genome duplication events (anchor points) had been identified previously (complete list available upon request) [101].

2.4.2 Gene Ontology functional classes

Duplicated genes were assigned to functional categories according to the Gene Ontology (GO) annotation. The GO annotation for *Arabidopsis thaliana* was downloaded from TAIR (version 24 June 2005) [168]. We studied genes belonging to the Biological Process (BP) and the Molecular Function (MF) classes of the GO tree. Rather than considering all categories from different levels in the gene ontology, we used the plant specific GO Slim process and function ontologies [169]. In these GO Slim ontologies, categories close to the leaves of the GO hierarchy are mapped onto the more general, parental categories. A gene pair is included in a functional class only when both genes of the pair have been assigned to that particular functional class. Functional classes containing fewer than 200 pairs of duplicated genes were excluded from the analysis.

2.4.3 Microarray expression data

This study was based on gene expression data generated with Affymetrix ATH1 microarrays (Affymetrix, San Diego, CA, USA) [170] during various experiments, all of which are publicly available from the Nottingham Arabidopsis Stock Centre (NASC) [171, 172]. Two datasets were examined that both comprise microarrays that were replicated at least once. The first set includes 153 microarrays that were generated under a broad range of experimental conditions, including, for example, diverse knockout mutants and chemical and biological perturbations (Additional data file 1). Raw data were subjected to robust multi-array average (RMA) normalisation, which is available through Bioconductor [173, 174]. The probe set data of all arrays were simultaneously normalised using quantile normalisation, which eliminates systematic differences between different chips [115, 122, 123]. The log-transformed values were used instead of the raw intensities because of the

variance-stabilising effect of this transformation. Because of the high sequence similarity of recently duplicated genes and the risk of artificially increased correlation due to cross-hybridisation, we selected expression data only from those genes for which a unique probe set is available on the ATH1 microarray (probe sets that are designated with an `'_at'` extension, without suffix). Next, the genes were non-specifically filtered based on expression variability by arbitrarily selecting the 10,000 genes with the highest interquartile range. This was done in an attempt to filter out those genes that show very little variability in gene expression, thereby artificially increasing the overall expression correlation. The mean intensity value was calculated for the replicated slides, resulting in 66 data points for every gene. Next, for each of the 16 different experimental conditions, a treated plant and its corresponding wild-type plant (control experiment without treatment, knock-out or perturbation) were identified (Additional data file 1). To adjust the data for effects that arise from variation in technology rather than from biological differences between the plants, for every gene the intensity value of the wild-type was subtracted from that of the treated plant. The final dataset contained 49 expression measures per gene. For each of the six subsets of duplicates described above 1,279, 8,510, 550, 708, 109, and 8,389 gene pairs, respectively, remained after filtering the microarray data.

The second dataset contains the expression data of genes in 63 plant tissues that were generated within the framework of the AtGenExpress project (Additional data file 2) [151]. The `'mas5calls'` function in Bioconductor was used to study tissue-specific gene expression [173, 174]. This software evaluates the abundance of each transcript and generates a *'detection p value'*, which is used to determine the detection call, indicating whether a transcript is reliably detected (present) or not (absent or marginal). The parameters used correspond to the standard Affymetrix defaults in which a gene with a p value of less than 0.04 is marked as *'present'* [175, 176]. We again selected only expression data from those genes for which a unique probe set is available on the ATH1 microarray. The dataset contains triplicated microarrays and we assigned a gene to be present if it was assigned with a present call in at least one of the three samples. In all other cases an absent call was assigned. We plotted both the absolute (or expression breadth) and relative (or expression divergence of two duplicates) number of tissues in which the genes of a duplicated gene pair are expressed. The latter is defined as the number of tissues in which a gene has a present call divided by the total number of present calls of the duplicated gene pair. Pairs of genes without any present calls were removed from the dataset, resulting in 6,193, 37,838, 1,387, 4,736, 269, 37,438 genes, respectively, for each of the six subsets described above. Both of the above described datasets are available upon request.

2.4.4 Correlation analysis

To measure the expression divergence of two duplicated genes, the Spearman Rank correlation coefficient ρ was calculated. We chose to use this non-parametric statistic because our dataset is a compilation of data from uncorrelated experiments, and might therefore contain outliers. The formula used was:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

where D is the difference between the ranks of the corresponding expression values of both duplicated genes and N is the number of samples. In evaluating and comparing the distributions of the correlation coefficients of the expression of a set of genes, we used the Mann-Whitney U test (two sided, not paired) that is incorporated in the statistical package R [152].

2.4.5 Regression analysis

The relation between expression correlation, measured as the Spearman correlation coefficient, and time, measured as the number of synonymous substitutions per synonymous site K_S , was studied using 'locfit', an R package to fit curves and surfaces to data, using local regression and likelihood methods [152, 177]. We hereby included all duplicated genes with a K_S value smaller than or equal to 3.7 (see above). A local regression model was fitted to the data of each of the functional classes of genes and we looked for biases in expression divergence between the different functional classes by interchanging the fitted models. The model fitted to the data of a particular class was fitted to the data of another class and the quality of the fit was evaluated by assessing the relation between the residuals and fitted values. Residuals that show a clear trend (which is reflected in a non-random distribution around $Y = 0$ with zero mean) indicate that the fitted regression model is inappropriate (that is, the model fitted to the data of the former class is not applicable to the data of the latter).

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a description of dataset 1. Additional data file 2 is a description of dataset 2. Additional data file 3 presents scatterplots of genes belonging to different functional classes. Supplemental material is also available online at [178].

Acknowledgements

This work was supported by a grant from the European Community (FOOD-CT-2004-506223-GRAINLEGUMES) and from the Fund for Scientific Research, Flanders (3G031805). SDB is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. SM is a Research Fellow of the Fund for Scientific Research, Flanders. We would like to thank Todd Vision and Wolfgang Huber for fruitful discussions.

Authors' contributions

TC designed the study, analysed data, and wrote the paper. SDB analyzed data. JR designed the study. SM analyzed data. YVdP designed the study, supervised the project, and wrote the paper.

3

Identification of Novel Regulatory Modules in Dicotyledonous Plants using Expression Data and Comparative Genomics

Klaas Vandepoele, Tineke Casneuf¹ and Yves Van de Peer
Genome Biology (2006) 7, R103

Birds of a feather, flock together

¹Contribution of Tineke Casneuf to this publication: design of the methodology with respect to expression data, assembly and quality-control of the microarray dataset from public resources, followed by their pre-processing and calculation of detection calls. Help on graphical presentation of the results.

Abstract

Transcriptional regulation plays an important role in the control of many biological processes. Transcription factor binding sites (TFBSs) are the functional elements that determine transcriptional activity and are organised into separable *cis*-regulatory modules, each defining the cooperation of several transcription factors required for a specific spatio-temporal expression pattern. Consequently, the discovery of novel TFBSs in promoter sequences is an important step to improve our understanding of gene regulation. Here, we applied a detection strategy that combines features of classic motif overrepresentation approaches in co-regulated genes with general comparative footprinting principles for the identification of biologically relevant regulatory elements and modules in *Arabidopsis thaliana*, a model system for plant biology. In total, we identified 80 TFBSs and 139 regulatory modules, most of which are novel, and primarily consist of two or three regulatory elements that could be linked to different important biological processes, such as protein biosynthesis, cell cycle control, photosynthesis and embryonic development. Moreover, studying the physical properties of some specific regulatory modules revealed that *Arabidopsis* promoters have a compact nature, with cooperative TFBSs located in close proximity of each other. These results create a starting point to unravel regulatory networks in plants and to study the regulation of biological processes from a systems biology point of view.

3.1 Background

Regulation of gene expression plays an important role in a variety of biological processes such as development and responses to environmental stimuli. In plants, transcriptional regulation is mediated by a large number ($> 1,500$) of transcription factors (TFs) controlling the expression of tens or hundreds of target genes in various, sometimes intertwined, signal transduction cascades [179, 180]. Transcription factor binding sites (TFBSs; or DNA sequence motifs, or motifs for short) are the functional elements that determine the timing and location of transcriptional activity. In plants and other higher eukaryotes, these elements are primarily located in the long non-coding sequences upstream of a gene, although functional elements in introns and untranslated regions have been described as well [181, 182]. Moreover, regulatory motifs organise into separable *cis*-regulatory modules (CRMs; modules for short), each defining the cooperation of several TFs required for a specific spatio-temporal expression pattern (for a review, see [183]). As a consequence of this complex organisation, understanding the combinatorial nature of transcriptional regulation at a genomic scale is a major challenge, as the number of possible combinations between TFs and targets is enormous. On top of this, it is important to realise that not all motifs present in a promoter

are functional elements or simultaneously active, since the cooperation between TFs is context dependent [184]. In the absence of already characterised TFBSs or systematic genome-wide location (that is, chromatin immunoprecipitation-chip) data revealing interactions between TFs and target genes, sequence and expression data are the only sources of information that can be combined to identify CRMs [185–187].

The discovery of regulatory motifs and their organisation in promoter sequences is an important first step to improve our understanding of gene expression and regulation. Since co-expressed genes are likely to be regulated by the same TF, the identification of shared and thus overrepresented motifs in sets of potentially co-regulated genes provides a practical solution to discover new TFBSs. Complementarily, the identification of significantly conserved short sequences (or footprints) in the promoters of orthologous genes in related species points to candidate regulatory motifs for a particular gene [188]. In yeasts and animals both overrepresentation of motifs in co-regulated genes and comparison of orthologous sequences have been successfully applied to delineate regulatory elements (for an overview, see [189, 190]); in plants, however, mainly analyses on co-regulated genes for particular biological processes (for example, stress, hormone and lightresponse, cell cycle control) have been reported [180].

Two problems interfering with comparative approaches for the detection of regulatory motifs in orthologous plant sequences are the limited amount of genomic sequence information for related species (but see [191]) and the high frequency of both small- and large-scale duplication events that hamper the delineation of correct orthologous relationships [128, 192]. Finally, the correct identification of functional TFBS is more complex in higher eukaryotes compared to prokaryotes or yeast because of the longer intergenic sequences. Consequently, characterising properties of regulatory elements and modules is not trivial due to the inclusion of large amounts of false positives in sets of putative target genes. To overcome these problems, several approaches integrate local sequence conservation between orthologous upstream regions to exclude non-conserved regions from the search space and to make more accurate predictions about the presence of regulatory signals [193–198]. Nevertheless, this methodology requires that genomic data from closely related species are available and that correct (one-to-one) orthologous relationships can be identified for nearly all genes.

Here, we present a detection strategy that integrates features of classic approaches looking for overrepresented motifs with general comparative footprinting principles for the systematic characterisation of biologically relevant TFBSs and CRMs in *Arabidopsis thaliana*, a dicotyledonous plant model system. In a first

stage, a classic Gibbs-sampling approach is used to identify TFBSs in sets of co-expressed genes. Next, these TFBSs are presented to an evolutionary filter to select functional regulatory elements based on the global conservation of TFBSs in target genes in a related species, *Populus trichocarpa* (poplar). In a second stage, a two-way clustering procedure combining the presence/absence of motifs and expression data is used to identify additional new TFBSs. The Gene Ontology (GO) vocabulary combined with the original expression data is used to functionally annotate sets of genes containing a particular regulatory element or module. As a result, 80 TFBSs are reported, of which more than half correspond with previously described plant *cis*-regulatory elements. More interesting, we were able to identify numerous regulatory modules driving different biological processes, such as protein biosynthesis, cell cycle, photosynthesis and embryonic development. Finally, the physical properties of some modules are characterised in more detail.

3.2 Results and Discussion

3.2.1 General overview

The input data for our analysis were genome-wide expression data and the genome sequence from *Arabidopsis*, plus genomic sequence data from a related dicotyledon, poplar [199]. Whereas the expression data are required for creating sets of co-regulated genes that serve as input for the detection of TFBSs using MotifSampler (see Methods), the genomic sequences are used to delineate orthologous gene pairs between *Arabidopsis* and poplar, forming the basis for the evolutionary conservation filter. This filter is used to discriminate between potentially functional and false motifs and is based on the network-level conservation principle, which applies a systems-level constraint to identify functional TFBSs [200, 201]. Briefly, this method exploits the well-established notion that each TF regulates the expression of many genes in the genome, and that the conservation of global gene expression between two related species requires that most of these targets maintain their regulation. In practice, this assumption is tested for each candidate motif by determining its presence in the upstream regions of two related species and by calculating the significance of conservation over orthologous genes (see Methods and Figure 3.1A). Whereas the same principle of evolutionary conservation is also applied in phylogenetic footprinting methods to identify TFBSs, it is important to note that, here, the conservation of several targets in the regulatory network is evaluated simultaneously. This is in contrast with standard footprinting approaches, which only use sequence conservation in upstream regions on a gene-by-gene basis to detect functional DNA motifs.

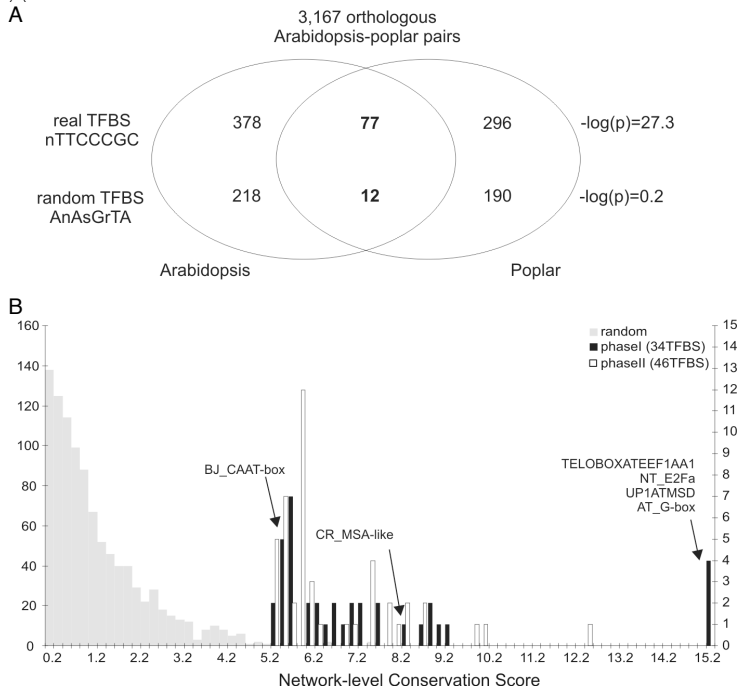


Figure 3.1: Network-level conservation filter. (A) The occurrence of a candidate TFBS in the set of orthologous *Arabidopsis*-poplar gene pairs was determined and the significance of the overlap is measured using the hypergeometric distribution [201]. The NCS is defined as the negative logarithm of the hypergeometric p value. (B) Distribution of NCS values for 1,000 randomly generated TFBSs (grey) and the motifs found using the co-expression (black) and the two-way clustering (white) procedure. The left and right y-axis show the frequency for the random and the potentially functional TFBSs, respectively.

After applying motif detection on a set of co-expressed *Arabidopsis* genes in a first stage, all TFBSs retained by the network-level conservation filter are subsequently combined with the original expression data to identify CRMs and additional regulatory elements ('two-way clustering'; Figure 3.2). Both objectives were combined because it has been demonstrated that the task of module discovery and motif estimation is tightly coupled [202]. We reasoned that, for a group of genes with similar motif content but with dissimilar expression profiles, additional TFBSs may exist that explain the apparent discrepancy between motif content and expression profile.

Whereas the procedure for detecting TFBS in co-expressed genes combined

with the evolutionary filter is highly similar to the methodology described by Pritsker and co-workers [200], the second stage of TFBS detection using the two-way clustering procedure is, to our knowledge, novel. The inference of regulatory modules is related to the work of Kreiman [195], although, in the current study, no *a priori* physical constraints were used to exhaustively search for CRMs.

3.2.2 Identification of individual TFBSs using co-expressed genes

Applying the Cluster Affinity Search Technique (CAST) algorithm to the data set measuring the expression of 19,173 *Arabidopsis* genes over 489 different experiments (1,168 Affymetrix ATH1 slides; see Additional data file 5) yielded 122 clusters of co-regulated genes covering 5,664 genes (see Methods). After running MotifSampler, applying the network-level conservation filter and removing redundant motifs (see Methods), 34 motifs with a significant (p value < 0.01) Network-level Conservation score (NCS) were retained (Figure 3.1B). Interestingly, 25 of the identified TFBSs can be functionally annotated based on overrepresented GO Biological Process or Molecular Function terms in the set of putative target genes (Table 3.1). Overall, nearly 60% (20/34) of all motifs correspond with known plant regulatory elements. Throughout this paper, for motifs corresponding with known regulatory elements described in PLACE [203] and PlantCARE [204] the original name is used, whereas for new elements the consensus motif will be used.

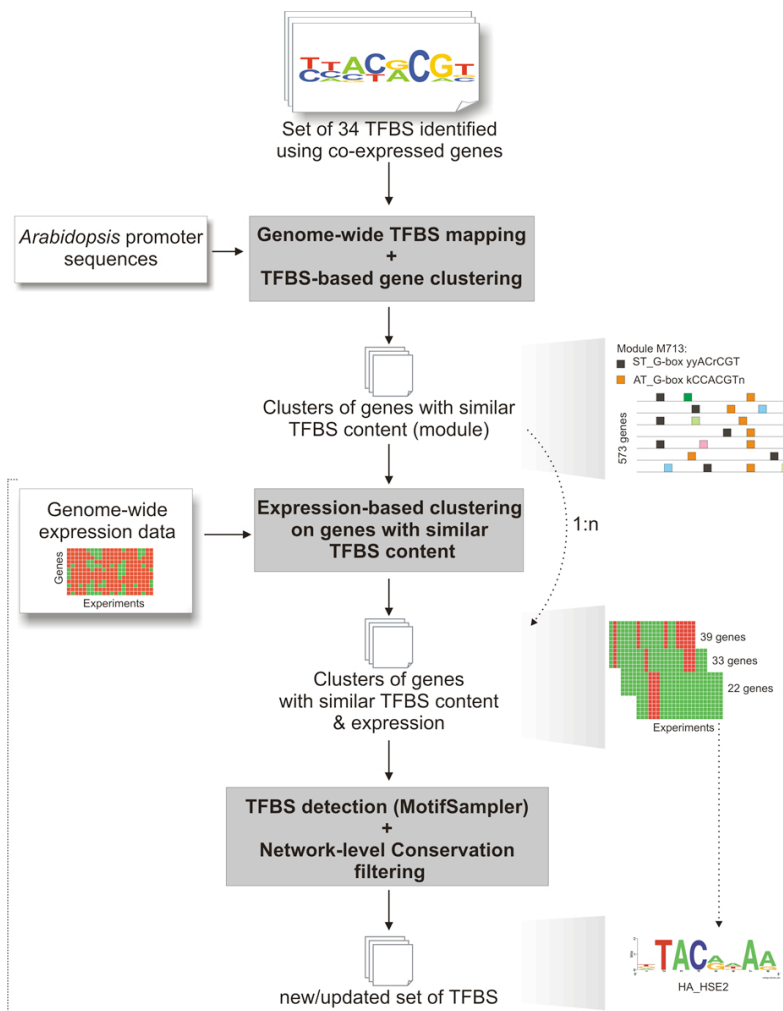


Figure 3.2: Detection of TFBSs using two-way clustering. Starting from the available set of 34 TFBSs identified using sets of co-expressed genes (see text for details), clusters of genes with similar TFBS combinations in their promoter are delineated. Next, within each set of genes with similar TFBS content, groups of co-expressed genes are identified. Finally, motif detection is applied and evolutionarily conserved TFBSs are retained. The panel on the right shows the identification of the TFBS HA_HSE2 involved in zygotic embryogenesis. The top picture depicts a subset of all 573 Arabidopsis genes containing the module consisting of two distinct G-boxes. The two images below show the three groups of co-expressed genes and the newly identified TFBSs found in a set of 22 genes containing both G-boxes in their promoter and showing embryo-specific expression. Note that the section indicated with the dotted line corresponds with the motif-detection approach applied on co-expressed genes in the first stage.

Table 3.1: Overview of the TFBSs identified using co-expressed genes

TFBS motif*	NCS†	Known motif	Site‡	Functional enrichment targets: GO Biological Process or Molecular Function§
nrCAAnTC (a)	5.77	BJ_CAAAT-box	TGCAAAATCT	GO:0008152 metabolism $8.58E^{-04}$ (1.2) GO:0003824 catalytic activity $8.91E^{-05}$ (1.2)
GTACAwry (b)	5.64			GO:0007275 development $2.89E^{-02}$ (1.6) GO:0003824 catalytic activity $2.98E^{-03}$ (1.2)
TTCkwTs	5.79	BOXIINTPATPB	ATAGAA	
sGCtGAGA	5.77			GO:0015980 energy derivation by oxidation of organic compounds $4.82E^{-02}$ (2.7) GO:0008152 metabolism $1.43E^{-03}$ (1.2) GO:0003824 catalytic activity $2.89E^{-03}$ (1.1)
kCCACGTn (4)	17.54	AT_G-box HV_ABRE6 PH_boxII	GCCACGTGGA GCCACGTACA TCCACGTGGC	GO:0015979 photosynthesis $2.48E^{-04}$ (4.2) GO:0048316 seed development $2.64E^{-03}$ (3.6) GO:0009793 embryonic development (sensu Magnoliophyta) $6.15E^{-03}$ (3.5)
yCAITTrT (c)	8.7	GM_Unnamed.6	GCATTTTATCA	GO:0003700 transcription factor activity $2.94E^{-03}$ (1.3) GO:0030528 transcription regulator activity $1.64E^{-02}$ (1.3) GO:0003677 DNA binding $3.86E^{-02}$ (1.2)
ynTTATCC	6.75	SREATMSD AT_I-box	TTATCC CCTATCCT	
nGTTGACw (d)	5.31	ZM_O2-site	GTTGACGTGA	GO:0006952 defense response $2.99E^{-04}$ (1.9) GO:0009607 response to biotic stimulus $3.56E^{-04}$ (1.7) GO:0016301 kinase activity $7.52E^{-11}$ (1.7)
TTTGChrA	6.13			GO:0016773 phosphotransferase activity, alcohol group as acceptor $1.14E^{-02}$ (1.6) GO:0016772 transferase activity, transferring phosphorus-containing groups $2.60E^{-02}$ (1.5)
rATyTGGG	5.58			
TtTwTATA	9.35	AT_TATA-box	TATATAA	GO:0019748 secondary metabolism $2.76E^{-02}$ (2.1)

IDENTIFICATION OF NOVEL REGULATORY MODULES IN DICOTYLEDONOUS PLANTS
USING EXPRESSION DATA AND COMPARATIVE GENOMICS

TFBS motif*	NCS†	Known motif	Site‡	Functional enrichment targets: GO Biological Process or Molecular Function§
ATArwACA (e) nTTCCCGC (5)	5.79 27.27	OS_Unnamed_2 NT_E2Fa	CCATGTCATATT TTTCCCGC	GO:0006519 amino acid and derivative metabolism $1.35E^{-02}$ (1.8) GO:0003700 transcription factor activity $3.36E^{-02}$ (1.3) GO:0006261 DNA-dependent DNA replication $6.48E^{-04}$ (6.2) GO:0000067 DNA replication and chromosome cycle $1.06E^{-07}$ (5.5) GO:0006260 DNA replication $3.57E^{-05}$ (5.1)
TKAGAwA	8.86	BO_TCA-element3	TCAGAAAGAGG	GO:0006464 protein modification $4.52E^{-02}$ (1.7) GO:0003824 catalytic activity $5.20E^{-03}$ (1.1)
AAACCCTA (13)(f)	40.06	TELOBOXATEEF1AA1	AAAGCCCTAA	GO:0042254 Ribosome biogenesis and assembly $9.86E^{-13}$ (4.4) GO:0007046 ribosome biogenesis $5.67E^{-12}$ (4.3) GO:0008248 premRNA splicing factor activity $3.20E^{-04}$ (3.9) GO:0003824 catalytic activity $2.93E^{-02}$ (1.1)
mGnyAAAG (g) GAnChkmG	6.38 6.29			GO:0003729 mRNA binding $1.00E^{-02}$ (3.1) GO:0003735 structural constituent of ribosome $3.69E^{-02}$ (1.7) GO:0006412 protein biosynthesis $3.15E^{-03}$ (1.7)
TCnCTCTC	8.98	LE_5UTRPy-richstretch	TTTCTCTCTCTCTC	GO:0003777 microtubule motor activity $9.90E^{-03}$ (2.7) GO:0050789 regulation of biological process $2.27E^{-03}$ (1.4) GO:0016772 transferase activity, transferring phosphoruscontaining groups $7.89E^{-03}$ (1.4) GO:0003824 catalytic activity $4.51E^{-03}$ (1.1)
wmGTCmAim ynCAACGG	7.16 8.39	CR_MSA-like	YCYAACGGYYA	GO:0003774 microtubule motor activity $3.17E^{-03}$ (3.4) GO:0003774 motor activity $8.55E^{-03}$ (2.9)
nmGAtyCr	5.66			GO:0006944 membrane fusion $2.32E^{-02}$ (4.5) GO:0003735 structural constituent of ribosome $2.77E^{-03}$ (1.9) GO:0005198 structural molecule activity $7.11E^{-04}$ (1.9)
CGkCGmCh AGGCCCAw (9)	7.68 21.94	OS_GC-motif5 UP1ATMSD	CGGGGCCCT GGCCCAWWW	GO:0007046 ribosome biogenesis $3.56E^{-14}$ (4.3) GO:0042254 ribosome biogenesis and assembly $2.28E^{-14}$ (4.3) GO:0003735 structural constituent of ribosome $8.66E^{-29}$ (3.3)
AykyATwA	6.09			

TFBS motif*	NCS†	Known motif	Site‡	Functional enrichment targets: GO Biological Process or Molecular Function§
CTGnCTCy	6.91			GO:0016301 kinase activity $3.44E^{-02}$ (1.3) GO:0003676 nucleic acid binding $3.48E^{-02}$ (1.2) GO:0005488 binding $2.60E^{-03}$ (1.2)
TsTCGnTT	7.22			GO:0003824 catalytic activity $5.10E^{-03}$ (1.1)
TmAsTGAn	7.76	OS_GTCAdirectrepeat	TAAAGTCATAA CTGA TGA	GO:0016491 oxidoreductase activity $3.85E^{-03}$ (1.5) GO:0008152 metabolism $5.74E^{-03}$ (1.2) GO:0003824 catalytic activity $5.70E^{-04}$ (1.2)
yyACrCGT (2)	6.56	ST_G-box	TCACACGTGGC	GO:0009605 response to external stimulus $4.80E^{-02}$ (1.6) GO:0006950 response to stress $3.42E^{-02}$ (1.6)
mATATTTT	5.51	GM_Nodule-siteI	GATATATTAA TATTTT ATTATTATA	
CCAATnCm	5.78	CAATBOX1	CAAT	
rKTCaWgM	5.42	HV_ATC-motif	GCCAAATCC	
ssCGCCnA (2)	9.13	E2F1OSPCNA	GCGGGGAAA	GO:0003824 catalytic activity $6.17E^{-05}$ (1.2) GO:0000067 DNA replication and chromosome cycle $4.74E^{-02}$ (3.0) GO:0006259 DNA metabolism $2.15E^{-03}$ (2.3) GO:0007049 cell cycle $4.29E^{-02}$ (2.2)
TTTATGnG	7.1			
TCaWATAA	6.74			

* Numbers in parentheses indicate the number of clusters (containing co-expressed genes) in which the motif was independently identified. The letters in parentheses refer to the updated TFBS identified using the two-way clustering: (a) GCAAnTCn; (b) GTACmwGy; (c) yCAATTAT; (d) mKTTGACT; (e) ATrrwACA; (f) AAACCCCTA; (g) mGnCAAAAG.

† Network-level Conservation score.

‡ Residues in bold indicate the matching position between the known motif and the motif found in this study. Known motifs were retrieved from PLACE [203] and PlantCARE [204].

§ Only the first three GO categories according to the highest enrichment score are shown. The enrichment score is shown as number in parentheses.

The telo-box (TELOBOXATEEF1AA1) is the TFBS with the highest NCS value (40.06), indicating that this motif is highly conserved in orthologous target genes between *Arabidopsis* and poplar. The GO annotation reveals that this motif is highly enriched in the promoter of genes involved in ribosome biogenesis and assembly (p value $< 10^{-12}$; 4.4-fold enrichment), confirming the role of the telo-box in regulating components of the translational machinery [205]. Other motifs with high NCS values together with their functional annotation correspond to well-described plant TFBSs, such as the E2F box and the MSA element involved in DNA replication and microtubule motor activity during the cell cycle [206], the UP1 box mediating the transcription of protein synthesis [207], and the G box inducing the transcription of photosynthesis genes in response to light [208]. The observation that 71% of these motifs are located within the first 500 base-pairs (bp) upstream of the translation start site (Additional data file 1) for conserved orthologous *Arabidopsis*-poplar targets confirms previous findings that *Arabidopsis* promoters are generally compact [209,210].

3.2.3 Combining motif and expression data to identify additional TFBSs

Although the motif detection approach using co-expressed genes revealed a first set of TFBSs, it is clear that expression data alone are insufficient to unravel the complex nature of transcriptional regulation in higher plants. Therefore, we applied a two-way clustering procedure combining motif and expression data to identify additional regulatory elements. We again used MotifSampler combined with the networklevel conservation filter to identify potential TFBSs in clusters of co-expressed genes, but now also incorporated the prior knowledge about the presence of particular TFBSs in a gene's promoter. Thus, first all genes with a particular motif combination (module) in the *Arabidopsis* genome were identified after which the expression profiles of these genes were used to delineate subgroups of co-expressed genes, which were then again presented to the motif detection routine (MotifSampler and network-level conservation filter; Figure 3.2). The rationale behind this approach is that additional TFBSs may exist that explain the different expression patterns within the set of genes containing the same module. As shown below, these new motifs can be missed in the first detection stage on co-expressed genes since the fraction of genes containing this TFBS within the set of co-expressed genes is too small for reliable detection by MotifSampler. By evaluating all possible combinations (from two up to four motifs) using all 34 initial TFBSs, we found 1,249 modules containing more than 40 genes. Next, we determined groups of co-expressed genes for each set of genes characterised by a specific module using the CAST algorithm (as described before). In total,

695 regulons, containing genes with a particular module and similar expression profiles, were found, covering 4,100 *Arabidopsis* genes. Note that the way of grouping genes with identical modules is compatible with the combinatorial nature of transcriptional control in higher eukaryotes, since the presence of additional TFBSs in a gene's promoter does not interfere with the gene clustering based on TFBS content (for example, gene *i* with motifs A, B and C can theoretically occur in the clusters containing module A-B, A-C, B-C and A-B-C; see Methods).

After running MotifSampler and the network-level conservation filter on all regulons, 46 new TFBSs were found (Additional data file 6). Again, the high fraction (25/46, or 54%) of TFBSs with similarity to previously described ones indicates that we most probably identified an extra set of genuine regulatory elements. As an illustration, we discuss the discovery of the HA_HSE2 motif, which is an element inducing gene expression during zygotic embryogenesis [211]. Initially, 573 *Arabidopsis* genes were grouped containing a combination of two distinct G-boxes in their promoters (AT_G-box kCCACGTn and ST_G-box yyACrCGT; Table 3.1). Subsequent clustering of the expression profiles of these genes, enriched for the GO terms embryonic development (*sensu* Magnoliophyta) and seed development (both with p value $< 10^{-2}$; 7.4fold and 8.1-fold enrichment, respectively), yielded three regulons, of which one showed expression in seeds, a second one expression in leaves and shoots, and a third one expression in the globular and heart stage embryo. Running the motif detection routine on the 22 genes in this last regulon resulted in the discovery of the HA_HSE2 motif (NCS 7.91). This motif was not identified in the first TFBS detection run using expression data only, since the genes in this regulon were part of a big set of 645 co-expressed genes not yielding any significant TFBSs. This finding confirms that splitting up co-expressed genes into smaller subsets based on prior knowledge of motif content can enhance the identification of new TFBSs.

3.2.4 Inferring functional regulatory modules

To get a general overview of the involvement of all 80 TFBSs (34 from co-expressed genes in the first stage plus 46 from two-way clustering in the second stage) and the derived CRMs in different biological processes, we identified all modules with two to four motifs (containing at least 20 *Arabidopsis* genes) and again used overrepresented GO terms for functional annotation. Briefly, we selected all *Arabidopsis* genes with a particular motif combination present in their upstream regions and verified whether any GO Biological Process term was significantly enriched within this set of putative target genes.

IDENTIFICATION OF NOVEL REGULATORY MODULES IN DICOTYLEDONOUS PLANTS USING EXPRESSION DATA AND COMPARATIVE GENOMICS

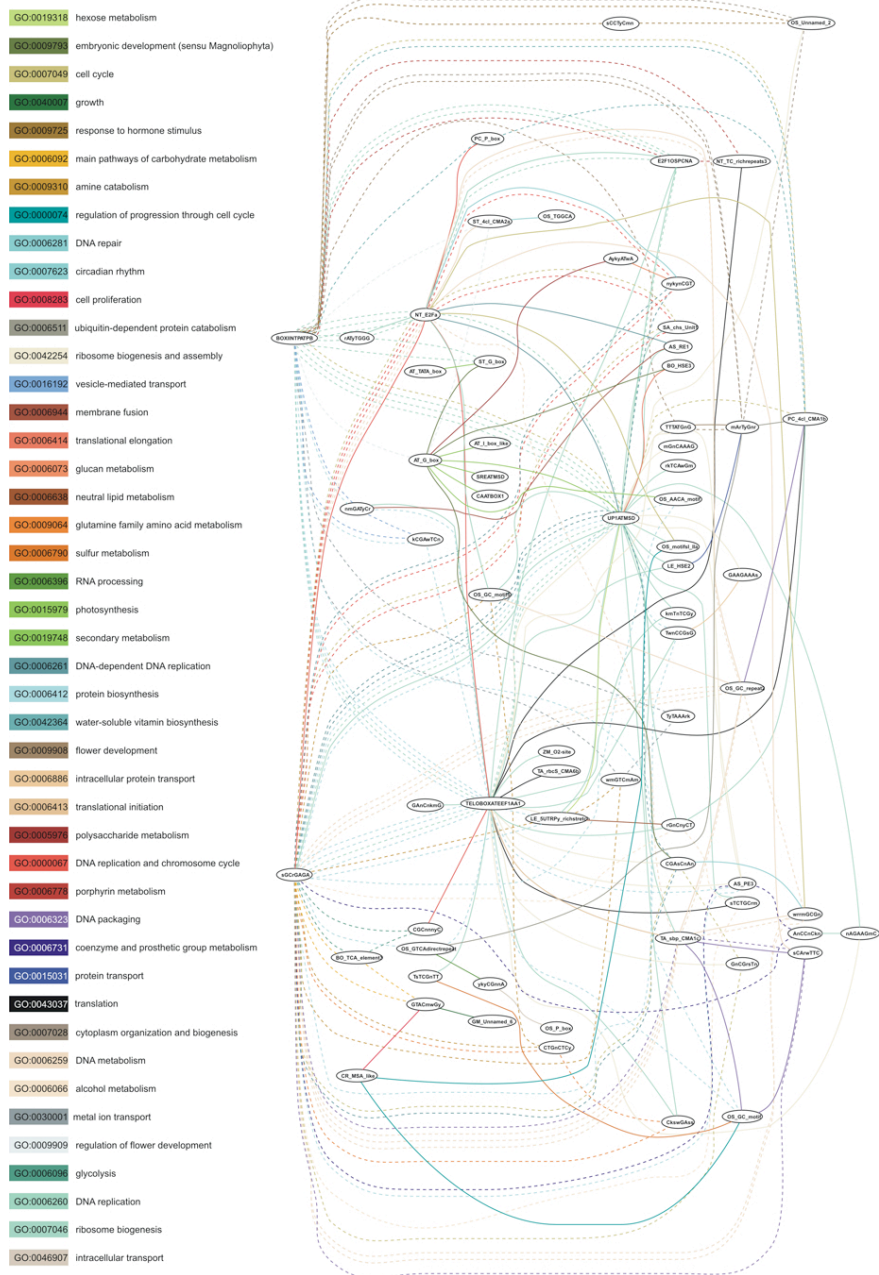


Figure 3.3: Motif synergy map for 139 modules with significant GO Biological Process annotation. The full and dotted lines connect motifs cooperating in modules containing two and three TFBSs, respectively. Line colours indicate the GO Biological Process enrichment for *Arabidopsis* genes containing this module (see also Additional data file 7).

Figure 3.3 shows the motif synergy map depicting the cooperation of different TFBSs for which the GO enrichment score is stronger for the module than for the individual TFBS (within that module). Applying this criterion is necessary to specifically identify the functional properties of the module, because the GO enrichment for many modules is caused by the presence of an individual TFBS and not by the specific TFBS combination in the CRM. In total, 139 modules with significant functional GO Biological Process enrichment were identified, of which 97 consist of a combination of two and 42 of three TFBSs (Additional data file 7). Moreover, 69 identified TFBSs in this study could be allocated to one or more CRM with significant functional annotation. The module with the strongest GO enrichment in the synergy map consists of a telo-box and the UP1 motif and targets protein biosynthesis (p value $< 10^{-51}$) and ribosome biogenesis (p value $< 10^{-25}$) genes (for example, 40S and 60S ribosomal proteins, translation initiator factors). In total, 851 *Arabidopsis* genes contain this module and the expression coherence [187] of these genes ($EC = 0.14$; see Methods) illustrates that this module is responsible for similar expression profiles in a large number of these genes. Detailed information about target genes and functional annotation for the different CRMs can be consulted on our website [178].

Analysing the topology of the motif synergy map reveals some highly connected TFBSs (for example, UP1ATMSD, TELOBXATEEF1AA1, sGCrGAGA, BOXIINTPATPB, AT_G-box kCCACGTn), which control, in cooperation with other TFBSs, different biological processes. A set of modules contain a G-box and confirm its role in controlling light-dependent processes such as photosynthesis (module 2.M6107, AT_G-box kCCACGTn + I-box-like ATAATCCA; module 2.M6144, AT_G-box kCCACGTn + OS_AACA_motif; module 2.M6069, AT_G-box kCCACGTn + SREATMSD) and embryonic development (module 2.M6103, AT_G-box kCCACGTn + CGAsCnAn; module 2.M6125, AT_G-box kCCACGTn + BO.HSE3 box). The cooperation between the G-box and the I-box-like motif in the module with GO enrichment 'photosynthesis' targets genes coding for chlorophyll binding proteins, different photosystem I reaction centre subunits, photosystem II associated proteins, and ferredoxin. The high expression of these genes in plant tissues exposed to light suggests a function for this module as a composite lightresponsive unit [212]. Combining the clusters of co-expressed genes used in the first detection stage with the targets of the different modules



Figure 3.4: Correlation between *cis*-regulatory modules and clusters of co-expressed genes. Rows depict co-expression clusters with their corresponding cluster number and brief description, if available, whereas columns show modules with their corresponding GO descriptions. The number of genes within each co-expression cluster is indicated in parentheses. Only expression clusters enriched for one (or more) modules are shown. Enrichment was calculated using the hypergeometric distribution and *p* values were corrected for multiple hypotheses testing with the false discovery rate method (*q* value) [213].

(Figure 3.4) shows a highly significant overlap of expression cluster 3 with the photosynthesis modules 2.M6069, 2.M6144, 2.M6107 and 2.M6081 (AT_G-box kCCACGTn + UP1 box). These strong associations indicate that these motif combinations are involved in (light-regulated) primary energy production.

Three modules (2.M6086, 2.M6103 and 2.M6125) targeting genes involved in embryonic development (>7-fold GO enrichment; Additional data file 7) are strongly associated with expression cluster 9, which shows high transcriptional activity in seedlings and embryo (Figure 3.4). The presence of these modules, all containing a G-box, in some well-described embryogenesis genes within this expression cluster (for example, late embryogenesis-abundant proteins, zinc-finger protein PEI1 and NAM transcriptional regulators [214,215]) confirms our finding that these modules play an important role in transcriptional control during embryo development.

The motif sGCrGAGA is involved in 26 different modules and is, to our knowledge, a new TFBS. Whereas the full set of *Arabidopsis* genes containing this motif shows a functional enrichment for 'energy derivation by oxidation of organic compounds' (Table 3.1), more than a quarter of all modules (7/ 26) containing this regulatory element seem to have a role in transcriptional control of sugar, amino acid or alcohol metabolism. Examples of biosynthesis pathways mediated by these

modules according to the GO Biological Process annotation include glycolysis, amine catabolism and branched chain family amino acid metabolism (Additional data file 7).

Another module (2.M6825) controls the progression through the cell cycle and consists of a combination of the known MSA element together with the OS_GC motif. A large number of genes associated with mitosis and cytokinesis, such as those encoding B-type cyclins, kinesin motor proteins and microtubule and phragmoplast-associated proteins, contain this CRM and are linked with expression cluster 62 (Figure 3.4). Comparing the occurrence of this module in a set of approximately 1,000 periodically expressed genes determined in *Arabidopsis* cell suspensions by Menges and co-workers [216] confirms a strong enrichment towards M-phase specific genes (hypergeometric probability distribution; p value $< 10^{-21}$). Nevertheless, because the frequency of the individual MSA element is higher in the set of M-phase specific genes compared to the occurrence of the module (87/198 MSA element and 40/198 module, respectively), this indicates that the presence of the individual MSA box is sufficient for M-phase expression during cell division and that additional cooperative elements only moderately mediate the level of transcription, as recently shown [217]. Likewise, despite the fact that several modules (for example, 2.M547, 2.M6460 and 2.M6451) consisting of the NT_E2Fa motif and one or more cooperative TFBS are targeting genes involved in DNA replication (>10 -fold enrichment) and are strongly associated with expression cluster 44 (Figure 3.4) containing many DNA replication genes (for example, DNA replication licensing factor, PCNA1-2), it is currently unclear whether additional motifs, apart from one or more E2F elements, are essential for transcriptional induction during S-phase in plants [210].

Another module driving endogenous light-regulated response contains the ST_4cl-CMA2a and OS_TGGCA boxes and targets genes involved in circadian rhythm (2.M8255, 'circadian rhythm' >24 -fold enrichment). Examples of genes containing this module are CONSTANS, a zinc finger protein linking day length and flowering [218], as well as APRR5 and APRR7, pseudo-response regulators subjected to a circadian rhythm at the transcriptional level [219]. One of the TFBSs within this module, motif OS_TGGCA with sequence [GT]C [AT]A [AG]TGG, is highly similar to the SORLIP3 motif (CTCAAGTGA; Pearson correlation coefficient (PCC) = 0.56 between linearised PWM and SORLIP3), a sequence found to be overrepresented in light-induced promoters [220].

3.2.5 Properties of *cis*-regulatory modules

Due to the frequent nature of large-scale duplication events in plants, a one-to-one orthologous relationship with poplar could be ensured for only a minority of *Arabidopsis* genes (17%). Therefore, applying across-species conservation on a genome-wide scale to predict functional TFBSs, as done in mammals and yeast, is not straightforward in plants. Similarly, studying cooperative TFBSs within regulatory modules also suffers from the inclusion of potentially false-positives when selecting genes in one species containing a putative module. Therefore, we exploited the conservation of TFBSs between *Arabidopsis* and poplar orthologs to study the properties of some modules in more detail. Based on all 139 modules and the set of 3,167 (one-to-one) orthologous genes between *Arabidopsis* and poplar, we only retained 30 modules with five or more conserved target genes for further analysis. By applying this stringent filtering step of five or more conserved orthologous targets, we wanted to study the physical properties - motif order and spacing - of CRM in a set of *Arabidopsis* target genes enriched for functional TFBSs (and with a minimum number of false-positives; data not shown). Since no *a priori* information about such properties was included in the identification of TFBSs and CRMs, we used this data set to verify whether such constraints exist and are used by the transcriptional apparatus to control gene expression in plants.

First, for each module the overrepresented motif order was quantified in all conserved target genes (for example, 9/11 of all conserved *Arabidopsis* target genes for module 2.M7010 contain pattern [TELOBOXATEEF1AA1 *spacer* UP1ATMSD *spacer* start codon]). Grouping all these results indicates that, on average, 68% (136/200) of all *Arabidopsis* targets contain an overrepresented motif order (Additional data file 8). Nevertheless, the observation that, on average, approximately 64% of the orthologous poplar targets contain the same motif order suggests that, although a preferred motif order might be present for some modules (Additional data file 2), this configuration is evolutionarily rather weakly conserved. Measuring the distance between cooperative TFBSs reveals that, for 11/30 modules, the average distance is significantly smaller than expected by chance (Additional data file 8). Moreover, the overall distribution of distances between TFBSs measured for all 200 targets within these 30 modules is, in both *Arabidopsis* and poplar, significantly different from a random distribution (Mann-Whitney U test p value < 0.001; Figure 3.5). This indicates that, like in other eukaryotic species (for example, see [195, 221, 222]), the distance between cooperative motifs within a module is important for functionality.

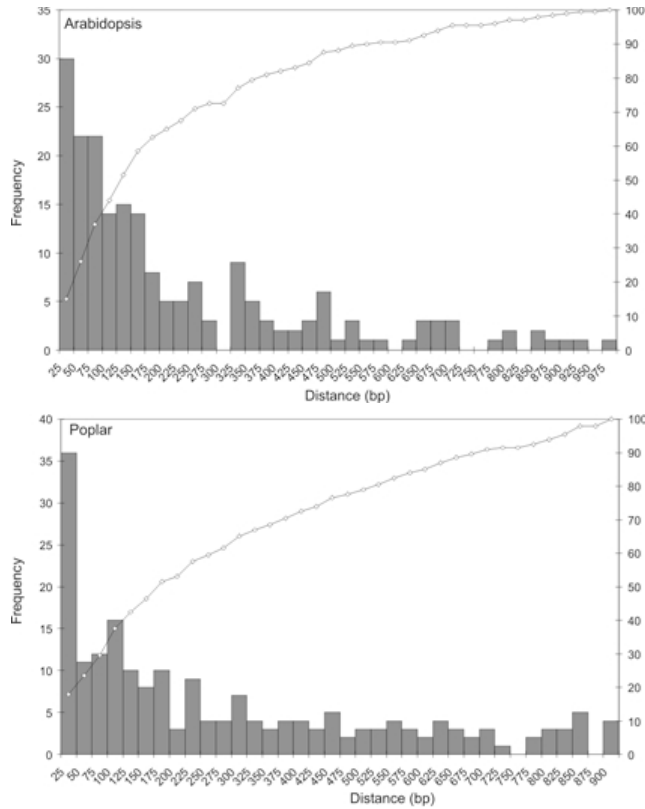


Figure 3.5: Motif distance distributions for 30 conserved modules in orthologous target genes between *Arabidopsis* and poplar. Motif distance distributions for 30 conserved modules in orthologous target genes between *Arabidopsis* and poplar. For all modules, the distance (in bp) between cooperative TFBS was measured in 200 conserved orthologous target genes and plotted in a histogram for *Arabidopsis* and poplar. The white boxes denote the cumulative fraction.

3.3 Conclusions

The results of this study confirm that TFBS detection using expression data within an evolutionary context offers a powerful approach to study transcriptional control [195, 197, 200]. Especially, the exploitation of sequence conservation between related species offers a good control against false-positives when performing motif detection on co-regulated genes [223–226]. Using clusters of co-expressed genes, MotifSampler, two-way clustering and the network-level conservation principle, 80 distinct TFBSs could be identified, of which 45 correspond to known plant *cis*-regulatory elements. From these, 139 regulatory

modules with biological functional annotation could be inferred and several CRMs were highly associated with distinct expression patterns. Despite the limited amount of comparative sequence data for dicotyledonous plants, which hinders the systematic identification of conserved and probably functional binding sites within a promoter, the regulatory modules identified here suggest that, like in yeast and animals, combinatorial transcriptional control plays an important role in regulating transcriptional activity in plants. For sure, the application of more advanced CRM detection methods (for example, [202, 227, 228]) integrating physical constraints acting on CRMs (as shown here) on more detailed expression data will lead to the discovery of additional plant CRMs. Finally, the sequencing of additional and less diverged plant species in the near future [229] should provide a more solid comparative framework to study the organisation and evolution of transcriptional regulation within the green plant lineage.

3.4 Methods

3.4.1 Expression data

A total of 1,168 Affymetrix ATH1 microarrays monitoring the transcriptional activity of more than 22,000 *Arabidopsis* genes in different tissues and under different experimental conditions were retrieved from the Nottingham Arabidopsis Stock Centre (NASC [171]; 1,151 slides) and The Arabidopsis Information Resource (TAIR [168]; 17 slides). An overview of all data sets is shown in Additional data file 5. Raw data were normalised using the MicroArray Suite 5.0 (MAS) implementation in Bioconductor ('mas5' function) [173]. To remove potentially cross-hybridising probes, only genes for which a unique probe set is available on the ATH1 microarray (probe sets with a '_at' extension without suffix) were retained. Next, the genes were filtered based on the detection call that is assigned to each gene by the 'mas5calls' function implemented in Bioconductor. This software evaluates the abundance of each transcript and generates a detection p value indicating whether a transcript is reliably detected (p value < 0.04 for present value). Only genes that were called present in at least 2% of the experiments were retained for further analysis. Finally, the mean intensity value was calculated for the replicated slides, resulting in 489 measurements for 19,173 genes in total.

3.4.2 Clustering of expression data

To group genes with similar expression profiles, we used the CAST algorithm with the PCC as affinity measure [230]. Advantages of CAST clustering over more classic algorithms such as hierarchical or K-means clustering are that only two

parameters have to be specified (the affinity measure, here defined as $PCC \geq 0.8$, and the minimal number of genes within a cluster, here set to 10) and that it independently determines the total number of clusters and whether a gene belongs to a cluster. We used an additional heuristic to choose the gene with the maximum number of neighbours (that is, the total number of genes having a similar expression profile) to initiate a new cluster. An overview of the cluster stability when randomly removing experiments from the complete expression data set is given in Additional data file 3.

3.4.3 Detection of transcription factor binding sites

For each cluster S , grouping n_S co-regulated genes returned by the CAST algorithm, we used MotifSampler [231] to identify an initial set of TFBSs. We restricted the search to the first 1,000 bp upstream of the translation start site. For some genes the upstream sequence was shorter because the adjacent upstream gene is located within a distance smaller than 1,000 bp. The parameters used were 6th order background model (computed from all *Arabidopsis* upstream sequences), $-n$ 2 (number of different motifs to search for), $-r$ 100 (number of times the MotifSampler should be repeated) and $-w$ (length of the motif) set to 8nt. For each cluster, the 20 best and non-redundant motifs (represented as a position weight matrix (PWM)) according to their log-likelihood score were retained using MotifRanking (default parameters; shift parameter $-s$ set to 2).

To create a non-redundant set of all motifs found in the different clusters of co-expressed genes, we first compared the similarity between two motifs as the PCC of their corresponding PWM. Each motif of length w was represented using a single vector, by concatenating the rows of its matrix (obtaining a vector of length $4 * w$). Subsequently, the PCC between every alignment of two motifs was calculated, as they are scanned past each other, in both strands [195, 232]. Then, all motifs with a $PCC > 0.75$ were considered as similar and only the motif with the highest NCS (see below) was retained.

The presence of a motif (represented by its corresponding PWM) in a DNA sequence was determined using MotifScanner, which uses a probabilistic sequence model (default parameters; prior probability $-p$ set to 0.1). Both MotifRanking and MotifScanner, together with MotifSampler, are part of the INCLUSIVE package [233].

3.4.4 Clustering based on TFBS content

To group genes containing similar motifs in their promoter and incorporating the possibility that not all motifs in a promoter are functional, we generated all groups of genes having two or more motifs in common. Starting from the set of nonredundant motifs mapped on all promoters, all motif combinations from two to four motifs were generated and only clusters with at least 20 genes containing that combination were retained. Note that, for a particular motif combination, the presence of additional motifs in a gene's promoter was ignored, resulting in the creation of overlapping clusters.

3.4.5 Network-level conservation score

We identified 3,167 orthologous *Arabidopsis*-poplar gene pairs through phylogenetic tree construction (see below). Due to the high frequency of gene duplication in both *Arabidopsis* and poplar [101, 234, 235], we preferred to apply phylogenetic tree construction to delineate orthologous relationships instead of sequence similarity approaches based on reciprocal best hit (for example, [201, 236]). Whereas the latter only uses similarity or identity scores to define putative orthology and is highly sensitive to incomplete associations due to in-paralogs, tree construction methods use an evolutionary model to estimate evolutionary distances and give a significance estimate through bootstrap sampling.

For each candidate TFBS and for all *Arabidopsis*-poplar orthologs, we first identified the set of *Arabidopsis* genes that have at least one occurrence matching the PWM in their upstream regions. Then, we also identified the poplar genes that have at least one occurrence matching the PWM in their upstream regions. Next, we calculated the overlap of matches in orthologs between both sets of sequences. Note that the matches can be anywhere in the upstream region and on any strand. For both *Arabidopsis* and poplar, the search was again restricted to the first 1,000 bp upstream from the translation start site or to a shorter region if the adjacent upstream gene is located within a distance smaller than 1,000 bp. The statistical significance of the overlap, which will be high for PWM representing functional TFBSs according to the network-level conservation principle, is measured using the hypergeometric distribution (for details, see [201]). Because the NCS, which is defined as the negative logarithm of the hypergeometric p value, is a relative measure of network-level conservation, the observed scores are compared against a distribution of scores obtained from random motifs. Thousand random motifs were generated by running the MotifSampler on clusters containing randomly selected genes. All NCS values larger than 5.3, which correspond to the 99th percentile of the random NCS distribution, were considered as significant.

3.4.6 Orthology determination

The full proteomes (that is, all proteins in a genome) of *Arabidopsis*, poplar, rice, and *Ostreococcus tauri*, together with proteins inferred from cDNA sequences for *Pinus taeda*, *Pinus pinaster* and *Physcomitrella patens* were used to delineate gene families using protein clustering. First, an all-against-all sequence comparison was performed using BLASTP [164] and relevant hits were retained [21]. Briefly, two proteins are considered homologous only when they share a substantially conserved region on both molecules with a minimum amount of sequence identity. In this manner, multi-domain proteins for which the sequence only partially overlaps because of shared single protein domains, which occasionally leads to significant *E* values in BLAST searches, are not retained as homologs. The proportion of identical amino acids in the aligned region between the query and target sequence is recalculated to $I' = \text{IxMin}(n_1/L_1, n_2/L_2)$, where L_i is the length of sequence i and n_i is the number of amino acids in the aligned region of sequence i . This value I' is then used in the empirical formula for protein clustering proposed by Rost [237]. Finally, all valid homologous protein pairs are subject to a simple-linkage clustering routine to delineate protein gene families. *Arabidopsis* and rice sequences were downloaded from TIGR (releases 5.0 and 3.0, respectively), *Ostreococcus* sequences from [238,239], poplar sequences from the JGI consortium [240], and pine and moss data from the Sequence platform for Phylogenetic analysis of Plant Genes database (SPPG) [241]. The coding sequences for *Ostreococcus* and poplar correspond to the genes predicted by the EuGene gene prediction software [242].

For all 7,038 gene families containing one or more *Arabidopsis* and poplar gene (and covering in total 20,273 and 31,894 genes, respectively), protein multiple alignments were created using T-coffee [243]. Alignment columns containing gaps were removed when a gap was present in >10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right of the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. Next, the median value for all these values was calculated. If this median was ≥ 0 , the column was considered as containing homologous amino acids. Neighbor-Joining phylogenetic trees were constructed with PHYLIP [244] using the Dayhoff PAM matrix and 100 bootstrap samples. Trees were rooted if a non-dicotyledonous species was present within the gene family. In total, 3,167 orthologous gene pairs were identified as speciation

nodes in the trees grouping one *Arabidopsis* and one poplar gene with high bootstrap support (≥ 70). An overview of the one-to-many and many-to-many orthologous relationships is shown in Additional data file 4. Note that these 3,167 orthologous gene pairs are not biased towards a particular functional GO class and thus can be used to estimate the conservation of candidate TFBSs between both plant genomes.

3.4.7 Functional annotation

GO [245] associations for *Arabidopsis* proteins were retrieved from TIGR [246]. The assignments of genes to the original GO categories were extended to include parental terms (that is, a gene assigned to a given category was automatically assigned to all the parent categories as well). All GO categories containing less than 20 genes were discarded from further analysis. Enrichment values were calculated as the ratio of the relative occurrence in a set of genes to the relative occurrence in the genome. The statistical significance of the functional enrichment within sets of genes was evaluated using the hypergeometric distribution adjusted by the Bonferroni correction for multiple hypotheses testing. Corrected p values smaller than 0.05 were considered significant. Only CRMs with significant GO Biological Process annotation and an enrichment score higher than 5 were retained in the final data set.

3.4.8 Expression coherence

The expression coherence, which is a measure of the amount of expression similarity within a set of genes, was calculated as described by Pilpel and co-workers [187]. Here, the PCC was used as a measure for similarity between expression profiles instead of the Euclidian distance used in the original implementation. Based on the similarity between expression profiles for 1,000 random genes (1,000 x 999 x 0.5 gene pairs), a PCC threshold of 0.5 (corresponding with the 95th percentile of this random distribution) was used to detect significantly co-expressed genes.

Additional data files

The following additional data are available with the online version of this paper and at http://bioinformatics.psb.ugent.be/supplementary_data/. Additional data file 1 is a figure showing the location of 34 conserved motifs (found in co-expressed

genes) in *Arabidopsis* promoters (2,445 genes) and of all conserved motifs in *Arabidopsis* promoters with more than 3 kb un-annotated upstream space (with distance <1,000 bp between position in *Arabidopsis* and poplar; 125 genes). Additional data file 2 is a figure giving an overview of the motif organization in orthologous *Arabidopsis* (left) and poplar (right) targets for module 2.M7010. Additional data file 3 is a figure showing the stability of clusters of co-expressed genes when randomly removing experiments from the complete expression data set. Additional data file 4 is a figure that gives an overview of the number of one-to-many and many-to-many orthologous relationships in the phylogenetic trees. Additional data file 5 is a table giving an overview of the 489 *Arabidopsis* microarray experiments. Additional data file 6 is a table giving an overview of the TFBSs identified using two-way clustering. Additional data file 7 is a table giving an overview of the 139 *cis*-regulatory modules. Additional data file 8 is a table showing the motif order and spacing for 30 *cis*-regulatory modules.

Acknowledgements

We would like to thank Kathleen Marchal for stimulating discussions and technical help with MotifSampler, Lieven Sterck and Stephane Rombauts for help with the poplar gene annotation and the DoE Joint Genome Institute and Poplar Genome Consortium for the poplar genomic sequence data. This work was supported by a grant from the Fund for Scientific Research, Flanders (3G031805). KV is a postdoctoral fellow of the Fund for Scientific Research, Flanders.

4

In situ Analysis of Cross-Hybridisation on Microarrays and the Inference of Expression Correlation

Tineke Casneuf, Yves Van de Peer and Wolfgang Huber
BMC Bioinformatics (2007) In press

The Chinese use two brush strokes to write the word 'crisis.' One brush stroke stands for danger; the other for opportunity. In a crisis, be aware of the danger - but recognise the opportunity

Richard M. Nixon

Abstract

Microarray co-expression signatures are an important tool for studying gene function and relations between genes. In addition to genuine biological co-expression, correlated signals can result from technical deficiencies like hybridisation of reporters with off-target transcripts. An approach that is able to distinguish these factors permits the detection of more biologically relevant co-expression signatures. We demonstrate a positive relation between off-target reporter alignment strength and expression correlation in data from oligonucleotide genechips. Furthermore, we describe a method that allows the identification, from their expression data, of individual probe sets affected by off-target hybridisation. The effects of off-target hybridisation on expression correlation coefficients can be substantial, and can be alleviated by more accurate mapping between microarray reporters and the target transcriptome. We recommend attention to the mapping for any microarray analysis of gene expression patterns.

4.1 Background

Microarrays are a valuable tool in functional genomics research. The breadth of their applications is reflected by the myriad of computational methods that have been developed for their analysis in the last decade. One popular practice is to compare expression patterns of genes by calculating correlation coefficients on expression level estimates across a set of conditions. Many downstream analysis tools are based on the presence or absence of correlation in the expression profiles of genes, like the inference of co-expression [247–251], gene regulatory [252] and Bayesian networks [253–256] and the study of gene family evolution [130, 257]. From a biological point of view, these approaches are useful and informative, but here we show that if care has not been taken as to how these correlations are calculated and how the reporters for each transcript are selected, incorrect conclusions can be drawn.

A gene is represented on a microarray by one or more reporters, i. e. nucleotide sequences that are designed to uniquely match its transcript, or transcripts if different splice variants exist [104]. Affymetrix GeneChips are the most widely used microarray platform, and a wealth of data measured on these arrays is publicly available. Affymetrix reporters are 25-mer oligonucleotides whose sequence is complementary to the intended target. Each target is represented by a set of reporters, called *composite sequences* [104] or *probe set* [258]. Probe set size varies between 11 and 20, depending on the type of array, but is the same for the majority of the probe sets within one array. The signals of these different individual reporters are combined into one expression value for the probe set in a step called *summarisation* [115, 117, 258].

The composition of the probe sets and the identifier of their gene transcript is contained in what is referred to as a CDF, a chip description file. Affymetrix, as array manufacturer, provides this information, and thanks to the openness of their technology specification, users can also construct their own custom-made CDFs. For Affymetrix' CDFs, probe set compositions are considered static and probe set annotation dynamic: with an updated annotation of a genome, the assignment of a probe set to a particular target gene can change, but never the content of its reporters [259]. For custom-made CDFs, this restriction is not necessary, as reporters can be arbitrarily assigned to targets.

Microarray technology confronts researchers with various challenges. Our understanding of transcriptomes is incomplete, and our estimates of which transcripts exist in a genome are constantly evolving. Therefore, for the analysis of microarray data it is important to ascertain that a reporter does in fact measure the transcript it was intended to target when the array was designed. Another concern is cross-hybridisation, where transcripts other than the ones intended hybridise to a reporter. The signal that is obtained for such a reporter will be that of a combination of multiple different transcripts.

The widespread use of expression arrays encouraged different research groups to study the extent and effect of hybridisation of cDNA molecules to reporters with mismatches in more detail. The cardinal importance of reporter annotation was underscored by observations made and evaluation tools developed by several research groups [260–263]. Dai et al. [263] conducted a comparative analysis of GeneChip data with original and redefined probe set definitions and described a discrepancy of 30 to 50% difference in the lists of reported genes using various analyses. These authors provide up-to-date reporter mapping files for various types of GeneChips that match individual reporters to transcripts. Based on the same observation of problematic reporter annotation, Zhang et al. [262] conducted an in-depth analysis of the reporter assignment on specific microarrays and pinpointed consistent but inaccurate signals across multiple experiments resulting from problematic reporters that are either non-specific or miss their target. They concluded that up to around 10% of the reporters on widely used arrays are non-specific in that they target multiple transcripts and another 10% miss their target. Different efforts have also aimed to model hybridisation strength and extent of cross-hybridisation to improve the design of high affinity reporters that are less prone to cross-hybridisation [264–267]. In addition, tools have been developed to infer the extent of cross-hybridisation of individual reporter sets subsequent to data analysis [268].

The technical aspect of the microarray technology has also been tackled: Eklund et al. [269] reported that replacing cRNA with cDNA hybridisation targets substantially reduces cross-hybridisation. Alternative technologies to detect cross-hybridisation on microarrays have also been suggested [270].

Wren et al. [271] described a positive relationship between the observed signal and the amount of contiguous hydrogen bonds involved in duplex formation during reporter-transcript binding. Okoniewski and Miller [272] conducted a large-scale analysis to map all interactions between reporters, probe sets and transcripts on the HGUI33A array. First, a set of basic motifs were defined to identify families of interacting probe sets as in some cases a reporter can bind more than one transcript, or a transcript can bind more than one reporter. The motifs were then used to build a bipartite graph of interactions with the probe sets and transcripts as nodes and matches as edges. The authors were able to identify several hub probe sets, whose expression combines the signals of many available transcripts. A detailed investigation of the expression signals revealed that reporters targeting multiple transcripts had higher absolute expression signal than those targeting a unique transcript, and that probe sets that contain reporters with multiple matches had increased expression correlation between them.

A different approach *in situ* was taken by Wu et al. [265] for the construction of a free energy model for cross-hybridisation. These authors observed a clear relationship between the known concentrations of spiked-in transcripts in different experiments and the measured signals of reporters not designed to target these specific transcripts. Based on the sequences of these affected reporters, the authors constructed a free energy model to assess the sequence dependence of cross-hybridisation which can be used to refine the algorithms used in reporter design. These different studies intelligibly show that cross-hybridisation is a critical concern for microarray analysis. It is clear that a reporter can bind different transcripts or that a transcript can bind to different reporters if stable, partial binding occurs or if hairpin structures are formed [273]. As a result, the signals of the reporters a transcript binds will be similar and correlation coefficients, calculated on these signals during downstream analysis, will be artifactual. The *in situ* effect of sequence similarity on expression correlation is however not known. For this study we worked with the ATH1 Affymetrix GeneChip that was designed for the analysis of gene expression in *Arabidopsis thaliana*. *Arabidopsis* is the most commonly studied model plant organism and a wealth of high quality data has been generated with this GeneChip. We investigated the relationship between reporter-to-transcript sequence similarity and correlation of expression signals. We assessed the extent to which inclusion of off-target reporters in probe sets, i. e. reporters that are highly alignable to another transcript than the intended one, influences this correlation. The conventional probe set design, as defined by the manufacturer of the microarray was evaluated with respect to cross-hybridisation and compared to our custom-made probe set composition.

We show that numerous probe sets on a widely used commercial array contain off-target reporters, and that inclusion of these reporters in a probe set gives rise to a signal pattern that is highly similar to that of the unintended probe set. We

illustrate our findings with examples and demonstrate the effect of individual reporters through simulation. Furthermore, we put forward a novel method to detect unreliable probe set to transcript hybridisation events. Our results show that excluding reporters that align well to another transcript diminishes this effect to a substantial extent and provides a method to pinpoint the occurrence of cross-hybridisation in existing microarray datasets. We conclude from this study that reporter-to-transcript sequence alignment strength can be a source of error in studies of correlation of expression signals and that proper probe set composition is effective in minimising the effect of cross-hybridisation.

4.2 Results and Discussion

4.2.1 Two definitions of probe set annotation

The ATH1 is an Affymetrix GeneChip for the analysis of gene expression in the premier plant model organism *Arabidopsis thaliana*. A wealth of high quality data measured with this array is publicly available and has been widely used for various applications, such as the inference of gene co-expression networks and the study of functional aspects of the evolution of gene families [130,247–251,257] (reviewed in [274]).

For the Affymetrix CDF of the ATH1, a probe set was assigned to a gene if nine or more of its reporters had perfect sequence identity with the gene's transcript consensus sequence. If this condition was fulfilled for multiple genes, the probe set was assigned to all of them. In this way, 22,810 probe sets were assigned to more than 24,000 genes. A probe set can thus contain up to eight reporters that align perfectly to another gene's transcript without being assigned to it [259].

We built a custom-made CDF with alternative probe set definitions and annotations. We aligned each 25-mer reporter sequence to the predicted transcripts of *Arabidopsis thaliana* (see Methods for details). A reporter was assigned to a gene if it had perfect sequence identity with its transcript(s) and did not align to any other gene's transcript with zero or one mismatches. We removed reporters that had multiple hits in the genome, and reporters that had hits in the reverse complementary direction. Probe sets were defined as eight or more reporters all assigned to a particular gene's transcript(s). This resulted in 19,937 probe sets with unique assignments to 19,937 target genes. Table 4.1 shows some statistics on the probe set definitions. The approach we took is highly similar to the one introduced by Dai et al. [263].

In those cases where their probe set annotations are based on the UniGene database, Dai and colleagues require perfect hits to unigene clusters and unique hits of a reporter to a genomic location. For their CDFs that are based on databases other than UniGene, the rule of one transcript assignment per reporter does not

	CDF Affymetrix	Custom-made CDF
Number of probe sets:	22,810	19,937
Number of reporters:	251,078	217,811
Number of alignment scores:	6,926,739,864	6,008,969,868
Total number of transcripts in TAIR6:		27,588

Table 4.1: Statistics of probe set definitions. The first 2 rows contain the number of probe sets and reporters in the Affymetrix and the custom-made CDF. The number of reporters times the number of predicted transcripts, in the bottom row, results in the total number of reporter-to-transcript alignment scores (see also Figure 4.1).

apply [263], so reporters can be assigned to multiple transcripts. As this is currently the case for the ATH1 array, for which the CDF of Dai et al. is based on the TAIR annotation, we computed a custom CDF that requires uniqueness. Hence, we expect that our results can be generalised to other arrays for which Dai et al. have computed CDFs with 1:1 reporter-target mapping, and in the future, when their ATH1 CDF will be changed to unique 1:1 mapping (personal communication), it could be used instead of our custom CDF.

4.2.2 Off-target alignments

Our aim was to investigate the relationship between correlation coefficients of microarray gene expression profiles and potential off-target sensitivity of reporters and probe sets. Figures 4.1A and B explain our procedure of calculating the score for off-target sensitivity. For a probe set with n reporters designed to target gene X , and another gene Y , we computed the alignment scores $\{a_1, \dots, a_n\}$ of X 's reporters to Y 's transcript sequence(s) with *Needle* [275], a Needleman-Wunsch alignment [276] program. A global alignment algorithm was used to align the full length of the reporter to the target while allowing for gaps and hairpin-forming. Furthermore, we used an exact algorithm to ensure that the optimal alignment was reached. *Needle* scores an identical match with a positive score of 5 and penalises a mismatch score with -4 . The gap open penalty was set to -50 and gap extension penalty to -0.5 . The reporters have a length of 25, so a perfectly matching reporter will have a score of 125. Some interesting scores are shown in Table 4.2.

To quantify the potential off-target affinity of a probe set, different percentiles Q_{XY}^p were calculated of the reporter alignment scores $\{a_1, \dots, a_n\}$, where $p \in [0, 100]$ is the percentile, X is the intended target gene of the probe set and Y is the potential off-target. For the results presented in this paper, we used $p = 75$, but qualitatively equivalent results were obtained with other values of p .

This analysis was carried out for each probe set against every sequence of the

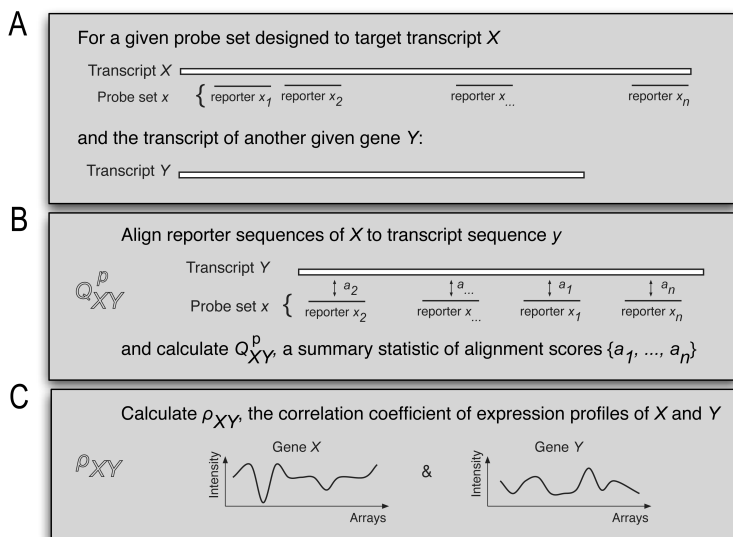


Figure 4.1: Illustration of our approach: A) for a given probe set x , assigned to measure the expression of gene X and the transcript of a given gene Y , two variables Q_{XY}^p and ρ_{XY} were calculated. B) Q_{XY}^p is a summary statistic (e.g. $p = 75$ for the 75% percentile) of the alignment scores of the reporters of X to the transcript of Y . C) ρ_{XY} is the correlation coefficient of the expression signals of genes X and Y . This procedure was repeated for each probe set against every other transcript of the *Arabidopsis* transcriptome.

transcriptome of *Arabidopsis* (as found in the TAIR6 sequence database <http://www.arabidopsis.org>), which results in a total number of 6,926,739,864 alignments for the Affymetrix CDF and 6,008,969,868 for the custom-made CDF (see Table 4.1). Figure 4.2 shows a histogram of the highest alignment scores of the pairs of the two CDFs.

4.2.3 Correlation of microarray expression profiles

Pearson correlation coefficients, ρ_{XY} were calculated for every pair of probe sets X and Y on two different ATH1 microarray datasets. One dataset contains expression data in 14 different plant tissues and the other is a dataset of nine stress conditions and consists of 60 datapoints (see Methods). Both datasets were generated by the AtGenExpress project [277].

Matches			Matches			Matches		
P	M	Score	P	M	Score	P	M	Score
25	0	125	22	3	98	17	0	85
24	0	120	21	2	97	20	4	84
24	1	116	20	1	96	19	3	83
23	0	115	19	0	95	18	2	82
23	1	111	21	3	93	17	1	81
22	0	110	20	2	92	20	5	80
23	2	107	19	1	91	16	0	80
22	1	106	18	0	90	19	4	79
21	0	105	21	4	89	18	3	78
22	2	102	20	3	88	17	2	77
21	1	101	19	2	87	16	1	76
20	0	100	18	1	86	19	5	75

Table 4.2: Table with some of the highest Needleman-Wunsch scores. P and M stand for the number of perfect and mismatch scores. Gap openings and extensions in the alignment were penalised with -50 and -0.5, respectively.

4.2.4 Probe set off-target sensitivity and expression correlation

The relation between expression correlation, ρ_{XY} and off-target sensitivity, Q_{XY}^{75} is shown in Figure 4.3. Figure 4.3A shows the results we obtained with all probe set pairs of the Affymetrix CDF and Figure 4.3C shows those of the custom-made CDF. These boxplots reveal a positive relation between the two variables: a gene whose expression is measured by reporters that align well to a different gene’s transcript tends to have an expression signal that is correlated with that of the other gene.

Because a positive trend between (reporter) alignment strength and expression correlation is not unexpected for functionally related genes like paralogous genes or genes that share protein domains, we defined a filtering criterion to set aside gene pairs that aligned to each other with BLAST [278] in at least one direction with an E-value smaller than 10^{-10} (see Methods). Figure 4.3B and Figure 4.3D show the data for the remaining probe set pairs of the Affymetrix and the custom-made CDF, respectively. For both, we see that for Q_{XY}^{75} values of up to around 70, the distribution of signal correlations of the probe set pairs is centred around zero. Pairs with higher Q_{XY}^{75} values are however accompanied by elevated signal correlation, even though for the gene pairs no functional relation is suggested by their sequence comparison. For a probe set with 11 reporters, the Q_{XY}^{75} summary statistic with $p = 75$ corresponds to the third strongest off-target reporter. A reporter alignment score value larger than 70 results from 15 or more perfect matches (cf. Table 4.2). Hence, our results imply that three or more well-

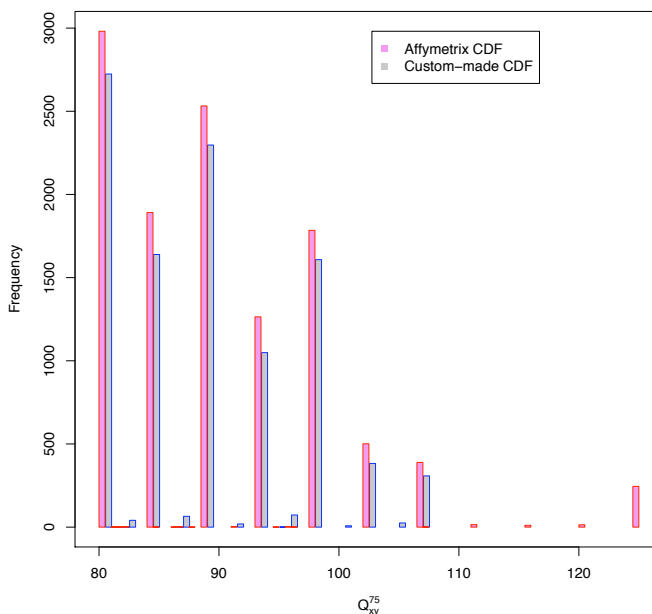


Figure 4.2: Custom-made versus Affymetrix CDF: barplot of the off-target sensitivity scores Q_{XY}^{75} of all probe set pairs in the Affymetrix (in pink) and the custom-made CDF (in light blue). This figure only shows pairs with an $Q_{XY}^{75} \geq 80$.

aligning off-target reporters in a probe set are associated with elevated expression correlation. Figures 4.3A and B also reveal that some probe sets in the Affymetrix CDF contain three or more reporters with perfect sequence identity to an off-target gene. These probe sets are in the rightmost boxes of these figures, corresponding to the score interval $(112, 125]$. The custom-made CDF does not contain such reporters, since all reporters uniquely map to their target gene's transcript and have at least two mismatches with any other sequence. As a result, the rightmost score interval in Figures 4.3C and D does not contain any probe sets, and the second-highest interval $(100, 112]$ contains only a few. A slight trend however remains. The results shown in Figure 4.3 were calculated on the tissue dataset, similar results were obtained for the stress dataset.

Different forces can give rise to the trend we observe here. First of all, genes with partially similar sequences can show biologically relevant expression correlation. Even though many such pairs will have been removed by the above filtering criterion, some may still remain in our dataset. Second, the trend can be due to cross-hybridisation, where the cDNA of a gene's transcript binds to both the

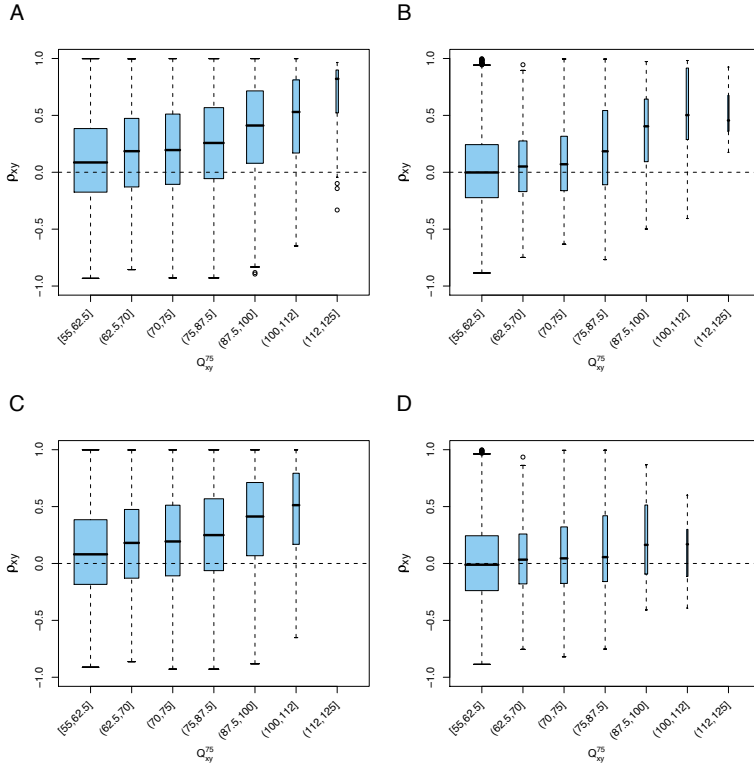


Figure 4.3: Boxplots depicting the expression correlation coefficients, ρ_{XY} stratified by off-target sensitivity score, Q_{XY}^{75} . Figures A and C show the data for all probe set pairs; for Figures B and D gene pairs with a BLAST hit in at least one direction with an E-value smaller than 10^{-10} were omitted. A-B) Results obtained with Affymetrix' CDF. C-D) Results obtained with the custom-made CDF. The widths of the boxes are proportional to the number of observations in each group. ρ_{XY} was calculated on the tissue microarray dataset. The plots show results for all pairs with $Q_{XY}^{75} \geq 55$.

reporters of its own probe set and those of other genes' probe sets. Both effects, functional relatedness and cross-hybridisation, can play at the same time.

4.2.5 Reporter off-target sensitivity and expression correlation

In an attempt to discern cross-hybridisation from functional relatedness and to identify incidences of unreliable reporter to transcript hybridisation, we designed a method that studies the behaviour of off-target sensitivity and signal correlation of different reporters within a probe set. For a probe set X and an off-target gene Y , we calculated the metacorrelation $\text{cor}(\rho_{X_iY}, a_i)$ between the alignment scores a_i of X 's reporters to Y 's transcript sequence and the Pearson correlation

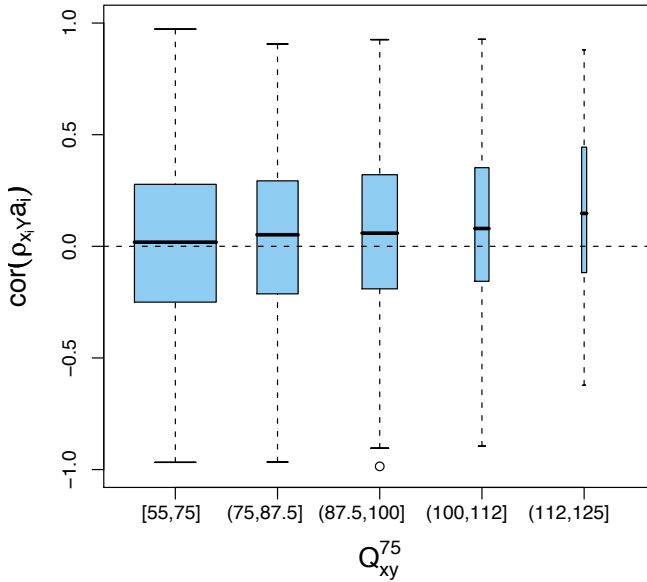


Figure 4.4: A boxplot showing the metacorrelation coefficients $\text{cor}(\rho_{X_i Y}, a_i)$ of all probe set pairs of the Affymetrix CDF, stratified by their off-target sensitivity score Q_{XY}^{75} . Only pairs with $Q_{XY}^{75} \geq 55$ are included. The correlation coefficients were calculated on the intensities measured in the tissue dataset.

coefficients of the reporters' signal patterns to the expression pattern of Y . We reasoned that if cross-hybridisation occurs, a positive trend between reporter to off-target correlation and the alignment score a_i can be detected. Conversely, lack of such a trend may indicate that cross-hybridisation is negligible.

Figure 4.4 depicts this metacorrelation coefficient for all probe set pairs with $Q_{XY}^{75} \geq 55$ of the Affymetrix CDF stratified by their off-target sensitivity score Q_{XY}^{75} . The results for the custom-made CDF are similar, except for the highest score interval $(112, 125]$, which does not occur with the custom-made CDF.

The distribution of the metacorrelations of most probe set pairs corresponds to a random distribution centred around zero. However, for those strata with high off-target sensitivity scores the distribution is shifted upwards. This means that within these probe sets some reporters do not correlate with the off-target, while others do, depending on their alignments score.

4.2.6 Examples

The metacorrelation method we developed was used to search for examples that illustrate our findings. Three examples are discussed in detail, each of which are presented in a row of Figure 4.5. The plots in the first column of this figure contain the summarised expression values of a probe set X (in blue) and an off-target gene Y (in orange) in the tissue dataset. The plots in the second column show the background corrected, normalised signal profiles of X 's reporters. The colour used to plot such a profile corresponds to the alignment score of that reporter to Y 's transcript and is explained in the legend in Figure B. In the third column, for each reporter ρ_{X_iY} , the Pearson correlation coefficient calculated between its signal profile and that of Y (orange in A-D-G) is plotted in function of its alignment score a_{X_iY} . The colours are identical to those used in the second column.

Probe set X in our first example is *245875_at*, which was designed to target gene *AT1G26240*, an extensin-like family protein. As shown in Figure 4.5A, the expression profile of this gene resembles that of *AT3G28550*, a protein that belongs to a zinc finger family. The Pearson correlation coefficient of these expression patterns is 0.63 in the tissue and 0.62 in the stress dataset. Figures 4.5B and C show that six of X 's reporters with $a_{X_iY} \geq 80$ have a signal profile that is highly correlated with that of *AT3G28550*. The remaining five have lower off-target sensitivity values and have a signal profile that is correlated less well with it. The Q_{XY}^{75} value of *245875_at* to *AT3G28550* is 89, the metacorrelation coefficient of the reporters of *245875_at* is 0.89.

The second example is of probe set *250857_at*, which was designed for *AT5G04790*, and gene *AT1G75180*. The function of both genes is unknown. Their ρ_{XY} is 0.70 and 0.89 in the tissue (in Figure 4.5D) and stress dataset respectively. Figures 4.5E and F reveal a positive relationship between off-target sensitivity and signal correlation. Interestingly, four reporters of probe set *250857_at* have 25 identical matches to *AT1G75180* and show an expression profile with $\rho > 0.8$. Two other reporters, with lower sensitivity to this off-target (107 and 89) also show high signal correlation to it. The Q_{XY}^{75} value of probe set *250857_at* to gene *AT1G75180* is 125, the metacorrelation coefficient of the reporters of *250857_at* is 0.62.

Figure 4.5G shows the expression patterns of probe set *258508_at* and *AT3G06650*. *258508_at* was designed to target *AT3G06640*, a protein kinase family protein. *AT3G06650* is a gene that encodes a subunit of the trimeric enzyme ATP citrate lyase. *AT3G06650* and *AT3G06640* are neighbouring genes that align for a stretch of about 50 base pairs with sequence similarity of $> 90\%$. The Pearson correlation coefficients of their expression profiles in the tissue and stress dataset are 0.30 and 0.16, respectively. Three reporters of *258508_at* have an off-target sensitivity to *AT3G06650* of 107 (Figure 4.5H and I). Two of them have a $\rho_{X_iY} \geq 0.6$, but the mean intensity of all three is higher than that of the other reporters. The Q_{XY}^{75}

value of this gene pair is 102.5, the metacorrelation coefficient of the reporters of probe set 258508_at is 0.55.

The examples presented here show that reporters that align best to the off-target Y have the most correlated signal with it and that the number of well aligning reporters plays an important role in the effect of cross-hybridisation. For example, the X probe set in our second example has several reporters with highly correlated signal profiles to the target: the four reporters that have perfect sequence similarity with it, as well as two others with alignment scores of 107 and 89. The Pearson correlation coefficient of the summarised expression pattern of this probe set pair is high in both expression datasets (0.70 and 0.89). In the first example five reporters show relatively high signal correlation to the off-target gene. The correlation of the summarised probe set values are 0.63 and 0.62. Different to these two, the probe set pair in our third example has a comparable Q_{XY}^{75} value but only two reporters show high signal correlation to gene Y . The correlation coefficient of this pair's expression pattern is much lower (0.30 and 0.16).

4.2.7 Effect of individual reporters on probe set summaries

It may come as a surprise that a few reporters out of 11 can affect the summarised expression profile of a probe set to the extent that their inclusion coerces it to resemble that of another gene. To better understand how this can happen, consider the following simulated data example. Assume that a gene A has a sinusoidal expression pattern over the course of 14 time points in an experiment.

Figure 4.6A shows the signal profiles of the 11 reporters of this gene's probe set, with data simulated using an established error model for microarray data [279]. The 11 reporters of a probe set B in Figure 4.6B show random signals without any underlying trend. Nine of the reporters of probe set C have identical signals as nine reporters of probe set B , while the remaining two reporters cross-hybridise with the transcript of gene A (Figure 4.6C). The summarised expression values obtained by applying the median polish method [122] are shown in Figure 4.6D. Interestingly, the Pearson correlation between probe set A and B is -0.07, while the correlation between A and C is 0.73. What is the explanation for this? The RMA method [115, 122, 123] exploits the fact that sensitivity to target abundance is strongly reporter-dependent and repeatable across arrays. RMA fits a model that explains the measured intensities as the product of a reporter effect and the target abundance. It estimates the model parameters, and hence the target abundance, with an outlier resistant method called *median polish*. These estimates can, however, be susceptible to subtle changes in the data, especially when the data from the reporters disagree, like here in our simulation [119].

We also explored other summarisation methods. With dChip [117, 280] for example, the effect of the two contaminating reporters is even stronger: the

correlation between A and B is 0.30, while it is 0.95 between A and C . The statistical model that dChip uses is similar to the one of RMA, however, there are differences in the variance assumptions and the robust estimation algorithm. Affymetrix' MAS 5 software uses an algorithm called one-step Tukey's Biweight [120]. This algorithm appears to be less influenced by the two off-target reporters: the correlation between probe set A and B is -0.22, while it is -0.19 between A and C .

4.3 Conclusions

Microarrays are an important source of functional data. Many inferential tools are based on the presence or absence of correlation in the expression profiles of genes, for example when inferring co-expression networks [247–251], in the study of the evolution of gene duplicates or families [130, 257] and in the inference of gene regulatory networks [252] or Bayesian networks [253–256].

Different research groups have pinpointed the critical concern of cross-hybridisation for microarray analysis [260–272]. Dai et al. [263] and Zhang et al. [262] highlighted problematic reporter annotation and underscored the importance of up-to-date reporter mappings. Zhang et al. [262] showed that about 10% of the reporters on widely-used arrays are non-specific in that they target multiple transcripts and approximately another 10% miss their target. Okoniewski and Miller [272] constructed a network of different levels of interactions between reporters and transcripts, as some reporters are able to bind more than one transcript, and some transcripts can bind more than one reporter. In this network they were able to identify several hub probe sets that show a higher absolute expression signal of reporters targeted by multiple transcripts than those that target a unique transcript because they combine the signals of many available transcripts. Moreover, their analysis revealed that probe sets whose reporters have multiple matches also show higher expression correlation with each other. Wu et al. [265] described a linear relationship between spiked-in concentrations and the measured signals of reporters that were not designed to target these particular transcripts.

We described a positive relationship between the correlation of microarray gene expression profiles and the off-target sensitivity of microarray probe sets, as estimated by sequence alignment of microarray reporters to off-target genes. Probe sets that contain reporters that align well to off-target genes show correlated intensity values to these other genes (Figure 4.3A and C).

In many cases, this positive relationship is likely not due to functional relatedness of the genes, but to a cross-hybridisation artifact. Three lines of argument support this statement: first, the positive trend is present even between gene pairs that do not share longer stretches of sequence similarity and where the reporter to off-target alignment is only based on short near-matches (Figures 4.3A ver-

sus B and C versus D). Second, this effect can be observed within probe sets (Figures 4.4 and 4.5). Third, omitting reporters liable to cross-hybridisation results in decreased artifactual correlation coefficients between probe sets (Figures 4.3B versus D).

Different summarisation methods perform differently when dealing with cross-hybridising reporters: methods that do majority weighting of reporters, such as RMA [115], can become unstable when there are two disagreeing groups of reporters that are close to balancing each other and when small changes can lead to a flip of the majority from one side to the other. Examples for this are shown in Figures 4.5 and by simulation. Simpler methods that are based on averages or trimmed averages, such as MAS [120], appear to be less affected by this problem, however, such methods suffer from the serious disadvantage of an overall smaller sensitivity [119, 124]. The latter thus cannot be regarded as a solution for the cross-hybridisation problem.

The standard probe set definition, as made available by the manufacturer of the array, Affymetrix, was compared to a custom-made one. In Affymetrix' definition, a probe set is a fixed set of reporters that is annotated to those genes to which a particular number of its reporters align perfectly. Probe sets can contain up to a certain number of reporters with perfect sequence identity to an off-target gene. In the custom-made CDF, a probe set is a set of reporters that align perfectly and uniquely to one gene's transcript. The use of more stringent probe set mapping and annotation results in decreased artifactual correlation coefficients. This will improve the quality of downstream analysis results. Our probe set definition is highly similar to the one used by Dai et al. [263]. Our results support and provide further evidence for the beneficial effect of probe set reorganisation they and others [262] reported.

In conclusion, off-target sensitivity is a factor that should be taken into account when doing correlation analysis from microarray data. High-quality assignment of reporters to target genes is essential for inferring genuine biological expression correlations. The correlation coefficient calculated between alignment strength and expression correlation coefficients, the metacorrelation coefficient, is a novel method to identify instances of unreliable reporter behaviour.

4.4 Methods

All analyses, except for the alignments, were done with development versions of R 2.6.0 [281] and Bioconductor 2.1 [173] packages. An R package, *XhybCasneuf*, containing a reproducible compendium of the datasets and scripts used for this study, is made available and is distributed through Bioconductor <http://www.bioconductor.org>.

4.4.1 Two Chip Description Files

This analysis was carried out on the GeneChip *Arabidopsis* ATH1 genome array of Affymetrix (<http://www.affymetrix.com/products/arrays/specific/arab.affx>). For Affymetrix' annotation of the probe sets, a file was downloaded from the Affymetrix website (<https://www.affymetrix.com/Auth/analysis/downloads/na21/ivt/ATH1-121501.na21.annot.csv.zip>) on August 12th, 2007. Affymetrix requires a 100% match of reporter's sequence to a consensus gene sequence and assigns a probe set to a particular locus if nine or more of the reporters in the probe set match it. We filtered out probe sets which Affymetrix assigned to multiple transcripts in addition to those that are assigned to a gene model that is not present in the TAIR6 (<http://www.arabidopsis.org>) sequence database.

For the custom-made chip description file, *Exonerate* [282] was used to map reporters onto the genome and transcripts. The target sequences were the predicted transcripts from the TAIR6 release, including mitochondrial and chloroplast-encoded genes. These sequences include UTRs but not introns. The fasta file was downloaded from TAIR (ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets/) on August 10th, 2007. We selected reporters that have perfect sequence identity with a single target gene's transcript. Reporters that hybridise with one mismatch to another gene's transcript are filtered out. We also filtered out reverse complementary matching reporters, and reporters that hybridise multiple times on the genomic sequence. The latter was done in order to remove reporters that match unannotated sequences. We included probe sets in this study if they consisted of at least eight reporters which resulted in 19,937 unique probe sets. The custom-made CDF is also available and distributed through Bioconductor (<http://www.bioconductor.org>, *tinesath1cdf*).

4.4.2 Reporter-to-transcript alignments

Reporter-to-transcript alignment scores were obtained with *Needle*, a global Needleman-Wunsch [276] alignment tool [275]. The analysis was carried out on the TAIR6 release of the *Arabidopsis* genome. The target sequences were the predicted transcripts, including mitochondrial and chloroplast-encoded genes and include UTRs but not introns. These cDNA sequences were downloaded from TAIR¹ on November 9, 2006. We ran the alignment analysis twice, with a gap penalty of -10 and -50. The same conclusions were reached but our findings were stronger when this penalty was set to -50. This means that higher correlation coefficients can be observed for reporter-to-transcript alignments without gaps.

¹ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR6_genome_release/TAIR6_cdna_20060907

4.4.3 Microarray data

The microarray data we used were generated within the framework of the AtGenExpress project [277]. The first set is a subset of the development dataset² and contains the expression data of genes in 14 plant tissues. The second contains expression data of plants under nine different abiotic stress conditions <http://www.weigelworld.org/resources/microarray/AtGenExpress/Sample%20list%20%28Abiotic%20stress%29>, measured over six different time points. Both datasets were normalised using RMA [115, 122, 123], summarised using a median polish algorithm and averaged over replicates.

4.4.4 Identification of gene pairs with long stretches of sequence similarity

To identify possibly functionally related gene pairs, we carried out a within-genome, all-against-all BLASTP [278]. Gene pairs with an E-value smaller than 10^{-10} in at least one direction were set aside during different parts of this study.

4.4.5 Metacorrelation

The metacorrelation was obtained as follows: for a probe set pair X and Y , the Pearson correlation coefficient was calculated between the alignment scores of X 's reporters to the transcript sequence of Y and the (Pearson) signal correlation coefficient of these reporters to the expression pattern of Y . We used the non-parametric measure for this metacorrelation because of the limited number of datapoints for each observation.

Acknowledgements

This work was supported by a grant from the Fund for Scientific Research, Flanders (3G031805) and by the European Commission through a Marie Curie Host Fellowship program (MEST-CT-2004-513973). WH acknowledges support from the European Commission through the Integrated Project Heart Repair (LSHM-CT-2005-018630). Grateful acknowledgements are made to Richard Bourgon, Jörn Tödling and Stefanie De Bodt for fruitful discussions.

Authors' contributions

TC designed the study, analysed data, and wrote the paper. YVdP wrote the paper. WH designed the study, supervised the project, and wrote the paper.

²http://www.weigelworld.org/resources/microarray/AtGenExpress/AtGE_dev_samples.pdf

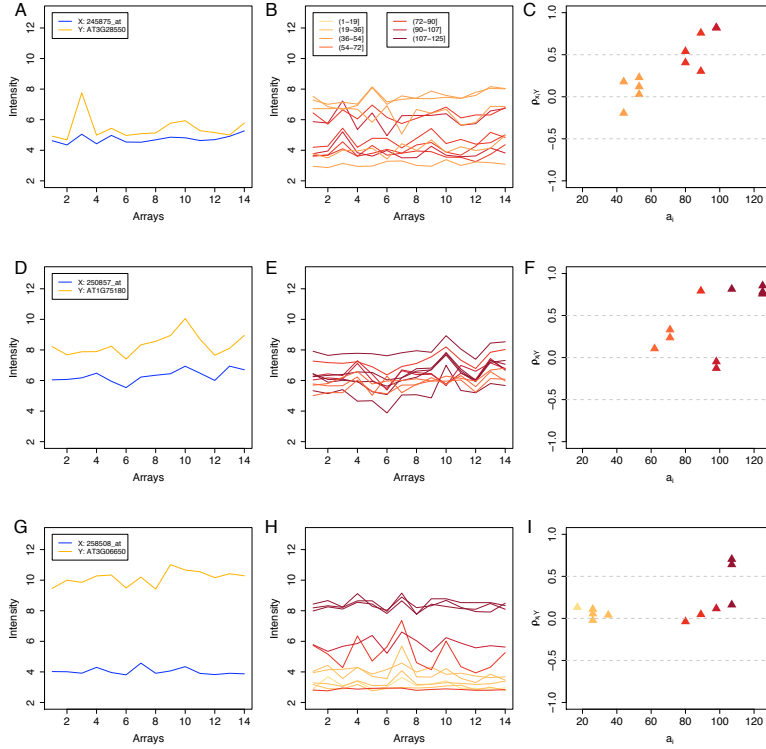


Figure 4.5: Three examples of cross-hybridisation: each of the three rows presents an example of cross-hybridisation. Each time, the first of the plots (A-D-G) shows the summarised expression values of probe set X (in blue) and probe set Y (in orange) in 14 different plant tissues. The plots in the second column (B-E-H) present the background corrected, normalised expression patterns of X 's reporters. The signal profile of the reporter is plotted in a colour that corresponds to its alignment score to Y and is explained in the legend of plot B. In the third column (C-F-I) for each of X 's reporters, $\rho_{X_i Y}$, calculated between its signal profile to that of Y , is plotted against its alignment score, $a_{X_i Y}$. Colours correspond to those used in the plot in the second column.

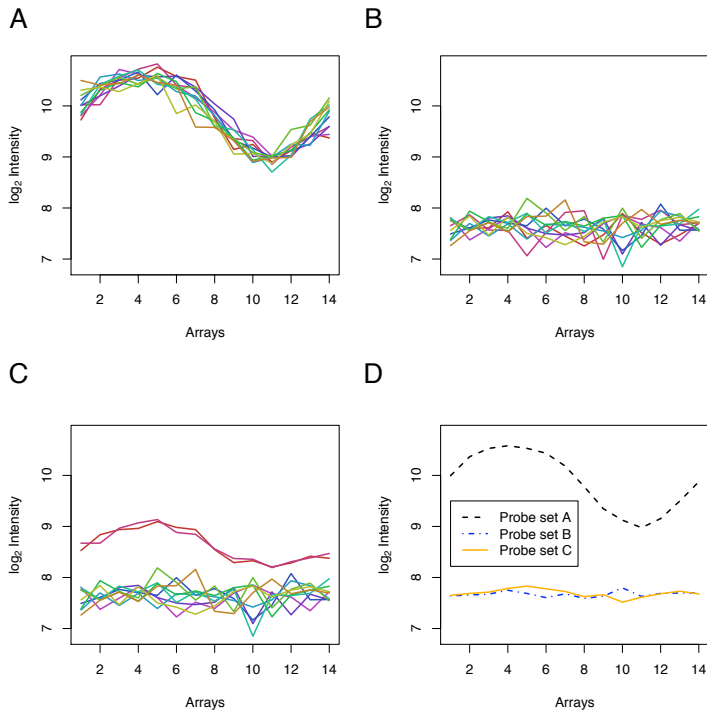


Figure 4.6: A) The expression profiles of the reporters of a probe set *A* that binds the transcript of a target gene with a sinusoidal expression pattern. Each reporter is drawn in a different colour. B) The expression profiles of eleven reporters of a probe set *B* that show random signals without any underlying trend. Each reporter is drawn in a different colour. C) Nine of the reporters of a probe set *C* have identical expression values as nine of those of probe set *B*. Two other reporters of this probe set cross-hybridise with the transcript of gene *A* and thus have a expression pattern that is highly similar to the reporters of probe set *A*. The expression values of these two reporters are coloured red. The other nine have the same colours as the corresponding reporters of probe set *B* in Figure 4.6B. D) The expression patterns of these three probe sets after summarisation with *median polish* [115, 122, 123].

5

Application of the Microarray Technology to the Study of Evolution and Functional Divergence of Duplicated Genes

Comparative biologists may understandably feel frustrated upon being told that they need to know the phylogenies of their groups in great detail, when this is not something they had much interest in knowing. Nevertheless phylogenies are fundamental to comparative biology; there is no doing it without taking them into account.

Felsenstein, 1985

Abstract

Our in-depth study of off-target transcript hybridisation to reporters on microarrays revealed a positive relation between alignment strength and expression correlation. The knowledge thereby gained about cross-hybridisation is applied here to the study of evolution and functional divergence of duplicated genes in *Arabidopsis thaliana*. The use of an up-to-date and, with respect to sequence similarity, stringent probe set definition allows accurate assessment of the divergence pattern of independent duplicates belonging to various functional categories. The molecular function of a gene and the biological process in which it functions are important players in the divergence tale of duplicated genes.

5.1 Background

Microarrays allow genome-wide monitoring of gene expression levels across different experimental conditions or treatments. The technology has empowered significant proceedings on various research topics, amongst them in the study of gene duplicate evolution, as discussed in the Background and Conclusions sections of Chapter 2 of this dissertation. In this Chapter 2, we described non-random expression divergence for duplicated genes generated by different duplication mechanisms and belonging to different functional classes. In the approach we took, we however did not incorporate information about the phylogenetic relationships of the genes under study. Because gene expression levels can be viewed as quantitative traits, improvement to our analysis could be introduced by making use of statistically more sophisticated, phylogenetic comparative methods, as proposed by Gu et al [283], Oakley et al [284] and Guo et al [285] or ANOVA-types (nested design) of approaches, as proposed by Duarte et al [148] and Gu et al [286]. However, some issues hinder their wide-spread application. As with any comparative method that studies evolution, one of the assumptions that is made is that each of the branches of a phylogenetic tree can be used as an independent data point when testing for the correlation of traits among different outer leaves of the tree [287, 288]. Therefore, in a bifurcating tree with n leaves, there will always be $2n - 1$ apparent degrees of freedom, but for two reasons, there may be actually be fewer. If an ancestral node is in state 0, then each of its radiating daughter nodes is constrained to either change to 1 or not change at all: they cannot change from 1 to 0. Second, the ancestral state is not known and is in fact estimated from the characteristics of the outer leaves, which also introduces dependence. In this respect, the current generation of sequence and expression data of closely-related species can complement our present-day knowledge and greatly increase confidence about the estimates of the ancestral state. For research on duplicated genes, this means that contemporary data points in the outer leaves,

that are connected via multiple internal nodes, are not independent. They share an evolutionary history up until their last common node, except from the relatively small amount of independent evolution since their split from this last common ancestor. Also, the exact divergence pattern of duplicated genes remains to be elucidated, as is the level of mutual dependence of duplicates. For instance, in the case of sub- or neo-functionalisation, selection pressure is hypothesised to constrain the divergence of both or one of the duplicates in order to maintain the functionality of the ancestral gene.

Our in-depth study of off-target transcript hybridisation to reporters on microarrays, described in Chapter 4, revealed a positive relation between alignment strength and expression correlation. We observed that genes that share long stretches of high sequence similarity are susceptible to severe bias of inferred co-expression relationships from microarray data. We learned that up-to-date and, more importantly stringent reporter-to-transcript assignment is of great importance when studying duplicate gene evolution.

In this chapter, insights in both the study of evolutionary traits [287, 288] and the microarray technology (Chapter 4) are applied to improve our analysis of divergence patterns of duplicated genes belonging to different functional categories in *Arabidopsis thaliana*. This analysis confirms that the molecular function of a gene and the biological process in which it is involved play an important role in the divergence rate of duplicates and is responsible for dissimilarities of divergence rates of genes belonging to different functional classes.

5.2 Results and Discussion

5.2.1 Independent duplicated genes

Regarding the selection of paralogs for our investigation in Chapter 2, namely all that were identified with BLASTP and passed the filtering step according to Li et al [21], is the approach we took an efficient, but pragmatical one. For the analysis in this Chapter, a possible improvement concerning the dependency of duplicated genes was implemented. Independent pairs were selected by taking into account the duplication relationships. For the paralogs identified with the above-mentioned method, clusters were formed of genes with mutual duplication relationships, the result of which is very similar to the construction of gene families. Independent duplicates were then selected in each cluster starting with the pair with the smallest K_S value. After excluding both genes from further selection, additional pairs were selected by extracting the gene pair with the smallest K_S value among the

remaining genes. A gene can thus only be assigned to one pair. The independence of these pairs lies in the fact that this way only the outer leaves of the phylogenetic tree will be studied. The selected gene pairs have evolved independently since the split from their last common ancestor.

5.2.2 Functional annotation and expression data

The goal of this investigation was to compare the divergence rates of duplicates with different molecular functions or that are involved in different biological processes. To that end, independent pairs were assigned to functional classes to which both members are annotated in the Gene Ontology (see Methods section). Their gene expression divergence patterns were then analysed by calculating the Spearman rank correlation coefficient on a data set of nine stress conditions (see Methods). These microarray data were pre-processed according to the probe set definition introduced in Chapter 4. For this custom-made CDF, a reporter was assigned to a gene if it had perfect sequence identity with the transcript and did not align to any other gene's transcript with zero or one mismatches. Reporters with multiple hits in the genome were removed, as are reporters with hits in the reverse complementary direction. Probe sets were defined as eight or more reporters being assigned to the same gene. This CDF contains 19,937 probe sets with unique assignments.

5.2.3 Expression divergence in different functional classes

Figures 5.1, 5.2 and 5.3 show the Spearman correlation coefficients of gene pairs in function of the time since their duplication. Duplicates with K_S values up to 1.5 are shown, as this is the estimated upper bound for duplicates generated by the last whole-genome duplication (3R) [83]. To show and compare the underlying divergence rate trends, a local regression line (blue solid) is fitted to each of the classes. The 95% confidence intervals are depicted with a dashed blue line. These plots reveal notable differences across the various functional classes.

Gene pairs with slow expression divergence

The duplicates in the classes depicted in Figure 5.1 show high correlated expression patterns: especially among the youngest pairs high expression correlation can be observed but also for the older the correlation coefficients are centred around 0.5. Duplicates that retain highly correlated expression patterns are genes involved in organelle organisation and biogenesis, biosynthetic processes (such as of macromolecules, shown here) and binding (nucleic acid and its child, DNA binding, shown here). Other classes of slowly diverging genes are genes with

transcription regulator and structural molecule activity.

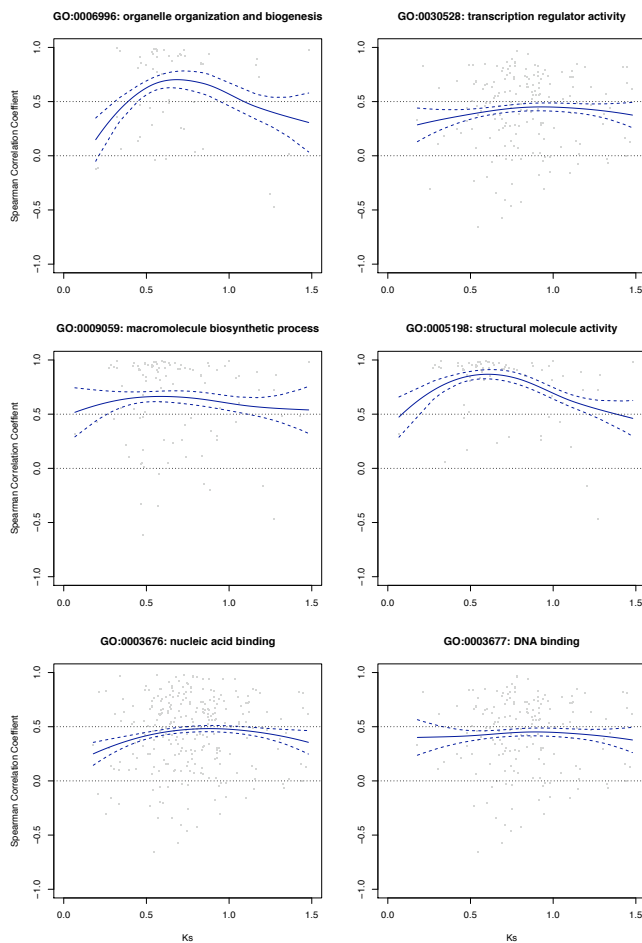


Figure 5.1: Duplicates belonging to functional classes of slowly diverging genes.

Gene pairs with quick expression divergence

For the gene pairs of some other classes, a considerable different picture can be observed: genes that are involved in developmental processes and biopolymer modification and enzymatic genes, such as those with kinase and oxidoreductase activity and response genes, like to chemical stimuli and stress, turn out to have diverged quickly after duplication. Few of the young pairs in these functional

classes have high correlated expression patterns and the correlation coefficients of the older duplicates is centred somewhere around 0.25.

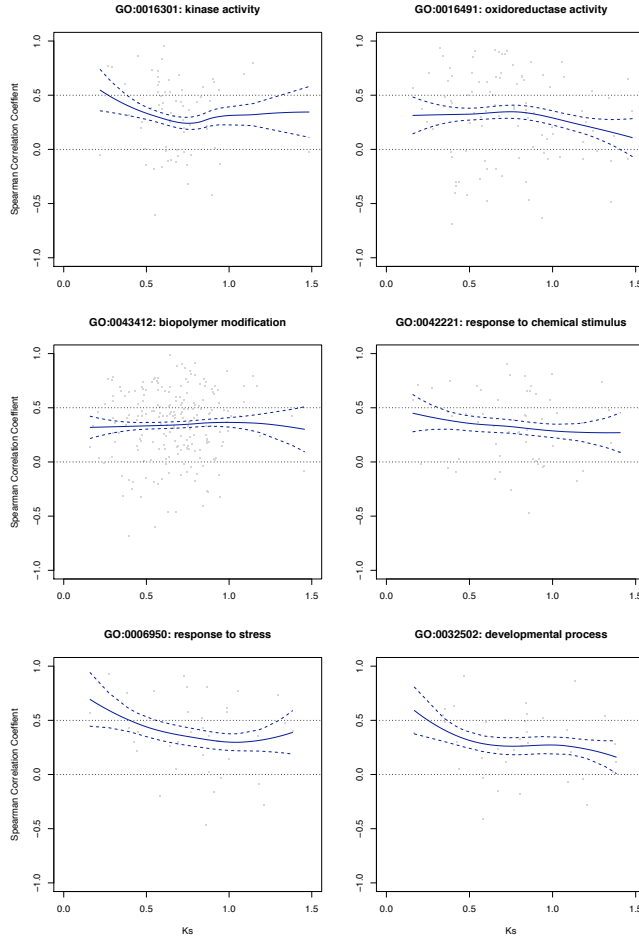


Figure 5.2: Duplicates belonging to functional classes of quickly diverging genes.

Gene pairs with intermediate expression divergence

Genes belonging to some other classes show divergence patterns with intermediate rates. Examples thereof include genes that are involved in cell communication and regulation of metabolic processes and genes with transporter and hydrolyse

activity.

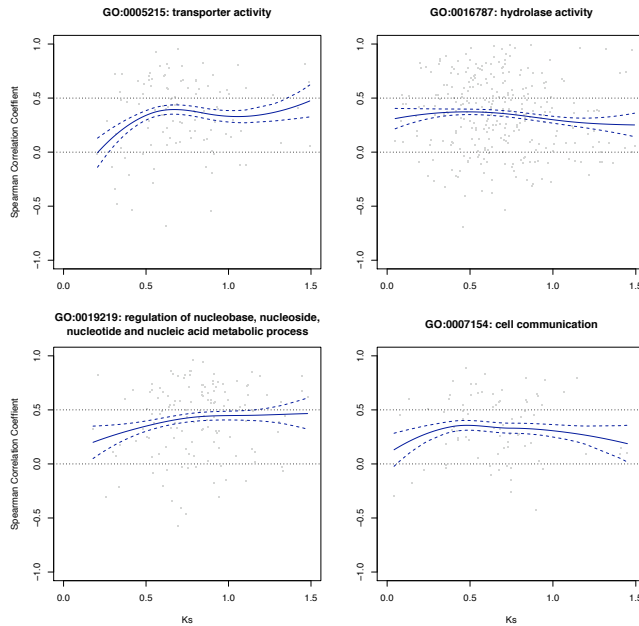


Figure 5.3: Duplicates belonging to functional classes of moderately diverging genes.

5.3 Conclusions

From our in-depth investigation of off-target transcript hybridisation to reporters on microarrays we learned that cross-hybridisation is a factor that should be taken into account when inferring expression relationships, as it possibly causes spurious correlations. Advances from this study are combined with improvements in the selection of duplicated genes for the study of evolution and functional divergence of duplicated genes in *Arabidopsis thaliana*.

Carried out on gene pairs that are selected based on statistically more sound criteria and expression data that was pre-processed with a probe set composition and definition that minimises the effect of cross-hybridisation, this study confirms the observations made in Chapter 2. Clear differences in divergence rates exist between duplicate pairs belonging to different functional classes. For some classes, fast expression diversification is selected against, probably due to the essential nature and sensitive regulation of these highly conserved processes.

Other genes that are involved in development and reactions against environmental changes or stress and enzymatic genes are confirmed to have diverged quickly after duplication which might suggest that the ancestors of *Arabidopsis* quickly put these newborn genes into use by means of altered and diverged expression patterns, as compared to their ancestral copy, to survive and cope with environmental changes.

Compared to the plots shown in this Chapter, the distributions of data points in Figure 2.5 are centred towards lower values of correlation coefficients. In the latter, numerous gene duplicates that are connected via one or multiple internal nodes are included. Most likely, these duplicates have diverged to such an extent that their expression patterns are unrelated and similar to random pairs with no expression correlation.

In short, the molecular function of a gene and the biological process in which it functions are important players in the divergence tale of duplicated genes.

5.4 Methods

5.4.1 Identification of independent duplicated genes

To identify duplicated genes, an all-against-all protein sequence similarity search was performed using BLASTP (with an E value cut-off of e^{-10}) [164], followed by the application of a criterion based on length and sequence similarity, according to Li et al. [21]. If different splice variants exist, the gene with the longest transcript sequence was selected. Transposons/able elements were filtered out by searching the annotations of the genes for *retrotransposon*, *retro*, *Mutator*, *hAT-like*, *hobo*, *mutator-like*, *CACTA-like*, *transposase*, *reverse*, *copla-like*, *retroelement*, *Athila*, *non-LTR*, *IS-element*, *IS4* and *hAT dimerization*, as according to Thomas et al [289].

To determine the time since duplication, the fraction of synonymous substitutions per synonymous site (K_S) was estimated. These substitutions do not result in amino acid replacements and are believed to be, in general, not under selection. Consequently, the rate of fixation of these substitutions is expected to be relatively constant in different protein coding genes and, therefore, to reflect the overall mutation rate. First, all pairwise alignments of the paralogous nucleotide sequences belonging to a gene family were made by using CLUSTALW [165], with the corresponding protein sequences as alignment guides. Gaps and adjacent divergent positions in the alignments were subsequently removed. K_S estimates were then obtained with the CODEML program [166] of the PAML package [167]. Codon frequencies were calculated from the average nucleotide frequencies at the three codon positions (F3 x 4), whereas a constant K_N/K_S (nonsynonymous

substitutions per nonsynonymous site over synonymous substitutions per synonymous site, reflecting selection pressure) was assumed (codon model 0) for every pairwise comparison. Calculations were repeated five times to avoid incorrect K_S estimations because of suboptimal local maxima.

In a next step gene clusters were formed of all gene pairs with mutual duplication relationships, according to the criteria defined above, that in practice are very similar to gene family clusters. The formation of this clusters was done in R (2.6, 2007-07-02 r42107) using the RBGL and Rgraphviz libraries. Within each cluster, independent duplicates were singled out, starting by selecting the pair with the smallest K_S value and proceeding with every increasing K_S value. A gene could only be picked once: once selected, it could not be included in another pair.

5.4.2 Gene Ontology functional categories

Independent duplicate gene pair were then assigned to Gene Ontology (GO) categories. The annotation file was downloaded from the GO website¹ on Thursday, August 23rd, 2007. The GO annotation of gene is the term it is assigned with in this file, in addition to each of the ancestors in the GO tree, a step that is conducted in R, with the GO package. An independent gene pair is assigned to a gene category if both members were annotated with the particular category. Assignments with all evidence codes were included, but even if a more stringent approach is taken where genes assigned with evidence codes 'IC', 'IDA', 'IGC', 'IGI', 'IMP', 'IPI', 'NAS', 'RCA' and 'TAS' were included, the results hold. However, application of this filtering step renders much fewer data points so that regression is difficult.

5.4.3 Microarray data

The microarray data used for this analysis were generated within the framework of the AtGenExpress project [277] and contains expression data of *Arabidopsis* plants under nine different abiotic stress conditions [290], measured over six different time points. The data were normalised using RMA [115, 122, 123], summarised using a median polish algorithm and averaged over replicates. Our custom-made CDF ('tinesath1cdf') was used, where each reporter is assigned uniquely to a sole transcript and are excluded if it aligns to a different transcript with 24 or more perfect matches (see Chapter 4).

¹http://cvswb.geneontology.org/cgi-bin/cvswb.cgi/go/gene-associations/gene_association.tair.gz

5.4.4 Correlation analysis

Expression correlation of two duplicated genes was calculated in R [152], using the non-parametric Spearman Rank correlation coefficient ρ :

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

where D is the difference between the ranks of the corresponding expression values of both duplicated genes and N is the number of samples.

5.4.5 Regression analysis

The relation between expression correlation, measured as the Spearman correlation coefficient, and time, measured as the number of synonymous substitutions per synonymous site K_S , was studied using 'locfit', an R package to fit curves and surfaces to data, using local regression and likelihood methods [152, 177]. All duplicated genes with a K_S value smaller than or equal to 1.5 were included.

6

Concluding Remarks

If we begin with certainties, we shall end in doubts; but if we begin with doubts, and are patient in them, we shall end in certainties.

Francis Bacon

Partnered by major technological advances, the accumulation of diverse genome sequences at the beginning of this century brought about various exciting breakthroughs in life sciences, such as the identification of the gene set of which genomes are comprised by genome annotation, perspectives on genomic structures as a result of major gene expansion and loss events by evolutionary and comparative genomics, the determination of biodiversity by population genomics, and the determination of genes' functions through functional genomics. This build-up of information enabled life science research to shift from gene-focused studies to exhaustive systems biology approaches in which the role that each individual gene plays in the greater picture of the functioning cell is revealed. Assumptions and hypotheses previously inferred from observational studies were to be assessed and verified.

Non-Random Divergence of Gene Expression

Significant advances have also been made in the study of duplicated genes. Based on experimental and observational evidence, gene duplication had since long been put forward as an important source of genetic novelty and had been linked to important advances in evolution. The availability of genome-wide expression data of various organisms allowed formal testing of the different models explaining the fate of duplicated genes that had been suggested previously. Functional analysis in human [143], yeast [142, 144], cotton [145] and *Arabidopsis* [130, 147, 148] had revealed that duplicated genes acquire distinct expression patterns quickly after duplication. The importance of factors playing in this process, such as mode of duplication and gene function, had however not been appreciated.

Gene duplication has been particularly prevalent in plants, including in the model system *Arabidopsis thaliana*, who has been shown to have undergone several rounds of large-scale, whole-genome duplication and many smaller, local duplication events [99–101, 149]. Investigation of gene and genome duplication events as well as the subsequent functional divergence of genes is of fundamental importance in the understanding of evolution and adaptation of organisms. Previous work conducted in our research group [83] involved the categorisation of duplicates in the *Arabidopsis* genome [101, 291, 292] and modelling of gene duplication events in this plant model organism [83]. To simulate the duplication dynamics of genes in *Arabidopsis* for different functional classes, we developed an evolutionary model in which all three whole-genome duplication events and a continuous mode of gene duplication were considered and fitted to the observed age distributions of duplicated genes [83]. Our study revealed that decay rates and retention are strikingly different for large- and small-scale events and highly biased towards the functional class to which a gene belongs. Results that were confirmed by subsequent studies [289, 293–295].

The above findings urged us to quantify the extent to which these factors direct the fate of retained duplicates [257]. Using microarray data, we conducted an in-depth analysis of expression divergence of duplicated genes with respect to time and mode of duplication and their encoded functions. Duplicated gene pairs were identified by conducting an all-against-all transcript sequence comparison and applying a criterion based on length and sequence similarity, according to Li et al. [21]. The number of synonymous substitutions per synonymous site were used as a proxy for their age. The duplicates were then divided according to their age and into sets of duplicates that arose through either large- or small-scale duplication events. An in-house developed tool, iADHoRe [291] for the detection

of homologous regions, was employed to this end. The extent of expression pattern divergence was assessed through the use of a publicly available perturbation microarray data set of *Arabidopsis*.

Our study revealed that genes that arose via large-scale events diverge slower than those that got duplicated through local small-scale events. In addition, analysis of their transcriptional absence and presence in fourteen different plant tissues showed that duplicates that the latter tend to diverge asymmetrically, in the sense that one member of the pair is expressed in a large number of tissues while the other is expressed in few tissues. The duplication mechanism clearly is an important determinant in the divergence tale of the newly-created gene pair. We hypothesised that small-scale duplicates, produced by unequal crossing-over and duplicative transposition [153], are likely much more prone to promoter disruption than genes duplicated through large-scale events. Similarly, translocation of a duplicate could also disrupt promoters and dislodges a gene from its transcriptional settings Both appear to result in the altered expression, and even asymmetric divergence of duplicates [88, 154]. A similar, more recent study, by Cusack and Wolfe investigated asymmetric sequence divergence of duplicated genes and reported that the degree of rate asymmetry of gene pairs where one copy has been relocated is greater than in pairs formed by local duplication events [296]. The authors take the asymmetry as evidence of natural selection's ability to discriminate between two duplicate copies and that it subjects them to different levels of purifying selection, or even permits adaptive evolution of one or both copies. Like us, they state that neighbouring duplicates share genomic context, i.e. *cis*-regulatory and distal elements, chromatin domain and gene neighbourhood, gene relocation has a strong impact on the asymmetry between genes, of protein evolutionary rates, in the case of their study. A similar observation on a small gene-family in diploid and tetraploid wheats has been made by Akhunov et al [297].

Our study also revealed remarkable differences in expression divergence rates of genes belonging to different molecular functions or that are involved in different biological processes. For instance, genes involved in signal transduction and response to stimuli, like stress and external stimuli, diverge relatively quickly, which could be the result of an evolutionary mechanism that the ancestor of *Arabidopsis* has evolved to meet the challenges of a changing environment. Particular highly conserved proteins, such as ribosomal proteins, or genes involved in conserved processes, such as biosynthesis pathways or photosynthesis, on the other hand turned out to maintain highly correlated expression patterns.

Our analyses have been limited to the study of general divergence patterns

and identification of factors that play a role in this. Future challenges involve making the distinction between diversification of the coding sequence and changes in regulatory control to which functional divergence can be attributed [298] and establishment of whether the asymmetric divergence patterns we observe result from non-, neo - or sub-functionalisation [299]. It would also be interesting to identify the co-evolution of genes that are for instance functioning in a particular pathway [130], or to study how the evolution of sequences of transcription factors affects the transcription binding pattern [300–303]. Sequence information on closely related species, together with expression compendia and genome-wide and across-species knowledge of regulatory elements, will allow the generation of reliable phylogenetic trees and the inference of the ancestral expression state, so as to identify expression pattern shifts and to establish the exact divergence histories.

Identification of Novel Regulatory Modules

A major goal in the post-genomic era is the identification of all functional elements of which genomes are comprised, including those that regulate expression. The elucidation of the exact dynamics by which transcription is regulated is important, as reflected in the examples of diseases that alterations of transcriptional components bring about [304]. Basic regulation of the timing, level and location of gene expression operates through binding of transcription factors to elements in the promoter regions of genes. The difficult task that the identification of these short and degenerate *cis*-regulatory elements constitutes can be alleviated by complementation with knowledge about evolutionary conservation (orthologous genes) and shared regulatory control (co-expressed genes). Comparative genomics is a powerful tool to improve the detection because functional noncoding sequences are often evolutionary conserved across species [194, 305–310]. Genes that share regulatory elements in their upstream regions are thought to be co-regulated and as a result show similar expression patterns [311–313]. By taking the reverse approach, i.e. treating expression levels as quantitative traits, calculating the correlation between expression patterns of genes and clustering them, shared elements can be identified.

The identification of novel regulatory elements in plants was the focus of our research conducted in Chapter 3 of this dissertation, in which advantage was taken from the accumulation microarray data for *Arabidopsis thaliana* and the availability of the genomic sequence of a related species, *Populus trichocarpa* [314]. First, transcription factor bindings sites were identified in sets

of co-expressed genes by applying a classic Gibbs-sampling approach. Functional elements were then selected by presenting the output to an evolutionary filter, based on conservation in orthologous genes in *Populus*. Next, a two-way clustering procedure that combined the presence or absence of motifs and expression data was used to identify additional novel regulatory elements. Sets of genes containing a particular element or motif, a combination of elements, were then annotated by making use of the Gene Ontology annotation. This resulted in the identification of 80 transcription factor binding sites and 139 regulatory modules, most of which are novel. These modules consist primarily of two or three regulatory elements that could be linked to different important biological processes, such as protein biosynthesis, cell cycle control, photosynthesis and embryonic development. Moreover, study of the physical properties of some specific regulatory modules revealed that *Arabidopsis* promoters have a compact nature, with cooperative transcription factor binding sites located in close proximity of each other.

Genes are expressed by the orchestrated interplay of numerous proteins, amongst which kinases, polymerases, transcription factors and coactivators, and events that include cellular signalling, activation and repression, DNA methylation, histone modification and chromatin remodelling [315–317]. Future work involves the development of computational approaches for the integration of information from the more than 600 genomes that have completely been sequenced at present (<http://gold.imbb.forth.gr/>, [318]) and various sources and levels of functional data [319], from technologies such as yeast two-hybrid screens, tandem affinity purification, cDNA and protein microarray experiments, chromatin immunoprecipitation assays, mass spectrometry, fluorescence microscopy and protein structure prediction. Comprehensive systems biology approaches will have to be taken to annotate all transcriptional regulatory elements, to provide in-depth views on entire gene regulatory mechanisms and to clarify the importance of *cis*-regulatory modification for adaptation and morphological and developmental evolution.

Cross-Hybridisation on Microarrays

Microarrays are valuable instruments for measuring gene expression on a genome-wide scale. Co-expression relationships thereby obtained are often used in systems biology to infer functional modules and regulatory networks. Many of the downstream analysis tools are based on the presence or absence of correlation in the expression profiles of genes, like the inference of co-expression [247–251], gene regulatory [252] and Bayesian networks [253–256] and the study of gene family evolution [130, 257]. From a biological point of view, these approaches

are useful and informative, but with the analysis presented in Chapter 4 of this dissertation, we show that if care has not been taken as to how these correlations are calculated and how the reporters for each transcript are selected, incorrect conclusions can be drawn.

The microarray technology confronts researchers with various challenges. Our understanding of transcriptomes is incomplete, and our estimates of which transcripts exist in a genome are constantly evolving. Because microarrays are often designed based on sequence information of early releases of a genome, it is important to ascertain that a reporter in fact picks up the mRNA of the target gene it was intended to pick at the time of array design. The cardinal importance of reporter annotation was underscored by studies conducted by several research groups [260–263]. Another concern is cross-hybridisation, where off-target transcripts, i.e. other than the intended ones, hybridise to a reporter. Cross-Hybridisation can in fact occur at the level of the reporter, when a single-stranded DNA sequence binds to a reporter which is not completely complementary, or at the probe set level, where a reporter of the probe set is complementary to an off-target transcript. The signal that is obtained for such a reporter or probe set will be that of a combination of multiple different transcripts. Cross-Hybridisation leads to spurious positive correlations and thus poses a critical concern to inferential tools, as these are often based on the presence or absence of correlation in the gene expression profiles. Different research efforts have aimed at investigating the phenomenon [264–268, 272].

In the study described in Chapter 4, we investigated the relationship between reporter-to-transcript sequence similarity and correlation of expression signals in different experimental datasets of the ATH1 GeneChip for *Arabidopsis thaliana*. We assessed the extent to which inclusion of off-target reporters in probe sets influences this correlation, and investigated the relation between expression correlation and reporter-to-transcript sequence alignment strength. To this end, we developed a custom-made probe set composition and annotation in which reporters are uniquely assigned to *Arabidopsis* transcripts, according to strict rules of transcript complementarity.

Our analysis showed that numerous probe sets on this widely used commercial array platform contain off-target reporters, and many show a signal pattern that is highly similar to that of unintended transcripts. In addition, a positive correlation was revealed between off-target alignment strength and the magnitude of the correlation with their off-target. Taken together this means that probe sets that contain reporters that align well to off-target genes show correlated intensity values to these other transcripts. As the positive trend can be observed even

between gene pairs that do not share longer stretches of sequence similarity but where the reporter to off-target alignment is only based on short near-matches and because the effect can be observed within probe sets we suggest that this positive relationship is likely not due to functional relatedness of the genes, but to a cross-hybridisation artifact.

We demonstrated that omitting reporters liable to cross-hybridisation results in decreased artifactual correlation coefficients between probe sets and thus conclude that careful reporter mapping alleviates cross-hybridisation effects to a substantial extent. Furthermore, we described a novel method for diagnosing individual probesets that are likely affected by off-target hybridisation.

In summary, our analysis represents a significant advance to analyses that rely on the absence or presence of expression correlation. Cross-Hybridisation of off-target transcripts to reporters is a serious concern that should be taken into account and can be alleviated by accurate mapping between microarray reporter and the target transcriptome.

7

Nederlandstalige Samenvatting –Summary in Dutch–

Bijgestaan door belangrijke technologische innovaties heeft de komst van verschillende genoomsequenties aan het begin van deze eeuw verscheidene grote doorbraken in de biologische wetenschappen met zich meegebracht. De beschikbaarheid van verschillende soorten genoom-wijde biologische data sets heeft het mogelijk gemaakt dat genetisch onderzoek kon evolueren van studies waarin het gen centraal staat naar grootschalige benaderingen, waarin de rol van elk individueel gen in de totaliteit van de functionerende cel blootgelegd wordt. Wetenschappers kregen instrumenten ter beschikking die hen in staat stelden assumpties te verifiëren en hypotheses te testen die eerder opgesteld waren aan de hand van experimentele studies.

Microarrays vormen een prominent voorbeeld van een toen ontwikkelde techniek, die het simultaan bestuderen van de expressie van een groot aantal genen toelaat [102, 103]. De technologie heeft belangrijke bevindingen mogelijk gemaakt, onder andere in het onderzoek van gedupliceerde genen. Afgeleid van cytologische experimenten verschillende decennia geleden, en bevestigd door recente sequenceringsprojecten, is de aanwezigheid van grote aantallen redundante kopijen een kenmerkende eigenschap van eukaryote genomen. Sinds decennia geniet gen duplicatie de erkenning als van primordiaal belang bij evolutionaire transitie en bij de aangroei in complexiteit van organismen. Gen duplicatie is bijzonder veel voorkomend in planten, waaronder het model organisme

Arabidopsis thaliana, waarvan aangetoond werd dat het verscheidende rondes van grootschalige, genoom-wijde duplicaties en talrijke kleinschalige, lokale duplicaties ondergaan heeft in de loop van zijn evolutie [99–101, 149].

De studie van gen en genoom duplicatiegebeurtenissen, evenals die van de daaropvolgende functionele divergentie van genen, is fundamenteel voor het begrijpen van evolutie en adaptatie van organismen. Twee studies uitgevoerd binnen onze onderzoeksgroep vullen elkaar in dat opzicht aan: in 2005 [83] toonden we aan dat genretentie na duplicatiegebeurtenissen gebiased is naar duplicatie mechanisme en de functie van het gen. De daaropvolgende studie, beschreven in Hoofdstuk 2 van dit proefschrift, benadrukt het belang van deze eigenschappen op de snelheid van expressie divergentie. We pasten de microarray technologie toe om de evolutie en functionele divergentie van gedupliceerde genen in *Arabidopsis thaliana* te analyseren [257]. Ons onderzoek toonde aan dat genen die ontstaan zijn door grootschalige gebeurtenissen relatief trager divergeren dan diegene die ontstaan zijn door kleinschalige, lokale duplicaties. Deze laatste vertonen daarenboven de neiging asymmetrisch te divergeren, in die zin dat één van beide geëxprimeerd wordt in een groot aantal weefsels, terwijl de andere slechts in een klein aantal weefsels tot expressie komt. Onze data geven ook aan dat de functie van een gen en het biologisch proces waarin het speelt een substantieel effect hebben op de divergentie snelheid.

In Hoofdstuk 3 werd microarray data analyse gecombineerd met comparative genomanalyse benaderingen ten einde nieuwe *cis*-regulatorische elementen en hun hogere orde combinaties, modules, te identificeren [314]. Idealiter worden zulke elementen geïdentificeerd met methodes die bepaling toelaten van de bindingsplaatsen voor transcriptiefactoren op het genoom, zoals ChIP-chip experimenten [320]. In de afwezigheid van zulke data op grote schaal kan de detectie uitgevoerd worden op sets van gecoëxprimeerde genen, daar deze waarschijnlijk door eenzelfde transcriptiefactor gereguleerd worden [311–313]. Wij hebben voordeel gehaald uit de toenemende hoeveelheid microarray data die beschikbaar werd voor *Arabidopsis* en uit het beschikbaar worden van de genoomsequentie van een gerelateerd species, *Populus trichocarpa*. In totaal hebben we 80 transcriptiefactor bindingsplaatsen en 139 regulatorische modules geïdentificeerd, waarvan de meeste voordien ongekend waren. Deze modules bestaan hoofdzakelijk uit twee à drie regulatorische elementen die gelinkt konden worden aan verschillende belangrijke biologische processen, zoals eiwitsynthese, controle van de cel cyclus, fotosynthese en embryologische ontwikkeling. Daarenboven heeft de studie van de fysische eigenschappen van sommige regulatorische modules aangetoond dat de promotors van *Arabidopsis* een compacte structuur hebben, waarbij samenwerkende transcriptiefactor bindingsplaatsen in elkaars

nabijheid ondergebracht zijn.

Expressie correlatie relaties, gemeten met behulp van microarray data worden in systeembioïogie frequent gebruikt voor deductie van functionele modules en regulatorische netwerken. Veel downstream analyse tools zijn gebaseerd op de aan- of afwezigheid van correlatie tussen expressie patronen van genen, zoals bij de deductie van coëxpressie [247–251], gen regulatorische [252] en Bayesiaanse netwerken [253–256], en de studie van de evolutie van gen families [130, 257]. Vanuit biologisch standpunt zijn deze benaderingen informatief en zinvol, maar onze analyse in Hoofdstuk 4 van dit proefschrift toont aan dat indien geen voorzorgen getroffen worden bij het berekenen van deze correlaties en bij het toewijzen van de probes aan de transcripten, onjuiste conclusies getrokken kunnen worden.

De microarray technologie confronteert wetenschappers met uiteenlopende uitdagingen. Onze kennis van transcriptomen is onvolledig en onze ramingen van welke transcripten in genomen leven evolueren continu. Daarom is het bij de analyse van microarray data belangrijk na te gaan of een probe werkelijk de expressie meet van het gen waaraan hij toegewezen werd toen de array werd ontwikkeld. Een andere bekommernis is cross-hybridisatie, waarbij transcripten, andere dan de geambieerde, op een probe hybridiseren. Het bekomen signaal van een zulke probe zal dat van een combinatie van verscheidene transcripten zijn.

Met behulp van de veelgebruikte Affymetrix GeneChip voor *Arabidopsis*, hebben we de relatie bestudeerd tussen signaal correlatie en probe-tot-transcript sequentie alignment sterkte. Daarnaast werd de invloed gemeten van de inclusie van een off-target probe in probe sets, i.e. een probe die niet enkel aligneert aan het toegewezen maar eveneens aan andere transcripten, op deze correlatie. Het traditioneel probe set ontwerp, zoals gedefinieerd door de fabrikant van de microarray, werd hiertoe vergeleken met een zelf gedefinieerde probe set annotatie, waarbij probes slechts aan één enkel transcript van *Arabidopsis* toegewezen werden, dit volgens strikte regels van off-target transcript complementariteit. Met betrekking tot de probe set compositie en annotatie, toonden onze data aan dat talrijke probe sets op dit veelgebruikte commercieel array platform off-target probes bevatten en dat veel van die probe sets een signaal patroon vertonen dat hoogst gelijkaardig is aan dat van die off-target transcripten. Met behulp van onze zelf gedefinieerde probe set annotatie toonden we aan dat het weglaten van probes die vatbaar zijn voor cross-hybridisatie resulteert in gedaalde kunstmatige correlatie coëfficiënten tussen probe sets. Met betrekking tot cross-hybridisatie werd een positieve relatie aangetoond tussen off-target alignment sterkte van probes en de omvang van signaal correlatie met die off-target. Tezamen betekent

dit dat probe sets die probes bevatten die goed aligneren aan off-target transcripten sterk gecorreleerde signaal patronen vertonen aan deze andere transcripten. Daar deze positieve trend daarenboven geobserveerd kan worden tussen genparen die geen grote stukken van sequentie similariteit vertonen, maar waarbij de probe tot off-target alignement enkel gebaseerd is op korte bijna-identieke sequenties, en aangezien het effect zichtbaar is binnen probe sets, suggereren wij dat dit positief verband niet te wijten is aan functionele verwantschap van de genen, dan wel aan een cross-hybridisatie artefact.

Onze analyse vertegenwoordigt een significante vooruitgang voor studies die gebaseerd zijn op de aan- of afwezigheid van expressie correlatie. Cross-hybridisatie van off-target transcripten op probes is een ernstige complicatie die in rekening gebracht moet worden en die opgeheven kan worden door nauwkeurig mappen van microarray probes op het doelwit transcriptoom. Daarenboven hebben wij in onze studie een nieuwe methode beschreven voor het diagnoserende van individuele probe sets die mogelijk vatbaar zijn voor off-target hybridisatie.

8

English Summary

Together with major technological advances, the accumulation of genome sequences at the beginning of this century brought about various exciting breakthroughs in life sciences. The availability of different types of genome-wide biological data sets enabled genetic research to shift from gene-focused studies to exhaustive systems biology approaches, in which the role that each individual gene plays in the greater picture of the functioning cell is revealed. Scientists were handed the tools to assess and verify assumptions and hypotheses that previously had been inferred from observational studies.

Microarrays constitute a prominent example of a platform that emerged to facilitate the expression profiling of a large number of genes simultaneously [102, 103]. Among other fields, this technology enabled significant advances in the study of duplicated genes. Inferred from cytological experiments many decades ago, and confirmed by recent genome sequencing projects, the high prevalence of redundant gene copies is a hallmark of eukaryotic genomes. Since decades, gene duplication had been granted acknowledgement as of paramount importance for evolutionary transitions and increases in organismal complexity. Gene duplication has been particularly prevalent in plants, including in the model system *Arabidopsis thaliana*, which has been shown to have undergone several rounds of large-scale, whole-genome duplication and many smaller, local duplication events [99–101, 149].

Investigating gene and genome duplication events as well as the subsequent

functional divergence of genes is of fundamental importance in the understanding of evolution and adaptation of organisms. Two studies conducted by our research group complement each other in this respect: in 2005 [83] we provided proof for a bias in gene retention after duplication events towards the duplication mode and a gene's functionality. The subsequent study, described in Chapter 2 of this dissertation, uncovered the importance of these characteristics on the expression divergence rate. To this end, we applied the microarray technology to analyse the evolution and functional divergence of duplicates in *Arabidopsis thaliana* [257]. We conducted a genome-wide investigation of expression divergence of duplicated genes with respect to time and mode of duplication and their encoded functions. Our study revealed that genes that arose via large-scale events diverge relatively slower than those that got duplicated through local small-scale events. Moreover, the latter tend to diverge asymmetrically, in the sense that one member of the pair is expressed in a large number of tissues while the other is expressed in few tissues. Our data also revealed that a gene's function, or the biological process it is involved in, also have a substantial effect the expression divergence rate.

In Chapter 3 of this dissertation, we combined microarray data analysis with a comparative genomics approach for the discovery of novel *cis*-regulatory elements and their higher-order combinations, modules [314]. Ideally, these elements are identified with a method that allows the determination of the locations to which transcription factors bind, like ChIP-chip assays [320]. In the absence of such systematic data, the detection can be carried out on sets of co-expressed genes, as these are likely to be regulated by the same transcription factor [311–313]. We took advantage of the accumulation of microarray data for *Arabidopsis thaliana* and the availability of the genomic sequence of a related species, *Populus trichocarpa*. In total, we identified 80 transcription factor binding sites and 139 regulatory modules, most of which are novel. These modules consist primarily of two or three regulatory elements that could be linked to different important biological processes, such as protein biosynthesis, cell cycle control, photosynthesis and embryonic development. Moreover, study of the physical properties of some specific regulatory modules revealed that *Arabidopsis* promoters have a compact nature, with cooperative transcription factor binding sites located in close proximity of each other.

Expression correlation relationships measured with microarray data are often used in systems biology to infer functional modules and regulatory networks. Many of the downstream analysis tools are based on the presence, or absence, of correlation in the expression profiles of genes, like the inference of co-expression [247–251], gene regulatory [252] and Bayesian networks [253–256]

and the study of gene family evolution [130, 257]. From a biological point of view, these approaches are useful and informative, but with the analysis presented in Chapter 4 of this dissertation, we show that if care has not been taken as to how these correlations are calculated and how the reporters for each transcript are selected, incorrect conclusions can be drawn.

The microarray technology confronts researchers with various challenges. Our understanding of transcriptomes is incomplete, and our estimates of which transcripts exist in a genome are constantly evolving. Therefore, for the analysis of microarray data it is important to ascertain that a reporter does in fact measure the transcript it was intended to target when the array was designed. Another concern is cross-hybridisation, where transcripts, other than the ones intended, hybridise to a reporter. The signal that is obtained for such a reporter will be that of a combination of multiple different transcripts.

Using the Affymetrix GeneChip for *Arabidopsis*, we investigated the relation between signal correlation and reporter-to-transcript sequence alignment strength and assessed the extent to which inclusion of off-target reporters in probe sets, i.e. reporters that align not only to their intended transcript but also to other transcripts, influences this correlation. The conventional probe set design, as defined by the manufacturer of the microarray, was compared to a custom-made probe set annotation, in which reporters are uniquely assigned to *Arabidopsis* transcripts, according to strict rules of transcript complementarity. With respect to probe set composition and annotation, our data revealed that numerous probe sets on this widely used commercial array platform contain off-target reporters, and that many show a signal pattern that is highly similar to that of unintended transcripts. With our custom-made probe set definition we demonstrated that omitting reporters liable to cross-hybridisation results in decreased artifactual correlation coefficients between probe sets. With respect to cross-hybridisation, a positive correlation was revealed between off-target alignment strength of reporters and the magnitude of the signal correlation to their off-target. Taken together this means that probe sets that contain reporters that align well to off-target transcripts show correlated signal patterns to these other transcripts. As the positive trend can be observed even between gene pairs that do not share longer stretches of sequence similarity but where the reporter to off-target alignment is only based on short near-matches, and because the effect can be observed within probe sets, we suggest that this positive relationship is likely not due to functional relatedness of the genes, but to a cross-hybridization artifact.

Our analysis represents a significant advance to analyses that rely on the absence or presence of expression correlation. Cross-hybridisation of off-target

transcripts to reporters is a serious concern that should be taken into account and can be alleviated by accurate mapping between microarray reporter and the target transcriptome. Furthermore, we described a novel method for diagnosing individual probesets that are likely affected by off-target hybridisation.

References

- [1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409(6822):860–921.
- [2] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, and Merrick JM. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 1995, 269(5223):496–512.
- [3] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. Life with 6000 genes. *Science*, 1996, 274(5287):563–567.
- [4] Blattner F, Plunkett 3rd G, Bloch C, Perna N, Burland V, Riley M, Collado-Vides J, Glasner J, Rode C, Mayhew G, et al. The complete genome sequence of *Escherichia coli* k-12. *Science*, 1997, 277(5331):1453–74.
- [5] The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 1998, 282(5396):2012–2018.
- [6] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 2000, 287(5461):2185–2195.
- [7] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, 408(6814):796–815.
- [8] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002, 420(6915):520–562.
- [9] Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 2004, 428(6982):493–521.

-
- [10] Blanc G, Barakat A, Guyot R, Cooke R, and Delseny M. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell*, 2000, 12(7):1093–101.
- [11] Venter J, Adams M, Myers E, Li P, Mural R, Sutton G, Smith H, Yandell M, Evans C, Holt R, et al. The sequence of the human genome. *Science*, 2001, 291(5507):1304–51.
- [12] Bailey J, Yavor A, Massa H, Trask B, and Eichler E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, 2001, 11(6):1005–17.
- [13] Friedman R and Hughes A. Pattern and timing of gene duplication in animal genomes. *Genome Res.*, 2001, 11(11):1842–7.
- [14] Taylor J, Van de Peer Y, Braasch I, and Meyer A. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. B-Biol. Sci.*, 2001, 356(1414):1661–79.
- [15] McLysaght A, Hokamp K, and Wolfe K. Extensive genomic duplication during early chordate evolution. *Nature Genet.*, 2002, 31(2):200–4.
- [16] Lynch M and Conery J. The origins of genome complexity. *Science*, 2003, 302(5649):1401–4.
- [17] Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvglise C, Talla E, et al. Genome evolution in yeasts. *Nature*, 2004, 430(6995):35–44.
- [18] Koonin E, Fedorova N, Jackson J, Jacobs A, Krylov D, Makarova K, Mazumder R, Mekhedov S, Nikolskaya A, Rao B, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, 2004, 5(2):R7.
- [19] Taylor JS and Raes J. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.*, 2004, 38:615–643.
- [20] Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, et al. Comparative genomics of the eukaryotes. *Science*, 2000, 287(5461):2204–2215.
- [21] Li WH, Gu Z, Wang H, and Nekrutenko A. Evolutionary analyses of the human genome. *Nature*, 2001, 409(6822):847–849.
- [22] Ohno S, Wolf U, and Atkin N. Evolution from fish to mammals by gene duplication. *Hereditas*, 1968, 59(1):169–87.

- [23] Spofford JB. Heterosis and the Evolution of Duplications . *Am. Nat.*, 1969, 103(932):407–432.
- [24] Ohno S. *Evolution by gene duplication*. Springer-Verlag, Berlin, Heidelberg, New York, 1970.
- [25] Kimura M and Ohta T. On Some Principles Governing Molecular Evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 1974, 71(7):2848–2852.
- [26] Ohta T. Multigene families and the evolution of complexity. *J. Mol. Evol.*, 1991, 33(1):34–41.
- [27] Piatigorsky J and Wistow G. The Recruitment of Crystallins: New Functions Precede Gene Duplication. *Science*, 1991, 252:1078–1079.
- [28] Prince VE and Pickett FB. Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.*, 2002, 3(11):827–837.
- [29] Holland P, Garcia-Fernndez J, Williams N, and Sidow A. Gene duplications and the origins of vertebrate development. *Dev. Suppl.*, 1994.
- [30] Garcia-Fernndez J and Holland P. Amphioxus hox genes: insights into evolution and development. *Int. J. Dev. Biol.*, 1996, Suppl 1.
- [31] Gu X, Wang Y, and Gu J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet.*, 2002, 31(2):205–9.
- [32] Panopoulou G, Hennig S, Groth D, Krause A, Poustka A, Herwig R, Vingron M, and Lehrach H. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.*, 2003, 13(6A):1056–66.
- [33] Meyer A and Schartl M. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.*, 1999, 11(6):699–704.
- [34] Taylor JS, Braasch I, Frickey T, Meyer A, and Van de Peer Y. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.*, 2003, 13(3):382–390.
- [35] Vandepoele K, De Vos W, Taylor J, Meyer A, and Van de Peer Y. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. U. S. A.*, 2004, 101(6):1638–43.

-
- [36] Meyer A and Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, 2005, 27(9):937–945.
- [37] Otto S and Whitton J. Polyploid incidence and evolution. *Annu. Rev. Genet.*, 2000, 34.
- [38] Paterson A, Bowers J, Burow M, Draye X, Elisk C, Jiang C, Katsar C, Lan T, Lin Y, Ming R, and Wright R. Comparative genomics of plant chromosomes. *Plant Cell*, 2000, 12(9):1523–40.
- [39] De Bodt S, Maere S, and Van de Peer Y. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.*, 2005, 20(11):591–7.
- [40] Seoighe C. Turning the clock back on ancient genome duplication. *Curr. Opin. Genet. Dev.*, 2003, 13(6):636–43.
- [41] Donoghue P and Purnell M. Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.*, 2005, 20(6):312–9.
- [42] Lawton-Rauh A. Evolutionary dynamics of duplicated genes in plants. *Mol. Phylogenet. Evol.*, 2003, 29(3):396–409.
- [43] Zhang J. Evolution by gene duplication: an update. *Trends Ecol. Evol.*, 2003, 18(6):292–298.
- [44] Hurles M. Gene duplication: the genomic trade in spare parts. *PLoS. Biol.*, 2004, 2(7):E206.
- [45] Gregory T. *The Evolution of the Genome*. Elsevier, San Diego, 2005.
- [46] Bzymek M and Lovett ST. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U. S. A.*, 2001, 98(15):8319–8325.
- [47] Jurka J. Evolutionary impact of human Alu repetitive elements. *Curr. Opin. Genet. Dev.*, 2004, 14(6):603–608.
- [48] Patel N and Prince V. Beyond the hox complex. *Genome Biol.*, 2000, 1(5):REVIEWS1027.
- [49] Eichler E, Hoffman S, Adamson A, Gordon L, McCready P, Lamerdin J, and Mohrenweiser H. Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res.*, 1998, 8(8):791–808.
- [50] Henikoff S, Greene E, Pietrokovski S, Bork P, Attwood T, and Hood L. Gene families: the taxonomy of protein paralogs and chimeras. *Science*, 1997, 278(5338):609–14.

- [51] Lootens S, Burnett J, and Friedman TB. An intraspecific gene duplication polymorphism of the urate oxidase gene of *Drosophila virilis*: a genetic and molecular analysis. *Mol. Biol. Evol.*, 1993, 10(3):635–646.
- [52] Antonell A, de Luis O, Domingo-Roura X, and Perez-Jurado LA. Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. *Genome Res.*, 2005, 15(9):1179–1188.
- [53] Horvath JE, Gulden CL, Vallente RU, Eichler MY, Ventura M, McPherson JD, Graves TA, Wilson RK, Schwartz S, Rocchi M, and Eichler EE. Punctuated duplication seeding events during the evolution of human chromosome 2p11. *Genome Res.*, 2005, 15(7):914–927.
- [54] Liu H, Li L, Zilberstein A, and Hahn CS. Segmental duplications containing tandem repeated genes encoding putative deubiquitinating enzymes. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 2004, pages 31–39.
- [55] Moran JV, DeBerardinis RJ, and Kazazian HHJ. Exon shuffling by L1 retrotransposition. *Science*, 1999, 283(5407):1530–1534.
- [56] Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, and Batzer MA. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U. S. A.*, 2006, 103(47):17608–17613.
- [57] Sakai H, Koyanagi KO, Imanishi T, Itoh T, and Gojobori T. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene*, 2007, 389(2):196–203.
- [58] Jiang N, Bao Z, Zhang X, Eddy SR, and Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, 2004, 431(7008):569–573.
- [59] Samonte RV and Eichler EE. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.*, 2002, 3(1):65–72.
- [60] Kidwell M. *Transposable elements*, pages 165–221. Elsevier, San Diego, 2005.
- [61] Craig N. Unity in transposition reactions. *Science*, 1995, 270(5234):253–4.
- [62] Raizada M, Nan G, and Walbot V. Somatic and germinal mobility of the rescuemu transposon in transgenic maize. *Plant Cell*, 2001, 13(7):1587–608.

-
- [63] Van der Hoeven R, Ronning C, Giovannoni J, Martin G, and Tanksley S. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell*, 2002, 14(7):1441–1456.
- [64] Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, 2002, 296(5565):92–100.
- [65] Leitch LJ and Bennett MD. Polyploidy in angiosperms. *Trends Plant Sci.*, 1997, 2(12):470–476.
- [66] Wendel J and Doyle J. *Polyploidy and evolution in plants*, pages 97–117. CABI Publishing, Wallingford, UK ; Cambridge, MA, 2005.
- [67] Masterson J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, 1994, 264(5157):421 – 424.
- [68] Otto SP and Whitton J. Polyploid incidence and evolution. *Annu. Rev. Genet.*, 2000, 34:401–437.
- [69] Wendel JF. Genome evolution in polyploids. *Plant Mol.Biol.*, 2000, 42(1):225–249.
- [70] Dobzhansky T. Genetics of natural populations. xix. origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. *Genetics*, 1950, 35(3):288–302.
- [71] Marsischky G, Filosi N, Kane M, and Kolodner R. Redundancy of *Saccharomyces cerevisiae* msh3 and msh6 in msh2-dependent mismatch repair. *Genes Dev.*, 1996, 10(4):407–20.
- [72] Wagner A. Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. *Biol. Cybern.*, 1996, 74(6):557–67.
- [73] Gibson T and Spring J. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.*, 1998, 14(2):46–9; discussion 49–50.
- [74] Gu Z, Steinmetz L, Gu X, Scharfe C, Davis R, and Li W. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 2003, 421(6918):63–6.

- [75] Fowler NL and Levin DA. Ecological Constraints on the Establishment of a Novel Polyploid in Competition with Its Diploid Progenitor . *Am. Nat.*, 1984, 124(5):703–711.
- [76] Mayer VW and Aguilera A. High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mutat. Res.*, 1990, 231(2):177–186.
- [77] Ganem NJ, Storchova Z, and Pellman D. Tetraploidy, aneuploidy and cancer. *Curr. Opin. Genet. Dev.*, 2007, 17(2):157–162.
- [78] Thorpe PH, Gonzalez-Barrera S, and Rothstein R. More is not always better: the genetic constraints of polyploidy. *Trends Genet.*, 2007, 23(6):263–266.
- [79] Boyers S, Diamond M, Lavy G, Russell J, and DeCherney A. The effect of polyploidy on embryo cleavage after in vitro fertilization in humans. *Fertil. Steril.*, 1987, 48(4):624–7.
- [80] Comai L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.*, 2005, 6(11):836–846.
- [81] Takahata N and Maruyama T. Polymorphism and loss of duplicate gene expression: a theoretical study with application of tetraploid fish. *Proc. Natl. Acad. Sci. U. S. A.*, 1979, 76(9):4521–4525.
- [82] Lynch M and Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*, 2000, 290(5494):1151–1155.
- [83] Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, and Van de Peer Y. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, 2005, 102(15):5454–5459.
- [84] Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, Cartry J, Duffraisse M, Holland L, Gronemeyer H, and Laudet V. Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet.*, 2006, 2(7):e102.
- [85] Rodriguez-Trelles F, Tarrío R, and Ayala FJ. Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. *Proc. Natl. Acad. Sci. U. S. A.*, 2003, 100(23):13413–13417.
- [86] Vandenbussche M, Theissen G, Van de Peer Y, and Gerats T. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Res.*, 2003, 31(15):4401–4409.

-
- [87] Hughes AL. The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.*, 1994, 256(1346):119–124.
- [88] Lynch M and Katju V. The altered evolutionary trajectories of gene duplicates. *Trends Genet.*, 2004, 20(11):544–549.
- [89] Hughes AL. *Adaptive evolution of genes and genomes*. Oxford University Press, New York, 1999.
- [90] Walsh JB. How often do duplicated genes evolve new functions? *Genetics*, 1995, 139(1):421–428.
- [91] Wagner A. The fate of duplicated genes: loss or new function? *Bioessays*, 1998, 20(10):785–788.
- [92] Nowak MA, Boerlijst MC, Cooke J, and Smith JM. Evolution of genetic redundancy. *Nature*, 1997, 388(6638):167–171.
- [93] Krakauer DC and Nowak MA. Evolutionary preservation of redundant duplicated genes. *Semin. Cell Dev. Biol.*, 1999, 10(5):555–559.
- [94] Pickett FB and Meeks-Wagner DR. Seeing double: appreciating genetic redundancy. *Plant Cell*, 1995, 7(9):1347–1356.
- [95] Gu X. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet.*, 2003, 19(7):354–356.
- [96] Wagner A. Robustness, evolvability, and neutrality. *FEBS Lett.*, 2005, 579(8):1772–1778.
- [97] Lynch M and Force A. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 2000, 154(1):459–473.
- [98] Meinke DW, Cherry JM, Dean C, Rounsley SD, and Koornneef M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science*, 1998, 282(5389):679–682.
- [99] Vision TJ, Brown DG, and Tanksley SD. The origins of genomic duplications in *Arabidopsis*. *Science*, 2000, 290(5499):2114–2117.
- [100] Bowers JE, Chapman BA, Rong J, and Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 2003, 422(6930):433–438.
- [101] Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, and Van de Peer Y. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.*, 2002, 99(21):13627–13632.

- [102] Huber W, Von Heydebreck A, and Vingron M. *Chapter 6. Analysis of microarray gene expression data*, pages 162–187. Wiley, 2003.
- [103] Draghici S. *Data analysis tools for DNA microarrays*. Chapman & Hall/CRC, London, UK, 2003.
- [104] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genet.*, 2001, 29(4):365–371.
- [105] Lipshutz RJ, Fodor SP, Gingeras TR, and Lockhart DJ. High density synthetic oligonucleotide arrays. *Nature Genet.*, 1999, 21(1 Suppl):20–24.
- [106] Huber W, Irizarry R, and Gentleman R. *Chapter 1. Preprocessing overview*, pages 3–12. Springer, New York, USA, 2005.
- [107] Draghici S. *Chapter 3. Image Processing*, pages 33–59. Chapman & Hall/CRC, London, UK, 2003.
- [108] Bergemann TL, Laws RJ, Quiaoit F, and Zhao LP. A statistically driven approach for image segmentation and signal extraction in cDNA microarrays. *J. Comput. Biol.*, 2004, 11(4):695–713.
- [109] Smyth G, Yang Y, and Speed T. Statistical issues in cdna microarray data analysis. *Methods Mol. Biol.*, 2003, 224.
- [110] Quackenbush J. Microarray data normalization and transformation. *Nature Genet.*, 2002, 32 Suppl:496–501.
- [111] Yang YH and Paquet A. *Chapter 4. Preprocessing Two-Color spotted arrays*, pages 49–69. Springer, New York, USA, 2005.
- [112] Boldstad B, Irizarry R, Gautier L, and W Z. *Chapter 3. Preprocessing high-density oligonucleotide arrays*, pages 11–32. Springer, New York, USA, 2005.
- [113] Draghici S. *Chapter 12. Data pre-processing and normalization*, pages 309–340. Chapman & Hall/CRC, London, UK, 2003.
- [114] Smyth G and Speed T. Normalization of cdna microarray data. *Methods*, 2003, 31(4):265–73.
- [115] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003, 4(2):249–264.

-
- [116] Huber W, von Heydebreck A, Sultmann H, Poustka A, and Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 2002, 18 Suppl 1:96–104.
- [117] Li C and Wong W. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.*, 2001, 98(1):31–36.
- [118] Hubbell E, Liu WM, and Mei R. Robust estimators for expression analysis. *Bioinformatics*, 2002, 18(12):1585–1592.
- [119] *Guide to Probe Logarithmic Intensity Error (PLIER) estimation.*, 2005. http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.
- [120] *Statistical Algorithms Description Document*, 2002. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.
- [121] Naef F, Lim DA, Patil N, and Magnasco MO. From features to expression: High-density oligonucleotide array analysis revisited, 2001. <http://asterion.rockefeller.edu/marcelo/Reprints/30features2expression-pre.pdf>.
- [122] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 2003, 31(4):e15.
- [123] Bolstad BM, Irizarry RA, Astrand M, and Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003, 19(2):185–193.
- [124] Cope LM, Irizarry RA, Jaffee HA, Wu Z, and Speed TP. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 2004, 20(3):323–331.
- [125] Bolstad BM. *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, 2004.
- [126] Tukey JW. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [127] Li WH, Gu Z, Cavalcanti ARO, and Nekrutenko A. Detection of gene duplications and block duplications in eukaryotic genomes. *J. Struct. Funct. Genomics*, 2003, 3(1-4):27–34.
- [128] Van de Peer Y. Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.*, 2004, 5(10):752–763.

- [129] Wolfe KH. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.*, 2001, 2(5):333–341.
- [130] Blanc G and Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell*, 2004, 16(7):1679–1691.
- [131] Force A, Lynch M, Pickett FB, Amores A, Yan YL, and Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 1999, 151(4):1531–1545.
- [132] Serebrowsky AS. Genes *scute* and *achaete* in *Drosophila melanogaster* and a hypothesis of gene divergency. *Compt. Rend. Acad. Sci. URSS*, 1938, 14:77–81.
- [133] Stoltzfus A. On the possibility of constructive neutral evolution. *J. Mol. Evol.*, 1999, 49(2):169–181.
- [134] Nadeau JH and Sankoff D. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 1997, 147(3):1259–1266.
- [135] Haldane JBS. The part played by recurrent mutation in evolution. *Am. Nat.*, 1933, 67(708):5–19.
- [136] Raes J and Van de Peer Y. Gene duplication, the evolution of novel gene functions, and detecting functional divergence of duplicates in silico. *Appl. Bioinformatics*, 2003, 2(2):91–101.
- [137] Robinson-Rechavi M and Laudet V. Evolutionary rates of duplicate genes in fish and mammals. *Mol. Biol. Evol.*, 2001, 18(4):681–683.
- [138] Kondrashov FA, Rogozin IB, Wolf YI, and Koonin EV. Selection in the evolution of gene duplications. *Genome Biol.*, 2002, 3(2):RESEARCH0008.
- [139] Van de Peer Y, Taylor JS, Braasch I, and Meyer A. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.*, 2001, 53(4-5):436–446.
- [140] Cronn RC, Small RL, and Wendel JF. Duplicated genes evolve independently after polyploid formation in cotton. *Proc. Natl. Acad. Sci. U. S. A.*, 1999, 96(25):14406–14411.
- [141] Hughes MK and Hughes AL. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.*, 1993, 10(6):1360–1369.

-
- [142] Gu Z, Nicolae D, Lu HHS, and Li WH. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.*, 2002, 18(12):609–613.
- [143] Makova KD and Li WH. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.*, 2003, 13(7):1638–1645.
- [144] Wagner A. Asymmetric functional divergence of duplicate genes in yeast. *Mol. Biol. Evol.*, 2002, 19(10):1760–1768.
- [145] Adams KL, Cronn R, Percifield R, and Wendel JF. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U. S. A.*, 2003, 100(8):4649–4654.
- [146] Blanc G and Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 2004, 16(7):1667–1678.
- [147] Haberer G, Hindemitt T, Meyers BC, and Mayer KFX. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of *Arabidopsis*. *Plant Physiol.*, 2004, 136(2):3009–3022.
- [148] Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, and dePamphilis CW. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol. Biol. Evol.*, 2006, 23(2):469–478.
- [149] Blanc G, Hokamp K, and Wolfe KH. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.*, 2003, 13(2):137–144.
- [150] Seoighe C and Gehring C. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet.*, 2004, 20(10):461–464.
- [151] Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, and Lohmann JU. A gene expression map of *Arabidopsis thaliana* development. *Nature Genet.*, 2005, 37(5):501–506.
- [152] R: a language and environment for statistical computing. <http://www.R-project.org>.
- [153] Taylor J and Raes J. *Small-scale gene duplications*, pages 289–327. Elsevier, San Diego, 2005.

- [154] Brown KE, Amoils S, Horn JM, Buckle VJ, Higgs DR, Merckenschlager M, and Fisher AG. Expression of alpha- and beta-globin genes occurs within different nuclear domains in haemopoietic cells. *Nat. Cell Biol.*, 2001, 3(6):602–606.
- [155] Williams EJB and Bowles DJ. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res.*, 2004, 14(6):1060–1067.
- [156] Perez-Martin J and de Lorenzo V. Clues and consequences of DNA bending in transcription. *Annu. Rev. Microbiol.*, 1997, 51:593–628.
- [157] Gerasimova TI and Corces VG. Chromatin insulators and boundaries: effects on transcription and nuclear organization. *Annu. Rev. Genet.*, 2001, 35:193–208.
- [158] Mishra R and Karch F. Boundaries that demarcate structural and functional domains of chromatin. *J. Biosci.*, 1999, 24(3):377–399.
- [159] Cockell M and Gasser SM. Nuclear compartments and gene regulation. *Curr. Opin. Genet. Dev.*, 1999, 9(2):199–205.
- [160] Ren XY, Fiers MWEJ, Stiekema WJ, and Nap JP. Local coexpression domains of two to four genes in the genome of *Arabidopsis*. *Plant Physiol.*, 2005, 138(2):923–934.
- [161] Rodin SN, Parkhomchuk DV, and Riggs AD. Epigenetic changes and repositioning determine the evolutionary fate of duplicated genes. *Biochemistry (Mosc)*, 2005, 70(5):559–567.
- [162] Adams KL and Wendel JF. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.*, 2005, 8(2):135–141.
- [163] Adams KL and Wendel JF. Novel patterns of gene expression in polyploid plants. *Trends Genet.*, 2005, 21(10):539–543.
- [164] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 1997, 25(17):3389–3402.
- [165] Thompson JD, Higgins DG, and Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 1994, 22(22):4673–4680.
- [166] Goldman N and Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 1994, 11(5):725–736.

-
- [167] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, 1997, 13(5):555–556.
- [168] The Arabidopsis Information Resource. <http://www.arabidopsis.org/>.
- [169] The Gene Ontology. <http://www.geneontology.org/>.
- [170] Affymetrix. <http://www.affymetrix.com/>.
- [171] Nottingham Arabidopsis Stock Centre. <http://affymetrix.arabidopsis.info/>.
- [172] Craigon DJ, James N, Okyere J, Higgins J, Jotham J, and May S. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.*, 2004, 32(Database issue):575–577.
- [173] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 2004, 5(10):R80.
- [174] Gautier L, Cope L, Bolstad BM, and Irizarry RA. affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 2004, 20(3):307–315.
- [175] Liu Wm, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho Mh, Baid J, and Smeekens SP. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, 2002, 18(12):1593–1599.
- [176] MicroArray Suite user guide. <http://www.affymetrix.com/support/technical/manuals.affx>.
- [177] Loader C. *Local Regression and Likelihood*. Springer, New York, 1999.
- [178] Bioinformatics and Evolutionary Genomics: Supplementary Data. http://bioinformatics.psb.ugent.be/supplementary_data/.
- [179] Venter M and Botha F. Promoter analysis and transcription profiling: Integration of genetic data enhances understanding of gene expression. *Physiol. Plant*, 2004, 120(1):74–83.
- [180] Wellmer F and Riechmann J. Gene network analysis in plant development by genomic technologies. *Int. J. Dev. Biol.*, 2005, 49(5-6):745–759.

- [181] Chaboute ME, Clement B, and Philipps G. S phase and meristem-specific expression of the tobacco RNR1b gene is mediated by an E2F element located in the 5' leader sequence. *J. Biol. Chem.*, 2002, 277(20):17845–17851.
- [182] Hong RL, Hamaguchi L, Busch MA, and Weigel D. Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. *Plant Cell*, 2003, 15(6):1296–1309.
- [183] Babu MM, Luscombe NM, Aravind L, Gerstein M, and Teichmann SA. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, 2004, 14(3):283–291.
- [184] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 2004, 431(7004):99–104.
- [185] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM. Systematic determination of genetic network architecture. *Nature Genet.*, 1999, 22(3):281–285.
- [186] Bussemaker HJ, Li H, and Siggia ED. Regulatory element detection using correlation with expression. *Nature Genet.*, 2001, 27(2):167–171.
- [187] Pilpel Y, Sudarsanam P, and Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, 2001, 29(2):153–159.
- [188] Wasserman WW and Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 2004, 5(4):276–287.
- [189] Li H and Wang W. Dissecting the transcription networks of a cell using computational genomics. *Curr. Opin. Genet. Dev.*, 2003, 13(6):611–616.
- [190] Siggia ED. Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.*, 2005, 15(2):214–221.
- [191] Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, and Freeling M. Conserved noncoding sequences in the grasses. *Genome Res.*, 2003, 13(9):2030–2041.
- [192] Vandepoele K, Simillion C, and Van de Peer Y. The quest for genomic homology. *Curr. Genomics*, 2004, 5:299–308.
- [193] Chang LW, Nagarajan R, Magee JA, Milbrandt J, and Stormo GD. A systematic model to predict transcriptional regulatory mechanisms based

-
- on overrepresentation of transcription factor binding profiles. *Genome Res.*, 2006, 16(3):405–413.
- [194] Kellis M, Patterson N, Endrizzi M, Birren B, and Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 2003, 423(6937):241–254.
- [195] Kreiman G. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res.*, 2004, 32(9):2889–2900.
- [196] Wang T and Stormo GD. Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc. Natl. Acad. Sci. U. S. A.*, 2005, 102(48):17400–17405.
- [197] van Noort V and Huynen MA. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.*, 2006, 22(2):73–78.
- [198] Van Hellefont R, Monsieurs P, Thijs G, de Moor B, Van de Peer Y, and Marchal K. A novel approach to identifying regulatory motifs in distantly related genomes. *Genome Biol.*, 2005, 6(13):R113.
- [199] Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. The genome of black cottonwood, *Populus trichocarpa* (torr. & gray). *Science*, 2006, 313(5793):1596–1604.
- [200] Pritsker M, Liu YC, Beer MA, and Tavazoie S. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.*, 2004, 14(1):99–108.
- [201] Elemento O and Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, 2005, 6(2):R18.
- [202] Zhou Q and Wong WH. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. U. S. A.*, 2004, 101(33):12114–12119.
- [203] Higo K, Ugawa Y, Iwamoto M, and Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, 1999, 27(1):297–300.
- [204] Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, and Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.*, 2002, 30(1):325–327.

- [205] Tremousaygue D, Manevski A, Bardet C, Lescure N, and Lescure B. Plant interstitial telomere motifs participate in the control of gene expression in root meristems. *Plant J.*, 1999, 20(5):553–561.
- [206] De Veylder L, Joubes J, and Inze D. Plant cell cycle transitions. *Curr. Opin. Plant Biol.*, 2003, 6(6):536–543.
- [207] Tatematsu K, Ward S, Leyser O, Kamiya Y, and Nambara E. Identification of cis-elements that regulate gene expression during initiation of axillary bud outgrowth in *Arabidopsis*. *Plant Physiol.*, 2005, 138(2):757–766.
- [208] Weisshaar B, Armstrong GA, Block A, da Costa e Silva O, and Hahlbrock K. Light-inducible and constitutively expressed DNA-binding proteins recognizing a plant promoter element with functional relevance in light responsiveness. *EMBO J.*, 1991, 10(7):1777–1786.
- [209] Bennetzen JL. Comparative sequence analysis of plant nuclear genomes:microcolinearity and its many exceptions. *Plant Cell*, 2000, 12(7):1021–1029.
- [210] Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GTS, Gruissem W, Van de Peer Y, Inze D, and De Veylder L. Genome-wide identification of potential plant E2F target genes. *Plant Physiol.*, 2005, 139(1):316–328.
- [211] Carranco R, Almoguera C, and Jordano J. A plant small heat shock protein gene expressed during zygotic embryogenesis but noninducible by heat stress. *J. Biol. Chem.*, 1997, 272(43):27470–27475.
- [212] Arguello-Astorga GR and Herrera-Estrella LR. Ancestral multipartite units in light-responsive plant promoters have structural features correlating with specific phototransduction pathways. *Plant Physiol.*, 1996, 112(3):1151–1166.
- [213] Storey JD and Tibshirani R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, 2003, 100(16):9440–9445.
- [214] Li Z and Thomas TL. PEI1, an embryo-specific zinc finger protein gene required for heart-stage embryo formation in *Arabidopsis*. *Plant Cell*, 1998, 10(3):383–398.
- [215] Takada S, Hibara K, Ishida T, and Tasaka M. The CUP-SHAPED COTYLEDON1 gene of *Arabidopsis* regulates shoot apical meristem formation. *Development*, 2001, 128(7):1127–1135.
- [216] Menges M, Hennig L, Gruissem W, and Murray JAH. Genome-wide gene expression in an *Arabidopsis* cell suspension. *Plant Mol.Biol.*, 2003, 53(4):423–442.

-
- [217] Li C, Potuschak T, Colon-Carmona A, Gutierrez RA, and Doerner P. ,textitArabidopsis TCP20 links regulation of growth and cell division control pathways. *Proc. Natl. Acad. Sci. U. S. A.*, 2005, 102(36):12978–12983.
- [218] Suarez-Lopez P, Wheatley K, Robson F, Onouchi H, Valverde F, and Coupland G. CONSTANS mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature*, 2001, 410(6832):1116–1120.
- [219] Matsushika A, Makino S, Kojima M, and Mizuno T. Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in *Arabidopsis thaliana*: insight into the plant circadian clock. *Plant Cell Physiol.*, 2000, 41(9):1002–1012.
- [220] Hudson ME and Quail PH. Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiol.*, 2003, 133(4):1605–1616.
- [221] Sudarsanam P, Pilpel Y, and Church GM. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res.*, 2002, 12(11):1723–1731.
- [222] Johnson DS, Zhou Q, Yagi K, Satoh N, Wong W, and Sidow A. De novo discovery of a tissue-specific gene regulatory module in a chordate. *Genome Res.*, 2005, 15(10):1315–1324.
- [223] Wang T and Stormo GD. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 2003, 19(18):2369–2380.
- [224] Grad YH, Roth FP, Halfon MS, and Church GM. Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D.pseudoobscura*. *Bioinformatics*, 2004, 20(16):2738–2750.
- [225] Sinha S, Blanchette M, and Tompa M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 2004, 5:170.
- [226] Monsieurs P, Thijs G, Fadda AA, De Keersmaecker SCJ, Vanderleyden J, De Moor B, and Marchal K. More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics*, 2006, 7:160.

- [227] Aerts S, Van Loo P, Moreau Y, and De Moor B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, 2004, 20(12):1974–1976.
- [228] Gupta M and Liu JS. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.*, 2005, 102(20):7079–7084.
- [229] Paterson AH. Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat. Rev. Genet.*, 2006, 7(3):174–184.
- [230] Ben-Dor A, Shamir R, and Yakhini Z. Clustering gene expression patterns. *J. Comput. Biol.*, 1999, 6(3-4):281–297.
- [231] Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, and Moreau Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, 2002, 9(2):447–464.
- [232] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, and Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 2005, 434(7031):338–345.
- [233] Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, Rouze P, De Moor B, and Marchal K. INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, 2002, 18(2):331–332.
- [234] Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, and Van de Peer Y. EST data suggest that poplar is an ancient polyploid. *New Phytol.*, 2005, 167(1):165–170.
- [235] De Bodt S, Theissen G, and Van de Peer Y. Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.*, 2006, 23(6):1293–1303.
- [236] Frazer KA, Elnitski L, Church DM, Dubchak I, and Hardison RC. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.*, 2003, 13(1):1–12.
- [237] Rost B. Twilight zone of protein sequence alignments. *Protein Eng.*, 1999, 12(2):85–94.

-
- [238] Bioinformatics and evolutionary genomics: Genomes. <http://bioinformatics.psb.ugent.be/genomes.php>.
- [239] Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, et al. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.*, 2006, 103(31):11647–11652.
- [240] DOE joint genome institute. <http://www.jgi.doe.gov/>.
- [241] Vandepoele K and Van de Peer Y. Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol.*, 2005, 137(1):31–42.
- [242] Schiex T, Moisan A, and Rouze P. Eugène: an eukaryotic gene finder that combines several sources of evidence. *Computational Biology: Selected Papers (Lecture Notes in Computer Science)*, 2001, Edited by: Gascuel O, Sagot M-F. Berlin: Springer-Verlag, 2006:111–125.
- [243] Notredame C, Higgins DG, and Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Evol.*, 2000, 302(1):205–217.
- [244] Felsenstein J. Phylogeny Inference Package (version 3.2). *Cladistics*, 1989, 5:164–166.
- [245] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, 2006, 34(Database issue):322–326.
- [246] The TIGR *Arabidopsis thaliana* Database. <http://www.tigr.org/tdb/e2k1/ath1/>.
- [247] Gutierrez RA, Lejay LV, Dean A, Chiaromonte F, Shasha DE, and Coruzzi GM. Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in *Arabidopsis*. *Genome Biol.*, 2007, 8(1):R7.
- [248] Wille A, Zimmermann P, Vranova E, Furholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, et al. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.*, 2004, 5(11):R92.
- [249] Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, and Loraine A. Transcriptional coordination of the metabolic network in *Arabidopsis*. *Plant Physiol.*, 2006, 142(2):762–774.

- [250] Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, and Benfey PN. A gene expression map of the *Arabidopsis* root. *Science*, 2003, 302(5652):1956–1960.
- [251] Williams EJ and Bowles DJ. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res.*, 2004, 14(6):1060–1067.
- [252] Chen G, Jensen ST, and Stoeckert CJJ. Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, 2007, 8(1):R4.
- [253] Friedman N, Linial M, Nachman I, and Peér D. Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, 2000, 7(3-4):601–620.
- [254] Husmeier D. Reverse engineering of genetic networks with Bayesian networks. *Biochem. Soc. Trans.*, 2003, 31:1516–1518.
- [255] Werhli AV, Grzegorzczak M, and Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 2006, 22:2523–2531.
- [256] Schafer J and Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 2005, 21:754–764.
- [257] Casneuf T, De Bodt S, Raes J, Maere S, and Van de Peer Y. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol.*, 2006, 7(2):R13.
- [258] *GeneChip® Expression Analysis Data Analysis Fundamentals*, 2006. http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf.
- [259] Transcript assignment for netaffx annotations, 2006. http://www.affymetrix.com/support/technical/manual/alignments_psl_manual.affx.
- [260] Roche FM, Hokamp K, Acab M, Babiuk LA, Hancock REW, and Brinkman FSL. ProbeLynx: a tool for updating the association of microarray probes to genes. *Nucleic Acids Res.*, 2004, 32(Web Server issue):471–474.
- [261] Talla E, Tekaiia F, Brino L, and Dujon B. A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization. *BMC Genomics*, 2003, 4(1):38.
- [262] Zhang J, Finney RP, Clifford RJ, Derr LK, and Buetow KH. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics*, 2005, 85(3):297–308.

-
- [263] Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, 2005, 33(20):e175–e175.
- [264] Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen MM, Lu G, Fang J, Liu WM, Ryder T, Kaplan P, Kulp D, and Webster TA. Probe selection for high-density oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.*, 2003, 100(20):11237–11242.
- [265] Wu C, Carta R, and Zhang L. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res.*, 2005, 33(9):e84.
- [266] Huang JC, Morris QD, Hughes TR, and Frey BJ. GenXHC: a probabilistic generative model for cross-hybridization compensation in high-density genome-wide microarray data. *Bioinformatics*, 2005, 21 Suppl 1:222–231.
- [267] Chen YA, Chou CC, Lu X, Slate EH, Peck K, Xu W, Voit EO, and Almeida JS. A multivariate prediction model for microarray cross-hybridization. *BMC Bioinformatics*, 2006, 7:101.
- [268] Flikka K, Yadetie F, Laegreid A, and Jonassen I. XHM: a system for detection of potential cross hybridizations in DNA microarrays. *BMC Bioinformatics*, 2004, 5:117.
- [269] Eklund AC, Turner LR, Chen P, Jensen RV, deFeo G, Kopf-Sill AR, and Szallasi Z. Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat. Biotechnol.*, 2006, 24(9):1071–1073.
- [270] Plutowski U and Richert C. A direct glimpse of cross-hybridization: background-passified microarrays that allow mass-spectrometric detection of captured oligonucleotides. *Angew. Chem. Int. Ed. Engl.*, 2005, 44(4):621–625.
- [271] Wren JD, Kulkarni A, Joslin J, Butow RA, and Garner HR. Cross-hybridization on PCR-spotted microarrays. *IEEE Eng. Med. Biol. Mag.*, 2002, 21(2):71–75.
- [272] Okoniewski MJ and Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 2006, 7:276.
- [273] Binder H. Thermodynamics of competitive surface adsorption on dna-microarrays. *J. Phys.: Condensed Matter*, 2006, 18:491–523.

- [274] Aoki K, Ogata Y, and Shibata D. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.*, 2007, 48(3):381–390.
- [275] Rice P, Longden I, and Bleasby A. The european molecular biology open source suite. *Trends Genet.*, 2000, 16(6):276–7.
- [276] Needleman SB and Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Evol.*, 1970, 48:443–453.
- [277] Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, and Lohmann JU. A gene expression map of *Arabidopsis thaliana* development. *Nature Genet.*, 2005, 37(5):501–506.
- [278] Altschul S, Gish W, Miller W, Myers E, and Lipman D. Basic local alignment search tool. *J. Mol. Evol.*, 1990, 215:403–410.
- [279] Rocke DM and Blythe D. A Model for Measurement Error for Gene Expression Arrays. *J. Comput. Biol.*, 2001, 8(6):557–569.
- [280] Li C and Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, 2001, 2(8):1–11.
- [281] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0. <http://www.R-project.org>.
- [282] Slater GSC and Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 2005, 6:31. <http://www.ebi.ac.uk/~guy/exonerate/>.
- [283] Gu X. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics*, 2004, 167(1):531–42.
- [284] Oakley T, Gu Z, Abouheif E, Patel N, and Li W. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Mol. Biol. Evol.*, 2005, 22(1):40–50.
- [285] Guo H, Weiss R, Gu X, and Suchard M. Time squared: Repeated measures on phylogenies. *Mol. Biol. Evol.*, 2006.
- [286] Gu Z, Rifkin S, White K, and Li W. Duplicate genes increase gene expression diversity within and between species. *Nature Genet.*, 2004, 36(6):577–9.

-
- [287] Harvey PH and Pagel MD. *The Comparative Method in Evolutionary Biology*. Oxford University Press, New York, 1991.
- [288] Felsenstein J. Phylogenies and the comparative method. *Am. Nat.*, 1985, 125:1–15.
- [289] Thomas B, Pedersen B, and Freeling M. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, 2006.
- [290] AtGenexpress data. <http://www.weigelworld.org/resources/microarray/AtGenExpress/Sample%20list%20%28Abiotic%20stress%29>.
- [291] Vandepoele K, Saeys Y, Simillion C, Raes J, and Van De Peer Y. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.*, 2002, 12(11):1792–1801.
- [292] Raes J, Vandepoele K, Simillion C, Saeys Y, and Van de Peer Y. Investigating ancient duplication events in the *Arabidopsis* genome. *J. Struct. Funct. Genomics*, 2003, 3(1-4):117–29.
- [293] Freeling M and Thomas B. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.*, 2006, 16(7):805–14.
- [294] Brunet F, Crollius H, Paris M, Aury J, Gibert P, Jaillon O, Laudet V, and Robinson-Rechavi M. Gene loss and evolutionary rates following whole genome duplication in teleost fishes. *Mol. Biol. Evol.*, 2006.
- [295] Chapman B, Bowers J, Feltus F, and Paterson A. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proc. Natl. Acad. Sci. U. S. A.*, 2006.
- [296] Cusack B and Wolfe K. Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol. Biol. Evol.*, 2006.
- [297] Akhunov E, Akhunova A, and Dvorak J. Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. *Mol. Biol. Evol.*, 2006.
- [298] Wapinski I, Pfeffer A, Friedman N, and Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 2007, 449(7158):54–61.

- [299] He X and Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, 2005, 169(2):1157–64.
- [300] Gu X. Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.*, 1999, 16(12):1664–74.
- [301] Dermitzakis E and Clark A. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, 2002, 19(7):1114–21.
- [302] Madan Babu M and Teichmann S. Evolution of transcription factors and the gene regulatory network in escherichia coli. *Nucleic Acids Res.*, 2003, 31(4):1234–44.
- [303] Dermitzakis E, Bergman C, and Clark A. Tracing the evolutionary history of drosophila regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.*, 2003, 20(5):703–14.
- [304] Maston G, Evans S, and Green M. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 2006.
- [305] Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, and Brenner S. Detecting conserved regulatory elements with the model genome of the japanese puffer fish, fugu rubripes. *Proc. Natl. Acad. Sci. U. S. A.*, 1995, 92(5):1684–8.
- [306] Santini S, Boore J, and Meyer A. Evolutionary conservation of regulatory elements in vertebrate hox gene clusters. *Genome Res.*, 2003, 13(6A):1111–22.
- [307] Liu Y, Liu X, Wei L, Altman R, and Batzoglou S. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.*, 2004, 14(3):451–8.
- [308] Emberly E, Rajewsky N, and Siggia E. Conservation of regulatory elements between two species of drosophila. *BMC Bioinformatics*, 2003, 4(1):57.
- [309] Hardison R, Oeltjen J, and Miller W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, 1997, 7(10):959–66.
- [310] Hardison R. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, 2000, 16(9):369–72.
- [311] Zhang M. Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.*, 1999, 23(3-4):233–50.

-
- [312] Niehrs C and Pollet N. Synexpression groups in eukaryotes. *Nature*, 1999, 402(6761):483–7.
- [313] Gasch A, Moses A, Chiang D, Fraser H, Berardini M, and Eisen M. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.*, 2004, 2(12):e398.
- [314] Vandepoele K, Casneuf T, and Van de Peer Y. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biol.*, 2006, 7(11):R103.
- [315] Fry C and Peterson C. Transcription. unlocking the gates to gene expression. *Science*, 2002, 295(5561):1847–8.
- [316] Ma Z, Shah R, Chang M, and Benveniste E. Coordination of cell signaling, chromatin remodeling, histone modifications, and regulator recruitment in human matrix metalloproteinase 9 gene transcription. *Mol. Cell. Biol.*, 2004, 24(12):5496–509.
- [317] Brown E, Malakar S, and Krebs J. How many remodelers does it take to make a brain? diverse and cooperative roles of atp-dependent chromatin-remodeling complexes in development. *Biochem. Cell Biol.*, 2007, 85(4):444–62.
- [318] Liolios K, Tavernarakis N, Hugenholtz P, and Kyrpides N. The genomes on line database (gold) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, 2006, 34(Database issue):D332–4.
- [319] Thurman RE, Day N, Noble WS, and Stamatoyannopoulos JA. Identification of higher-order functional domains in the human encode regions. *Genome Res.*, 2007, 17(6):917–927.
- [320] Ren B, Robert F, Wyrick J, Aparicio O, Jennings E, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert T, Wilson C, Bell S, and Young R. Genome-wide location and function of dna binding proteins. *Science*, 2000, 290(5500):2306–9.



Glossary

Allopolyploid	An organism from which the genome came about by the merging of two or more genetically different genomes
Aneuploidy	The occurrence of one or more extra or missing chromosomes leading to an unbalanced chromosome complement
Autopolyploid	An organisms from which the genome came about by the doubling of a complete single genome
Batch effect	Experimental factors that add systematic biases to the measurements of microarrays and that may vary between different subsets or stages of the experiment
cDNA microarray	Microarrays whose spots contain complementary DNA, generated via e.g. PCR amplification
Comparative genomics	The study of the similarities and differences in structure and function of hereditary information across taxa

Duplicative transposition	A mechanism of transposition that results in a copy of the element at both the excision and acceptor site. This can occur even after excision of the element by the process of gap repair
DNA transposon	Transposable elements that do not use a reverse-transcription step to integrate copies into the genome
Epigenetic	A heritable change that is not caused by a genetic mutation
Homologous genes	An all-or-nothing concept describing whether or not, two genes derived from the same gene in a common ancestor
Neo-functionalisation	When one of two duplicate genes acquires a mutation in coding or regulatory sequences that allows the gene to take on a new and for the organism useful function
Non-functionalisation	When one of two duplicate genes acquires a mutation in coding or regulatory sequences that renders it non-functional
Oligo(nucleotide)	A short fragment of a single-stranded DNA that is typically 5 to 50 nucleotides long.
Orthologous genes	Duplicate genes that originated by speciation
Paralogous genes	Duplicate genes that originated by gene duplication
Promoter	The nucleotide sequence upstream of a gene to which RNA polymerase attaches at the beginning of transcription
Purifying selection	Selection against deleterious alleles

Reporter	Or probe, the single stranded DNA sequence that is attached to a microarray
Retrotransposon	Transposable elements that use a reverse-transcription step to integrate copies into the genome; also known as retroposons
Terminal inverted repeat	Repeats that flank most DNA transposons and lie in an inverted orientation
Transposable elements	All mobile DNA segments in the genome, regardless of their mechanism of transposition

B

List of Acronyms

BLAST	Basic Local Alignment Search Tool
BP	Biological Process
CAST	Cluster Affinity Search Technique
CRM	<i>Cis</i> -Regulatory Module
GO	Gene Ontology
K_S	The number of synonymous substitutions per synonymous site
K_A	The number of nonsynonymous substitutions per nonsynonymous site
MF	Molecular Function
NCS	Network-level Conservation Score
PCC	Pearson Correlation Coefficient
PLIER	Probe Logarithmic Intensity Error
PWM	Position Weight Matrix
RMA	Robust Multiple-array Average
TFBS	Transcription Factor Binding Sites
TF	Transcription Factor



Publication List

Tineke Casneuf, Yves Van de Peer & Wolfgang Huber (in press) The Effect of Cross-Hybridization on Expression Correlations Inferred from Microarrays

Klaas Vandepoele, Tineke Casneuf & Yves Van de Peer (2006) Identification of novel regulatory modules in dicot plants using expression data and comparative genomics. *Genome Biology* **7**(11): R103

Tineke Casneuf, Stefanie De Bodt, Jeroen Raes, Steven Maere & Yves Van de Peer (2006) Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biology*. **7**(2): R13

Steven Maere, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper & Yves Van de Peer (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **102**(15): 5454-5459

Modeling gene and genome duplications in eukaryotes


Steven Maere*, Stefanie De Bodt*, Jeroen Raes, Tineke Casneuf,
Marc Van Montagu, Martin Kuiper, and Yves Van de Peer

Proc. Natl. Acad. Sci. U.S.A. **102**(15): 5454-5459

Abstract

Recent analysis of complete eukaryotic genome sequences has revealed that gene duplication has been rampant. Moreover, next to a continuous mode of gene duplication, in many eukaryotic organisms the complete genome has been duplicated in their evolutionary past. Such large-scale gene duplication events have been associated with important evolutionary transitions or major leaps in development and adaptive radiations of species. Here, we present an evolutionary model that simulates the duplication dynamics of genes, considering genome-wide duplication events and a continuous mode of gene duplication. Modeling the evolution of the different functional categories of genes assesses the importance of different duplication events for gene families involved in specific functions or processes. By applying our model to the Arabidopsis genome, for which there is compelling evidence for three whole-genome duplications, we show that gene loss is strikingly different for large-scale and small-scale duplication events and highly biased toward certain functional classes. We provide evidence that some categories of genes were almost exclusively expanded through large-scale gene duplication events. In particular, we show that the three whole-genome duplications in Arabidopsis have been directly responsible for ~90% of the increase in transcription factors, signal transducers, and developmental genes in the last 350 million years. Our evolutionary model is widely applicable and can be used to evaluate different assumptions regarding small- or large-scale gene duplication events in eukaryotic genomes.

*Contributed equally



Microarrays are a valuable source of large-scale and detailed information for functional genomics research. In the past decade their application helped to answer a myriad of scientific questions.

In a first section of this thesis, microarray data are used to study the fate of the numerous duplicated genes in the plant model organism, *Arabidopsis thaliana*. Different questions are addressed, such as how fast do duplicates diverge, does the rate of expression divergence depend on a gene pairs' duplication mechanism or function, and do different types of genes show distinct tissue expression divergence patterns?

In a second part of this thesis, a detection strategy that combines classic motif overrepresentation approaches with general comparative footprinting principles is applied for the identification of novel regulatory motifs in sets of co-expressed genes, delineated by means of microarray data.

Co-expression signatures are an important tool for studying gene functions and relations. In a third section the contribution of genuine biological co-expression and cross-hybridisation in correlated microarray signal profiles is quantified.

The last major part covers a revision of the work presented in the first section, in light of more recent methodological progress with (a) microarray data analysis and the potential pitfalls of cross-hybridisation, presented in section three, and (b) the treatment of the correlation structure within the set of duplicated genes.

The materials presented cover both the application of microarray data in gene expression studies and fundamental research of the use of the microarray technology for correlation analysis.