CAGED: a tool to investigate the relationship between gene expression and genome organization in *Arabidopsis thaliana*

Sachin J. Somers

May 2010

Department of Molecular and Cell Biology
University of Cape Town

Submitted in fulfillments of the requirements for the degree of Master of Science

Supervised by:
Cathal Seoighe

This work was supported by the National Bioinformatics Network (NBN). The views and conclusions in this document are those of the author and should not be interpreted as representing the views of the sponsoring institution.

Abstract

The distribution of genes in eukaryotic genomes is not random. Whole genome array experiments have shown that genes which are similarly expressed may cluster to form domains. These domains are identified by correlating gene expression across multiple experimental conditions. This analysis requires skill and knowledge of microarray highthroughput processing which would be time consuming for a researcher with minimum bioinformatics skills. Hence, various software and tools have been created to assist but even these can prove to be difficult to use. CAGED is an online tool which allows the researcher to visualize co-expressed *Arabidopsis* genes that are located close to one another in the genome. In this way it highlights genomic domains consisting of co-regulated genes The simplicity of its use is demonstrated in examples which highlight CAGED's ability to identify new genomic regions of interest as well as to investigate previously established gene clusters and re-examine them in a genomic context.

In addition to the construction of a tool to visualize gene co-expression in the genomic context, the data underlying the CAGED tool was used to investigate the relationship between physical location of gene pairs and the extent of co-expression. Gene co-expression is thought to be influenced by a number of factors, such as, gene distance, promoter sharing, gene pair orientation and gene function. A small percentage of gene pairs were shown to be highly co-expressed. These were found to occur predominantly in the parallel orientation and were located in very close proximity. We also investigated the co-expression of plastid-derived (mitochondrial and chloroplast) adjacent gene pairs and found that few of these gene pairs were highly co-expressed and lie much closer together than all non-duplicate adjacent gene pairs.

Table of Contents

List of Figures

Chapter 2: CAGED

Chapter 3: Analysis of the relationship between gene co-expression and co-location in *Arabidopsis thaliana*

List of Tables

Chapter 2: CAGED

Chapter 3: Analysis of the relationship between gene co-expression and co-location in *Arabidopsis thaliana*

List of Abbreviations

%	Percent
by	base pairs
CAGED	Chromosomal Analysis of Gene Expression Data
GO	Gene Ontology
ID	Identifier
Kbp	Kilobase pairs
Mb	Megabase pairs
NASC	Nottingham Arabidopsis Stock Center
OSC	Oxidosqualene Cyclase
PCA	Principal Component Analysis
Q-0	Quantile-Quantile
R	Pearson Correlation Co-efficient
RMA	Robust Multiarray Analysis
Std	standard
td	tandem duplicates
THAD	Thaliana-diol desaturase
THAN	Thalianol hydroxylase
THAS	Thalianol synthase

Acknowledgements

I would like to thank my supervisor Cathal for his supervision and guidance. Also I like to thank my colleagues, especially Vicky and Natasha for their bioinformatics assistance during my degree. Thanks to Rodger, our system administrator, for all his help with my server needs and assistance with getting CAGED up on the server. Finally, I would like to thank my family and friends who supported me.

Chapter 1: Background

1.1 Genome Organization of Prokaryotes and Eukaryotes

Various structural arrangements exist in cellular organisms to maintain the genome and its replication. Some functionally related genes were found to cluster together to form operons. Various models have been proposed to describe this behaviour: the Natal Model (Horowitz and Netzenberg, 1965), the Fisher Model (Fisher, 1930), Co-regulation Model (Jacob et al., 1960) and the Selfish Operon Model (Lawrence and Roth, 1996).

The Natal Model states that genes that occur in clusters are a result of gene duplication and divergence (Horowitz and Netzenberg, 1965). This model does apply to some genes such as prokaryotic MetB and MetC genes involved in methionine biosynthesis (Belfaiza et al., 1986) and, in eukaryotes, some globin genes (Maniatis et al., 1980) and HGH genes (Jones et al., 1995). After the 3D alignment of the nucleotide binding site from functionally distant hydrogenases, it was hypothesized that the enzymes originated from a common ancestor (Rossmann et al., 1974). Also, since hydrogenases occur in different operons, it was suggested that operons were assembled from genes that originated separately.

The Fisher Model proposes that clustering can result from an allele at one locus which produces a protein that interacts with a protein produced by a nearby locus. This interaction can result in selection against recombination (Fisher 1930). The model was expanded to include that in bacteria, genes were physically clustered to prevent recombination from disrupting the coadapted alleles (Stahl and Murray, 1966). The problem with this model is that it would require frequent recombination events to occur for the genes to become clustered in the first place.

In the Co-regulation Model, genes are clustered because the sharing of a promoter results in co-ordinated expression which is beneficial to the cell (Jacob et al., 1960). The shortcoming of this model is that in order for this promoter sharing model to exist there would need to be tremendous selection for the positioning of these genes next to each other such that during operon formation there would be some co-regulation of unlinked genes.

Finally, the Selfish Operon Model proposes that genes are organized into clusters which are beneficial to the genes rather than the organism (Lawrence and Roth, 1996). Genes are transferred as clusters either by vertical or horizontal gene transfer. The problem with this model is that it predicts that essential genes do not cluster, which contradicts the fact that there are cases in which essential genes do cluster eg. in *E.coli* (Gerdes et al., 2003). These genes have also been shown to be likely clustered whether they are in operons or not (Pal and Hurst, 2004).

The debate concerning which model or variation of a model which best suits the evolution of operons is currently ongoing (Hershberg et al., 2005; Price et al., 2005; Price et al., 2008; Yerushalmi and Teicher, 2007). Operons and genes which are functionally related have been reported to be in close proximity with each other. For example, in *E.coli,* genes and operons involved in sulphur metabolism are clustered together on the chromosome, creating a compartmentalization of the process (Rocha et al., 2000).

Although there is high protein sequence similarity between gene products across bacterial genomes, many bacterial genomes lack metabolic and regulatory gene systems (Koonin et al., 1996). While operons tend to be conserved there is a lack of conservation of gene order, even between closely related species such as *E.coli* and *H.Influenzae* (Tatusov et al., 1996). Comparisons between genomes demonstrate that operons are extensively rearranged during evolution while only operons which code for physically interacting proteins like ribosomal proteins are conserved (Mushegian and Koonin, 1996; Itoh et al., 1999; Dandekar et al., 1998)

A higher level of organization exists beyond that of the operons that allows for the co-ordinated expression of groups of genes. In bacteria, location and orientation of genes have been documented to influence gene co-transcription. Genes that are in close proximity to each other are co-regulated regardless of gene orientation but gene pairs transcribed in opposite directions (divergently) were shown to be the most highly co-regulated (Korbel et al., 2004). Bacterial chromosomes generally possess more genes on the leading strand than on the lagging strand to varying degrees across species, which could be explained by the specific polymerase that the particular organism has. (Korbel et al., 2004; Rocha, 2002). There is a higher percentage of essential genes than non-essential genes on leading strand (Rocha and Danchin, 2003). The reason for this was speculated to be that head on collisions between polymerases in the replication fork would lead to shortened RNA transcripts resulting in incomplete peptides thus non-functional proteins.

Repeat elements found in tandem duplicate genes can recombine by homologous recombination or illegal recombination. Inversion events lead to inverted repeats. Bacterial chromosomes possessing fewer inverted repeats were shown to be more stable than those which had more (Achaz et al., 2003). DNA strand bias was also shown to be negatively correlated with the number of inverted repeats in a chromosome (Achaz et al., 2003). The above mentioned observations suggest that genome structure and positioning of genes influences bacterial gene expression.

Operons are essential in prokaryotes (see above), but in eukaryotes, only a few organisms possess operons, such as nematodes (Spieth et al., 1993; Blumenthal et al., 2002), trypanosomes (Johnson et al., 1987), tunicates (Satou et al., 2006), flatworms (Davis and Hodgson, 1997) and to a very limited degree, fruitflies (Ben

Shahar et al., 2007) and humans (Lee, 1991). Eukaryotic genomes are packaged into chromatin domains. Chromatin comprises of lengths of approximately 148bp DNA wrapped around core histone proteins (two copies of H2A-H2B dimer and a H3-H4 tetramer) to form nucleosomes, which in turn are tightly coupled together. Gene expression is controlled by remodelling of chromatin between an open and closed state i.e. euchromatin and heterochromatin. In the euchromatic state, a region of nucleosomes is unwound, allowing gene expression which is inhibited in the heterochromatic state. A genome-wide analysis of *Drosophila,* revealed that a third of testes-specific genes were found to be clustered on chromosomes in groups of three or more (Boutanaev *eta',* 2002). Similar clustering trends were found in the embryo and adult head. Chromatin remodeling has been shown to be a contributing factor to gene co-expression as domains of high nucleosome occupancy correlated with high gene co-expression in *Saccharomyces cerevisiae* (Batada et al., 2007).

Gene co-expression studies in eukaryotes, using high-throughput data, have shown that many adjacent genes are highly co-expressed. The obvious explanation would be operons (discussed above) and tandem duplicates. However, studies have shown that beyond these, significant gene co-expression still occurs. Promoter sharing (Hurst et al., 2002; Spellman and Rubin, 2002), gene pairs with similar functions (Lee and Sonnhammer, 2003), gene pair orientation and intergenic distance (Cohen et al., 2000; Williams and Bowles, 2004) have been shown to participate in co-expression.

1.2 Co-expression of Neighbouring Genes

1.2.1 Investigations into gene distance and gene pair orientation

Gene co-expression in eukaryotic genomes has been extensively studied with respect to gene adjacency, gene proximity and gene orientation. Pearson correlation coefficient (R) computations of expression values of adjacent genes from high-throughput data such as microarrays and EST data, have generally been used as an indication of co-expression (Cohen et al., 2000; Williams and Bowles, 2004; Zhan et al., 2006). Tandem duplicates have been known to influence results of gene co-expression studies as their expression is highly correlated; consequently, many reports have compared results of data in the presence and absence of tandem duplicates. Gene pairs are transcribed in three orientations, parallel (--+-44--4—), convergent (--+4—) and divergent

Cohen and coworkers showed that adjacent genes were more highly co-expressed then non-adjacent genes in *Saccharomyces cerevisiae* (Cohen et al., 2000). Co-expressed gene pairs were also mostly divergently transcribed, but the authors were unable to determine if this was a result of sharing of upstream activation sites or if the promoters for the genes lay close together. Studies using yeast cell cycle microarray data identified sets of adjacent genes involved in ribosomal functions which are highly correlated, thus leading to the conclusion that they share a regulatory element (Kruglyak and Tang, 2000). Highly co-expressed adjacent gene pairs were shown to be conserved between yeast and *Candida* (Pal and Hurst, 2002). In addition, highly conserved gene pairs had smaller intergenic distance than non-conserved gene pairs.

In *Arabidopsis,* neighbouring genes were highly co-expressed even after tandem duplicates were excluded using microarray datasets from Affymetrix (Williams and Bowles, 2004). Neighbouring genes were described as genes within 10 genes of each other. Gene clusters of up to 20 genes were found and there was significant correlation found between co-expression of gene pairs for intergenic distances of up to 12kb. An examination of gene pair orientation found that the most highly co-expressed gene pairs were divergently orientated followed by parallel orientation. This suggested that like in *S. cerevisiae* (mentioned previously), there may be sharing of regulatory elements among highly co-expressed gene pairs.

Analysis using Affymetrix data from *A. thaliana* root and MPSS (Massive Parallel Signature Sequencing repository) data but with stricter criteria have also been performed (Ren et al., 2005). In this case, adjacent gene pairs were defined as gene pairs that physically lay next to each other on the same chromosome and genes with a Pearson correlation coefficient > 0.7 were considered co-expressed. Approximately 20% of the co-expressed gene pairs were tandem duplicates. It was noted that only 58 gene pairs were in common between both datasets and all gene pairs were found scattered across the genome (Ren et al., 2005). Co-expression domains of three and four genes were also found but were much rarer than gene pairs. When gene pair orientation was considered, there was no preference among gene pairs but parallel transcription was found to occur twice as often as the other two orientations. This was explained by parallel orientation having two sets of directions (parallel forward and parallel reverse). Co-expressed adjacent gene pairs were found to occur at distances of up to 12 kb in any orientation. Similar studies were conducted on the rice genome *Oriyza sativa* with similar findings (Ren et al., 2007). Interestingly, no microsynteny was found between rice and Arabidopsis co-expression domains which might indicate that these domains did not play a significant role in genome conservation.

The idea of arrangement of *Arabidopsis* genes into co-expressed groupings led to further investigation in which 128 Affymetrix microarrays, each from different experimental conditions were used (Zhan et al., 2006). Genes were defined as adjacent using the same criteria idea as Ren et.al 2005. 497 genes in 226 groups were found to be positively correlated while only 15 genes were found to be negatively correlated. Ninety two percent of the genes occurred in groups of 2 and 3 scattered across the genome. Highly co-expressed genes were found to be closer than less co-expressed gene pairs with respect to gene distance. Gene orientation results were the same as those previously reported (Ren et al., 2005).

1.2.2 Regulation by a common promoter

It has been reported that some adjacent genes are co-expressed because they share a common promoter (Hurst et al., 2002; Spellman and Rubin, 2002; Takai and Jones, 2004). The promoter will have to be positioned either upstream or downstream of the genes or between the genes it regulates. Recent reports in *Drosophila* have shown that approximately 12 % of the genome contains adjacent gene pairs that are separated by less than 350bp and are divergently arranged (Herr and Harris, 2004). These pairs also display higher levels of co-expression than gene-pairs in other orientations. Using genes involved in sphingolipid metabolism as an example, it was hypothesized that genes that were divergently arranged and highly co-expressed could share a common promoter.

Divergently transcribed gene pairs have also been shown to occur in 10% of human genes (Trinklein et al., 2004). Most of these gene pairs are also highly co-expressed while some have been shown to be down-regulated. The study also showed that there were promoter regions between a pair of genes which co-regulated both of them. The divergently transcribed gene pairs were also highly conserved in mouse. Studies involving five yeast species showed that there was a very low proportion of adjacent gene pairs that were conserved across species (Tsai et al., 2007). In addition, while there was sharing of transcription factors among some of the divergently transcribed gene pairs, the proportion was not as high as in mammalian genomes and was not the major reason for adjacent gene co-expression.

1.2.3 Clustering of genes sharing similar function

Many occurrences where genes which belonged to the same functional class in *Arabidopsis* were shown to be clustered together (Riley et al., 2007). This finding was determined in the absence of tandem duplicates as they are likely to form clusters and are also likely to belong to the same functional class. Investigation into the degree of clustering of genes involved in KEGG pathways was undertaken in sequenced eukaryotic genomes (Lee and Sonnhammer, 2003). *Saccharomyces cerevisiae* was determined to have the most clustering with the degree of clustering decreasing in the order of *Homo sapiens, Caernorhabditis elegans, Arabidopsis thaliana* and *Drosophila melanogaster.* Across the genomes there was no significant conservation of gene clustering.

Co-expression of genes involved in the same pathway would provide a plausible reason for gene clustering but the results varied among pathways. In *A. thaliana,* the strongest co-expression was detected between genes which produced products, such as the ribosomal proteins. Gene pairs from metabolic pathways were not highly co-expressed except for those involved in fatty acid biosynthesis and the TCA cycle (Williams and Bowles, 2004). Applying GO terminology to co-expressed non-homologous gene pairs in *Arabidopsis* to determine if they interact, revealed that 29% of these groups shared GO biological process level 3 terms, such as cellular physiological process and metabolism (Zhan et al., 2006). In contrast, *Drosophila* co-expressed genes demonstrated no GO function relatedness (Spellman and Rubin, 2002) while in humans, co-expressed genes belonging to the same GO category were infrequent (Fukuoka et al., 2004).

1.2.4 Co-expression of linked genes is disadvantageous

While co--expression studies of linked genes have generally indicated it is beneficial for a genome, there has been recent evidence reported in mammalian genomes which indicate otherwise (Liao and Zhang, 2008). The two models which have been used to describe co-expressed linked genes are the Adaptive Model (Singer et al., 2005; Hurst et al., 2002) and the Neutral Model (Eszterhas et al., 2002; Semon and Duret, 2006). The Adaptive Model states that it is beneficial for genes that require each other to be brought together by chromosomal rearrangement. When the linkage between the relevant genes is established, it is maintained by purifying selection. The problem with the Adaptive Model was when Gene Ontology (GO) was used to define protein function, there were few linked gene pairs with the same function (Fukuoka et al., 2004; Spellman and Rubin, 2002). The latter model describes the co-expression of linked genes as a result of the effect one gene has on the other, termed transcriptional inference, which might not always be advantageous. This arrangement is caused by cis-regulatory elements or chromatin structures. A variation of the Neutral Model states that the majority of gene expression differences occurring between species are either selectively neutral or nearly neutral, and therefore of no functional significance. The rate of evolution here is rather driven by the rate of mutation and removal of deleterious changes by negative selection (Semon and Duret, 2006). Some studies favoured the Neutral Model (Khaitovich et al., 2004) but was later disputed over technical errors (Liao and Zhang, 2006). The new model, by Liao and Zhang, postulates that it is disadvantageous for linked genes to be co-expressed.

Correlation was determined by applying a formula using Pearson's Correlation coefficient to human and mouse GeneAtlas V2 microarray datasets. Few linked genes were found adjacent to each other, including some over large distances (up to 10Mb). So the authors decided not to restrict study to only adjacent genes but also to include nonadjacent genes that were also highly correlated (up to

100Mb apart). Approximately 518 000 non-adjacent gene pairs from about 4800 genes were determined. Evolutionary conservation of linkage was tested to determine which model was best suited. If the Adaptive Model applied then linkages would be maintained through evolution. If the Neutral Model applied then there would be no difference in the conservation in linkage between highly co-expressed gene pairs and between gene pairs with low co-expression levels. Finally, if co-expression is disadvantageous, then the conservation of linkage would be broken several times for both highly co-expressed and weakly co-expressed gene pairs.

The conservation of the linked gene pairs was determined by using mammalian phylogeny to infer that if two linked human genes have their orthologs linked in the dog genome then the genes are also linked in the common ancestor of rodents. Rat and mouse genomes have undergone numerous rearrangements during evolution which allowed the authors to sort the linked gene pairs into those with conserved linkage and and those non-conserved linkage. Non-conserved linked gene pairs had higher co-expression levels than conserved linked gene pairs. Also, there was a higher proportion of highly co-expressed gene pairs with non-conserved linkage than highly co-expressed gene pairs with conserved linkage. This suggested that natural selection acted against the conservation of highly co-expressed linked gene pairs (Liao and Zhang, 2008). These findings disagree with both the Adaptive and the Neutral Model but agree with the detrimental model.

The model was described simply as follows: Two genes A and B are expressed but not together, with gene A performing at its optimum but not gene B. A mutation occurs which causes gene B to perform optimally as well as adjusting the expression profile of gene A. This allows a linkage between the genes to be established. But gene A is under strain as the current expression profile is not optimum, resulting in a breakage in the linkage. Gene A returns to its original expression state, while gene B remains the same (Liao and Zhang, 2008).

1.3 Gene duplication

When a gene is copied to produce an identical gene adjacent to the original gene, a tandem duplicate or tandem repeat is created. The tandem duplicate can be mutated through evolution to produce a gene with another novel function, by a process known as neo-functionalization or it could be silenced (non-functionalization) or the gene and its duplicate become complementary to each other, sharing the original gene's function (sub-functionalization). During whole genome duplication events, entire genomes are duplicated; alternatively parts of chromosomes can be duplicated (segmental duplication).

Lynch and Conery reported that most duplicate genes are silenced following duplication with only a few that evolved new functions (Lynch and Conery, 2000). Expression divergence was shown to occur at a faster evolutionary rate in duplicate genes as opposed to single-copy genes in *Drosophila* and yeast species (Gu et al., 2004). In *Arabidopsis,* it was shown that younger tandemly duplicated genes are more likely to be susceptible to intraspecific variation than older ones, though older segmental duplicates also displayed some level of intraspecific variation (Kliebenstein, 2008). The metabolic and biosynthetic pathways also displayed low levels of variation, while the defense pathways had higher levels. The variation levels of the defense pathways have been suggested to be as a means of adapting to infections and intruders (Kliebenstein, 2008).

Divergence in expression between duplicate gene pairs is positively correlated to synonymous and nonsynonyomous substitutions in humans (Makova and Li, 2003) and yeast (Gu et al., 2002). However in Arabidopsis there is a strong positive relationship between only expression divergence and synonymous substitutions with a weak relationship between expression divergence and nonsynonymous substitutions (Ganko et al., 2007).

Tandem duplication and whole-genome duplication events can contribute to gene family expansion, for example, polygalacturonases (PGs) in plants (Kim et al., 2006). Tandem duplicated genes are frequently present in gene clusters and studies indicate these genes are sources of gene co-expression (Lercher et al., 2003). However, removal of all tandem duplicates from datasets for genome-wide analyses, demonstrate that it is not the only contributing factor to gene co-expression (Mayor et al., 2004).

1.4 An Overview of Comparative Genomics in Arabidopsis

Comparative genomics is the study of genomic sequences in which species are compared to understand genomic structure, function and evolution. The completion of the sequencing of the *Arabidopsis thaliana* genome in 2000 has positioned it as a popular model system in the field of comparative genomics of plants as well as other eukaryotes (The Arabidopsis Genome Initiative, 2000). The genome was found to contain 25 498 protein-coding genes sharing the diverse range of functions found in other previously sequenced multi-cellular eukaryotes, *Drosophila* and *Caemohabditis elegans.* Approximately sixty-nine percent of the genome could be assigned functions according to sequence similarity with other organisms. The *Arabidopsis* genome is also rich in tandem duplicates and segmental duplications (approximately sixty percent) which were postulated to contribute to its large genome size. The authors hypothesized that the large number of duplicate regions was due to the plant's ancestral history of polyploidy (The Arabidopsis Genome Initiative, 2000).

Multiple research studies into the evolutionary history of the *Arabidopsis* genome produced varying results. Three whole genome duplication (tetraploidy) events were postulated after comparing genes from *Arabidopsis* to genes from other plants with a potential evolutionary tree being described (Bowers et al., 2003). Vision et. al instead identified three segmental duplication events by inferring from patterns of sequence divergence of duplicated blocks (Vision et al., 2000). Duplicated blocks were neighbouring genes that possessed high sequence similarity to other neighbouring genes elsewhere in the genome. An evolutionary model was also presented by Maere et. al which also supported the hypothesis of three whole genome duplication events (Maere et al., 2005).

The complete sequencing of *Arabidopsis* has enabled it to participate in many comparative genomic studies. Analysis of chromosomal homology by collinearity and synteny has provided an insight into the evolution of the chromosomes.

Collinearity is the conservation of the order of genes between different species. Synteny is described as the presence of two or more genes on the same chromosome which may or may not be linked. Collinearity and synteny has been identified between Arabidopsis and other plant genomes. Liu et. al investigated colinearity between the genomes of Arabidopsis and rice *(Oryza sativa)* using the complete Arabidopsis genome and BAC sequences for rice (Liu et al., 2001). Several small syntenous regions which were interrupted by non-collinear genes were found. Collinearity was conserved among homologous genes. Wang et. al developed a statistical approach to find collinearity and infer chromosomal homology between Arabidopsis and rice (Wang et al., 2006). Several syntenous segments were reported between the two genomes, encompassing -33% of Arabidopsis and -17% of rice, but the segments were very small in size (<0.6Mb). This observation was also seen in the previously mentioned study (Liu et al., 2001). This study concurred that extensive rearrangements occurred in both genomes after the divergence of monocots and dicots. The region surrounding the *bronze (bz)* gene in maize was compared to Arabidopsis. It was found to gene rich and to contain many putative homologues to Arabidopsis but no colinearity could be established (Fu et al., 2001).

Arabidopsis has also been compared to other members from its angiosperm family *Brassicaceae.* A high degree of colinearity was detected between Arabidopsis and a genetic linkage map, composed of two *Capsella* genomes, *C.rubella* and C. *grandiflora* (Boivin et al., 2004). Fourteen collinear segments were reported which comprised 85% of the Arabidopsis genome and 92% of the *Capsella* genetic linkage map. Several genome rearrangements caused by fusions, fissions and inversions were identified as the reason for the high colinearity. A system was proposed to provide a unified perspective of the *Brassicacaece* family. An ancestral karyotype with a chromosome number of n = 8 based on cytology and genetic maps of *Arabidopis lyrata* and *Capsella rubella* was created (Schranz et al., 2006). The karyotype was not based on

A.thaliana **as it has 5 chromosomes and would make comparison within the family *Brassicacaece* difficult.**

Chapter 2: CAGED

2.1 Introduction to CAGED

Analysis of high-throughput gene expression data in the context of genome organization is generally performed by custom scripts. The Arabidopsis genome, in particular, has been the focal point of many recent studies, with respect to gene co-expression as were reviewed in Chapter 12. There are a number of applications and tools that have been written to provide assistance in gene expression analysis and the identification of clusters of co-expressed genes (Coppe et al., 2006; Jen et al., 2006; Mutwil et al., 2008; Srinivasasainagendra et al., 2008). Generally, these tools are able to generate lists of highly correlated gene pairs based on preprocessed data and produce scatterplots of gene pairs. The lists can be filtered for potential gene co-regulation via algorithms or GO terminology. The shortcoming of many of these programs is that they require the user to be knowledgeable in its use. This might limit its use among plant biology researchers with limited bioinformatics experience. An intuitive and basic tool with a graphical interface would provide an insight into gene co-expression with a genomic view. CAGED was developed as an online visualization tool which displays co-expressed Arabidopsis genes in their genomic locations to facilitate the investigation of relationships between genome organization and gene expression. CAGED is available at http://web.cbio.uct.ac.za:9090.

2.2: Materials and Methods

2.2.1. Preprocessing of Data

2.2.1.1. RMA processing

CAGED is an online tool which contains a database of co-expressed genes with GO Slim annotations built using microarray datasets. Analysis was performed on 1758 Affymetrix ATH1 microarray datasets purchased from Nottingham Arabidopsis Stock Center (NASC). When comparing results across multiple arrays, two types of variation are encountered. The first is termed 'interesting variation' as it is the difference in expression between a gene or genes from different states, eg. normal and diseased. The second type of variation is `obscuring variation' which is an accumulation of errors that could be introduced during sample preparations, production and processing of the genome arrays. Normalization is used to remove obscuring variation which would otherwise lead to incorrect results.

Datasets were normalized using RMAExpress (http://rmaexpress.bmbolstad.com) which implements the robust multiarray analysis (RMA) algorithm (Irizarry et al., 2003). RMAExpress produces the same results as the RMA function in the Bioconductor package for the R programming language (R Development Core Team, 2008), but it is more memory efficient allowing larger batches of datasets to be processed. Briefly, RMA consists of three steps, background correction, quantile normalization and calculation of expression. In background correction, each chip has the probe level data adjusted using a statistical model. The observed intensity *s,* is made up of true intensity *x,* which is exponentially distributed and random noise y, which is normally distributed $(s = x + y)$. The true intensity, *x* is unknown and is calculated in the background correction step, given the information of the observed intensity, *s,* the rate a of *x,* the mean p and variance 6^2 of y (Wernisch, 2004).

Quantile normalization equalizes the distribution of the probe intensities for each array across all the arrays using a common set of mean quantiles (Bolstad et al., 2003). The idea behind the algorithm is: there are *n* arrays of length *p* which form a matrix *X,* where each array is a column. The columns of *X* are then sorted to form Xsort. The means across each row of Xsort are computed and then each element in that row gets assigned that mean to form X'sort. Finally, the columns in X'sort are returned to the order as in *X* which forms the matrix Xnormalized.

The calculation of expression is performed separately for each dataset using a linear model, *Yen= Pin ajn eijn* where *i* = 1, • • • ,*I*, *j* = 1, • • • ,*J, n* = 1, • • • ,N
pi **is the logarithm scale of the intensity level for array *i,* cti is the effect of the j^{th} probe and Eij is the error term (Bolstad et al., 2003).**

2.2.1.2. Deleted Residuals

Deleted residuals was the method of quality control applied to identify and remove potential outliers (Trivedi et al., 2005). For n genes and m datasets which form a matrix D where the genes are the columns. Each element Xis in the matrix D, where *i* is the index of the row and *j* is the index of the column, the mean for the 1417 row excluding the element Xis is calculated and then subtracted from Xis to produce a matrix D'. The standard deviation sdi is then calculated for each row and each element in that row is divided by sdi to produce a final matrix Ddelres A Python script was written to perform this operation using the RMA processed data.

2.2.1.3. Kolmogorov — Smirnoff (K-S) goodness-of-fit test

The Kolmogorov-Smirnoff test is generally used to compare a sample cumulative distribution with a theoretical cumulative distribution. The biggest difference between the two is known as the D **statistic, which is used to reject or accept the null hypothesis (Porkess, 2004). This test was applied to determine if a sample**

comes from a population with a specific distribution and remove outliers. To identify a cut-off D value, Persson et. al. processed all the Affymetrix Chips in the Gene Expression Omnibus (Persson et al., 2005). A cut-off of 0.15 was decided due to the change in slope of the K-S D curves between the 80th and 90th percentile. Our data displayed similar change in pattern and a cut-off of 0.15 was also chosen. A Python script using the K-S function from the Rpy package (http://rpy.sourceforge.net) was used to filter the results from the deleted residuals and remove outliers.

2.2.2. Building CAGED

2.2.2.1. Determination of co-expressed gene pairs

The image maps of CAGED display significantly co-expressed neighbouring gene pairs that have a Pearson correlation co-efficient (R) of 0.5 or higher. A Python script using the Rpy module was written to compute R for every neighbouring gene pair using the pre-processed data as the data source. Neighbouring gene pairs were defined as genes that lay within 100 000 by distance of each other. Annotations for Arabidopsis genes that included Probe Identifiers, gene names, chromosomal locations and gene orientation were downloaded from NASC (http://arabidopsis.info).

2.2.2.2. Determining Paralogs and Orthologs

It was likely that some of the co-expressed gene pairs were paralogs and it would be interesting to display these gene pairs on the image maps. To identify these pairs, an all-against-all BLASTP was performed with the Arabidopsis protein sequences (downloaded from TIGR (ftp://ftp.tigr.org)) against itself using default parameters. Gene pairs which had an E-value of < 10^{-5} were considered as paralogous. Rice *(oryza sativa)* is another commonly studied plant genome which is often compared to the Arabidopsis. Using protein sequences

downloaded from TIGR (ftp://ftp.tigr.org), a best reciprocal BLASTP hit match between rice and Arabidopsis was performed to produce a list of orthologs.

2.2.2.3. Assembling the Website

Image maps were generated using a Python script with the aid of the Sping drawing module (http://sping.sourceforge.net). Co-expressed gene pairs were represented by black arcs. Paralogous relationships were represented by green arcs and orthologous pairs in red CAGED was built using the Python web application framework Turbogears (http://www.turbogears.org). The interface and interactivity of the CAGED was constructed using the html-like Kid template component. The database section was built using SQLObject component. GOslim annotations were downloaded from TAIR (ftp://ftp.arabidopsis.org) and formatted via a custom Python script.

2.3 Results and Discussion

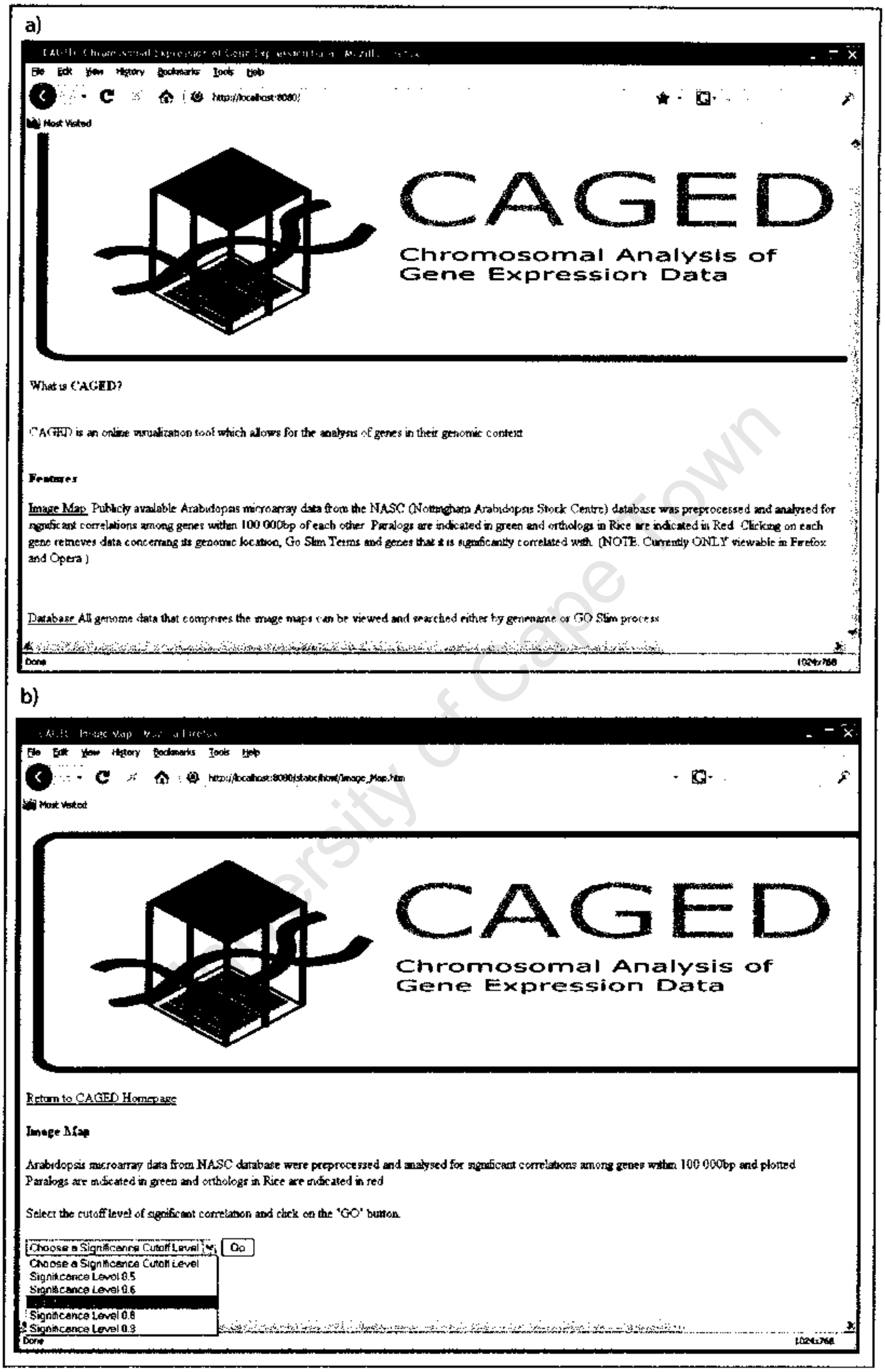
2.3.1 Exploring the features of CAGED

On loading the CAGED homepage (http://web.cbio.uct.ac.za:9090), the user is presented with two options: "Image Map" or "Database" as seen in Figure 1a. Image Maps are viewable on Mozilla Firefox and Opera web browsers only. These display Arabidopsis chromosomes with arcs connecting co-expressed genes that lie on the same chromosome. The CAGED database is comprised of highly co-expressed Arabidopsis genes with annotation information from NASC (http://arabidopsis.info). Selecting the "Image Map" option directs the user to a webpage with a drop-down box for selecting the significance cut-off level for Pearson correlation between gene pairs, ranging from 0.5 to 0.9, in 0.1 increments (Figure 1 b), which leads to the image map of choice (Figure 2a).

There are five dark green lines drawn across the image which represent Arabidopsis chromosomes. Co-expression of a gene pair is assessed on the basis of gene expression correlation across 1758 experiments obtained from the

NASC database (http://arabidopsis.info) and represented by arcs which link the genes. The height of the arcs is proportional to the Pearson correlation coefficient. Green arcs are drawn between coexpressed paralogs and co-expressed gene-pairs for which microsynteny is conserved in rice are shown in red. Mousing over to the beginning or end of each arc displays a pop-up of the name of the correlated gene. Clicking on the gene name reveals a webpage retrieved from the CAGED database containing information pertaining to the gene of interest's chromosomal location, GO Slim terms and other genes that it is correlated with together with the degree of correlation (Figure 3a).

CAGED also permits the user to search its database from the "Database" hyperlink. This leads to a webpage, as shown in Figure 2b, containing a drop-down box option for either browsing the database by gene name or GO Slim term. Choosing to browse by gene name will lead to a text box where the user has to input the name of the gene of interest. Browsing by gene name displays the same gene information as previously mentioned in the "Image Map" option (Figure 3a). If the gene is not listed, a "gene not found" message is displayed and the user can input another genename. The GO Slim option displays a drop-down box displaying all Go Slim terms (Figure 3b). Choosing the term of interest produces all genes that share that term. The examples that follow demonstrate the application of CAGED.


a)
File Edit View History Bookmarks Tools Help
http://localhost:8080/
Most Visited
CAGED
Chromosomal Analysis of Gene Expression Data
Done
1024x768

What is CAGED?

CAGED is an online visualization tool which allows for the analysis of genes in their genomic context

Features

Image Map Publicly available Arabidopsis microarray data from the NASC (Nottingham Arabidopsis Stock Centre) database was preprocessed and analysed for significant correlations among genes within 100 000bp of each other Paralogs are indicated in green and orthologs in Rice are indicated in Red Clicking on each gene retrieves data concerning its genomic location, Go Slim Terms and genes that it is significantly correlated with (NOTE. Currently ONLY viewable in Firefox and Opera)

Database All genome data that comprises the image maps can be viewed and searched either by genename or GO Slim process

b)
File Edit View History Bookmarks Tools Help
http://localhost:8080/static/html/Image_Map.htm
Most Visited
CAGED
Chromosomal Analysis of Gene Expression Data

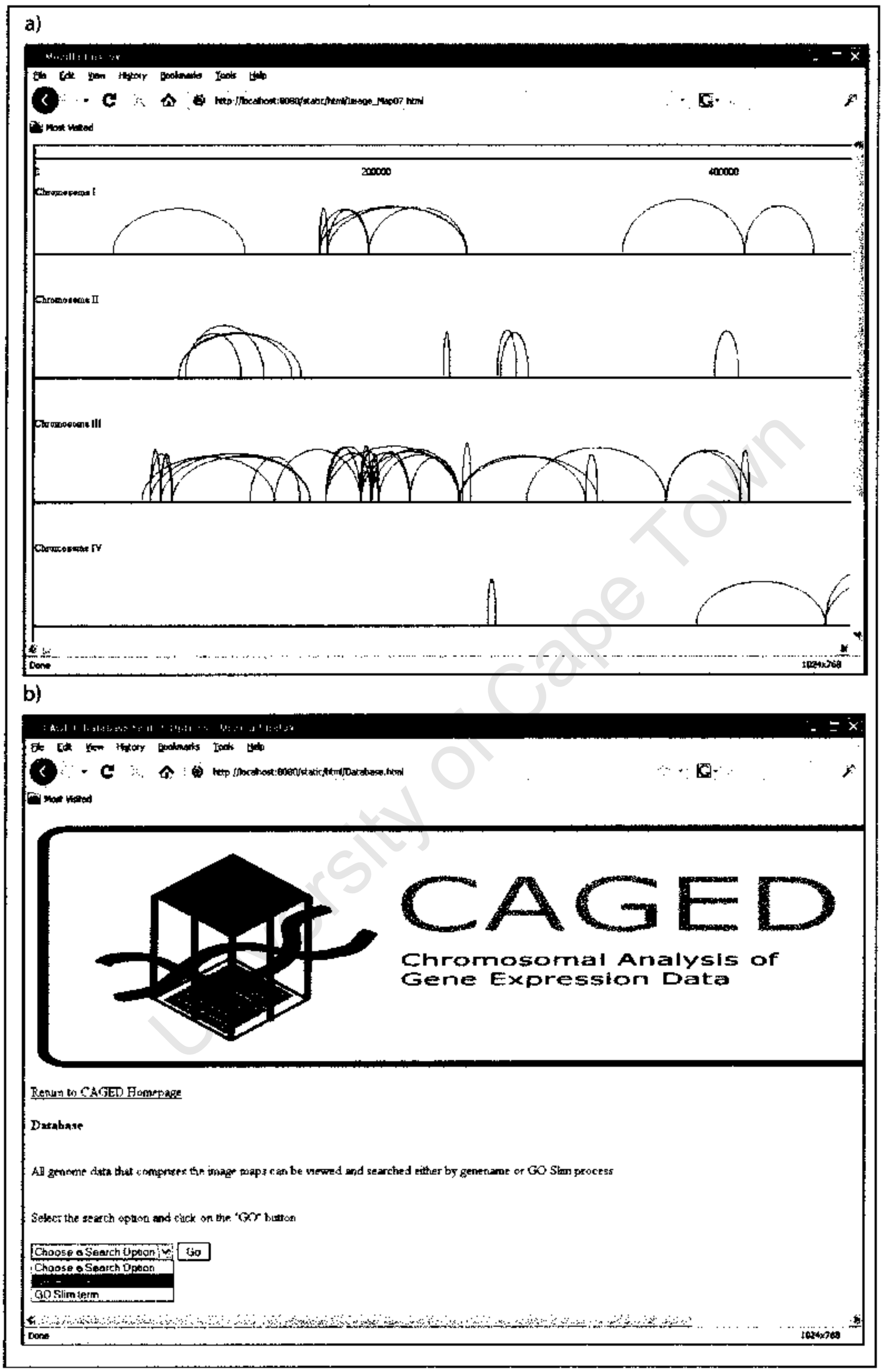
Return to CAGED Homepage

Image Map

Arabidopsis microarray data from NASC database were preprocessed and analysed for significant correlations among genes within 100 000bp and plotted Paralogs are indicated in green and orthologs in Rice are indicated in red

Select the cutoff level of significant correlation and click on the 'GO' button.

Figure 1: The features of CAGED: a) the homepage, and b) the Image Map selection menu.


a)
File Edit View History Bookmarks Tools Help
http://localhost:8080/static/html/image_Map07.html
Most Visited
200000
400000
Chromosome I
Chromosome II
Chromosome III
Chromosome IV
Done
1024x768
b)
File Edit View History Bookmarks Tools Help
http://localhost:8080/static/html/Database.html
Most Visited
CAGED
Chromosomal Analysis of Gene Expression Data
Choose a Search Option
Go
Choose a Search Option
GO Slim term
Done
1024x768

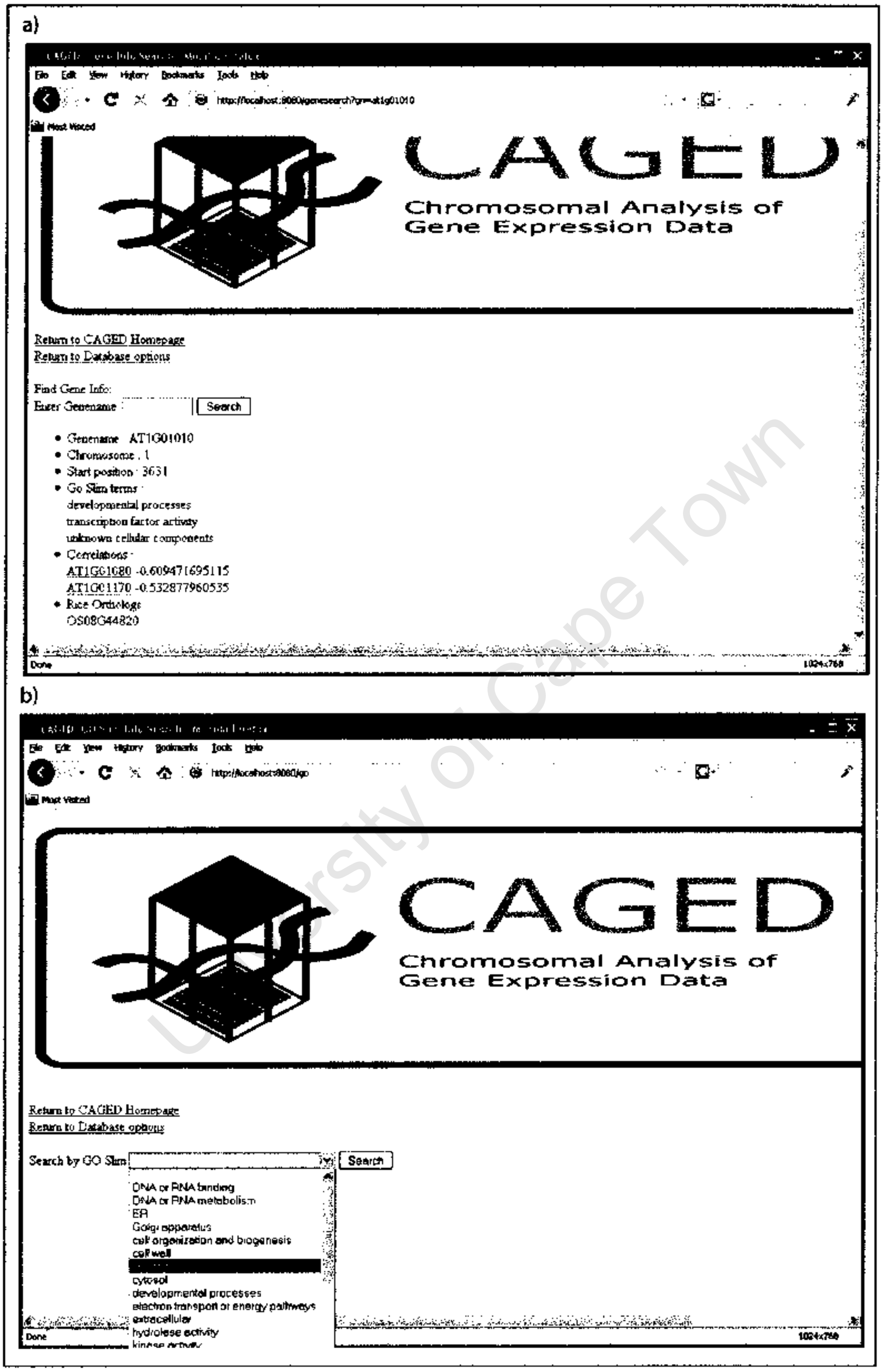
Return to CAGED Homepage

Database

All genome data that comprises the image maps can be viewed and searched either by genename or GO Slim process

Select the search option and click on the 'GO' button

Figure 2: CAGED features continued: a) Image map displaying gene pair correlation with a significance cut-off 0.7 (Green arcs indicate orthologs in *O.sativa.* Red arcs indicate paralogs)and b) the Database menu



Return to CAGED Homepage
Return to Database options

Find Gene Info:
Enter Genename:

- Genename AT1G01010
- Chromosome : 1
- Start position : 3631
- Go Slim terms :
  developmental processes
  transcription factor activity
  unknown cellular components
- Correlations :
  AT1G01080 -0.609471695115
  AT1G01170 -0.532877960535
- Rice Orthologs
  OS08G44820

Return to CAGED Homepage
Return to Database options

Search by GO Slim

Figure 3: CAGED Features continued: a)Gene information display page and b) GO Slim term selection menu

2.3.2 Visualizing previously established gene relationships in CAGED

Diverse gene relationships can be examined on CAGED. Viewing a previously established region of co-regulated genes may reveal some new insights. An operon-like gene cluster involved in triterpene synthesis was identified in Arabidopsis (Field and Osbourn, 2008). Triterpene is involved in the thalianol pathway which is responsible for plant resistance to disease and pests. Genes identified were oxidosqualene cyclase (OSC) genes which produce a variety of triterpenes. There were four OSC genes of interest found to occur in close proximity of each other on chromosome 5. As gene expression between the genes was highly correlated and all four of the genes occur in the root epidermis it was likely that they were functionally related. *At5g48010* was determined to encode a thalianol synthase (THAS) which is responsible for the conversion of 2,3-oxidosqualene into triterpene thalianol. *At5g48000* and *At5g47990* were both initially identified to encode cytochrome P450 enzymes. Further investigation revealed the adjacent gene *At5g48000,* to be a thalianol hydroxylase (THAN) which converts thalianol into thaliana-diol. Thaliana-diol is converted into a desaturated version by *At5g47990,* a thaliana-diol desaturase (THAD) and finally *At5g47980,* is a BARD family acyltransferase which may produce acylated desaturated thaliana-diol or lead to another modified product. Despite both Arabidopsis and oats both possessing triterpene synthesis gene clusters there was no evolutionary evidence to suggest that they had a common origin (Field and Osbourn, 2008).

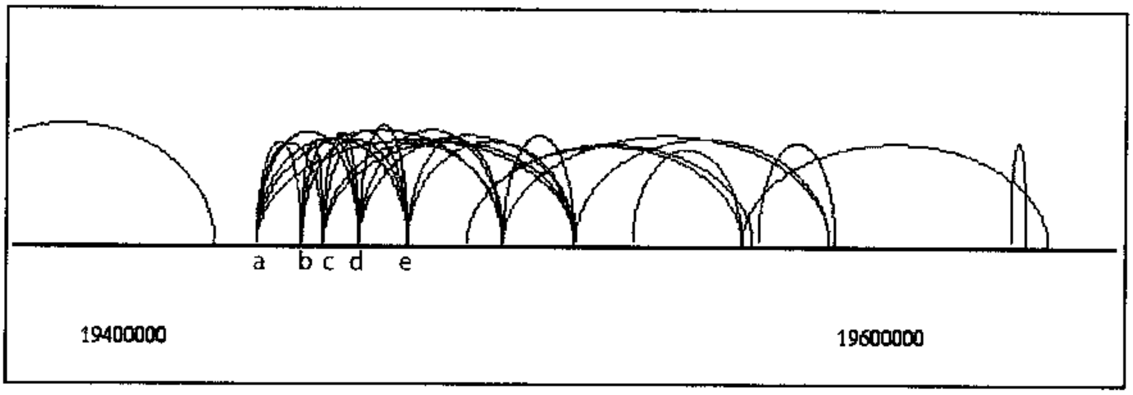


Figure 4: Genes involved in triterpene synthesis. a - *At1g47950*, b – *At1g47980*, c – *At1g47990*, d – *At1g48000* and e – *At1g48010*.

Because the expression of the genes in this cluster is highly correlated, it would be practical to begin by viewing the image map which displays arcs with a Pearson correlation coefficients of 0.7 or higher. The chromosomal region can be found by searching for chromosome V and then scrolling horizontally to the area between 19.4 and 19.6 Mbp (Figure 4). The region has several highly correlated gene pairs which include the genes that are involved in triterpene synthesis. *At5g48010* (THAS) should be the first gene to be explored since it is the initial gene to act in the pathway. Clicking on the gene retrieves the relevant data. Among the GO Slim terms listed are `other enzyme activity' and `other metabolic processes'. This corresponds to the synthase function of *THAS.* The highly correlated genes to *At5g48010* include those involved in triterpene synthesis, *At5g48000* (0.88), *At5g47990* (0.832) and *At5g47980* (0.771). This would suggest that these genes may be co-regulated. Following the trail of conversion of thalianol, the next gene to be investigated would be *At5g48000* (THAH). Clicking on this gene from the Correlations list of *At5g48010* retrieves the gene data page. *THAH* adds a hydroxyl group to thalianol with the most relevant GO Slim term listed as `electron transport or energy pathways'. The other GO Slim term, 'ER' (endoplasmic reticulum), coincides with the initial description of the *At5g48000* gene product as a cytochrome P450 enzyme, as the ER contains cytochrome P450 (CYP450) complexes.

The next gene in the pathway is *At5g47990* (THAD) which is also strongly correlated to *At5g48000* (0.824). Clicking on the gene name retrieves its data page. *At5g47990* is a desaturase but the GO Slim terms listed are not remotely relevant. But like *At5g48000,* the term 'ER' is listed, which also coincides with its broader description as a CYP450. The final gene in this cluster, *At4g47980* is listed under `Correlations' on the data page of *At5g47990.* This gene was determined to be a BAHD transferase which would correspond to 'transferase activity' listed under *At5g47980's* GO Slim terms. This demonstrates CAGED's usefulness as a tool to view previously published relationships between co-expressed gene clusters. New insights can also be gained by examining the areas surrounding the gene cluster. For example, *At5g47950* is also highly correlated to *At5g48000* (0.826) and functions as a transferase, according to its GO Slim terms. As this relationship has not been previously established, it would be worthwhile to determine if this gene would form part of the triterpene pathway, maybe in a similar capacity to, or a shared role with *At5g47980* with which its expression is significantly correlated. It should be noted that the gene pairs consisting of *At5g47950* and *At5g47980* and of *At5g47990* and *At5g48000* are classified as paralogs (indicated by green arcs).

2.3.3 Investigation into new gene relationships with CAGED

The 'Image Map' function in CAGED can also be used to find new areas of interest. Browsing at a strict level of significance for correlation between gene pairs (0.7) may present genes in close proximity to each other which could be functionally related. Using the tool we identified a region of interest between 10.2 Mbps and 10.4 Mbps on chromosome I, containing genes involved in photosynthesis (Table 1; Figure 5). Three of the genes had GO Slim terms linking them to chloroplast and responses to stress and one to abiotic or biotic factors.

Table 1: List of closely positioned genes with relevant GO Slim terms that may interact with each other during photosynthesis

Gene name	GO Slim Term
At1g29390	Chloroplast / response to stress
At1g29395	Chloroplast / response to stress / response to biotic or abiotic stimulus
At1g29460	Mitochondria
At1g29500	Nucleus
At1g29700	Chloroplast
At5g29720	Kinase

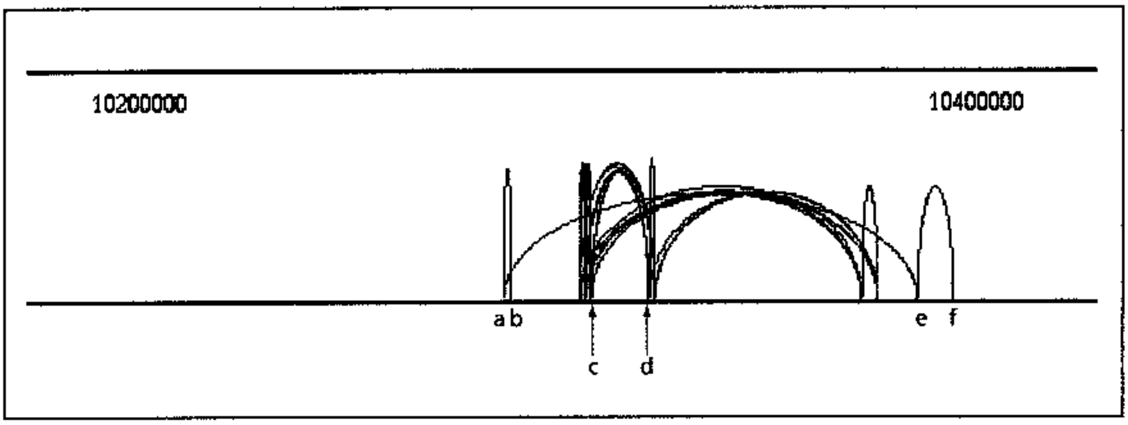


Figure 5: Genes that may be co-expressed together in response to plant stress. a – *At1g29390*, b – *At1g29395*, c – *At1g29460*, d - *At1g29500*, e – *At1g29700* and f – *At1g29720*.

While the genes are not adjacent to each other, they all lie within a 200 Kbp region and are highly correlated (as seen in Figure 5), as well as have related functions. This may suggest a coordinated expression. *At1g29390* and *At1g129395* are correlated (0.878) and possess the GO Slim terms, `chloroplast' and 'response to stress'. Both genes are also correlated to *At1g29700* (0.71) and 0.623) which shares the `chloroplast' GO Slim term, and to *At1g29460* (0.534), a gene involved in mitochondria function. *At1g29460* is also correlated to *At1g29500* (0.852) which has a role in the nucleus and *At1g29460* to *At1g29720* (0.751), a kinase. It should be observed that many of the gene pairs in this putative cluster are paralogous, notably the pairs, At1g29390 and Atl g29395, and At1g29460 and Atl g29500.

This gene arrangement may suggest that the genes in this region participate in retrograde signaling. Certain plants such as Arabidopsis have evolved mechanisms to respond to various stresses such as environmental changes. Retrograde signaling is coordinated transmissions from organelles, such as chloroplasts and mitochondria, to the nucleus to regulate the control of genes encoding organellar proteins. There are various possible pathways that could explain this phenonomen (Fernandez and Strand, 2008). The most relevant in this case describes redox reactions catalyzed by a protein kinase at Photosystem I and 11 (PSI and P511) which transmit signaling molecules to the nucleus when

there is an accumulation of reactive oxygen species (ROS) due to abiotic or biotic stresses. *At1g29390, At1g129395* and potentially *At1g29700* may encode chioroplast proteins that respond to such stresses by triggering a retrograde signaling pathway. Mitochondria! genes have been implicated in participating in retrograde pathways. Arabidopsis mutants deficient in propyl4RNA synthase (found normally in plastids and mitochondria), have reduced expression of nuclear photosynthesis genes (Pesaresi et al., 2006). *At1g29460* may be involved in an undescribed pathway with both *At1g29700* and *At1g29500* as the latter encodes a nuclear gene. It is also possible that *At1g29460* is involved in the same pathway as *At1g29390* and *At129395* as it is also correlated to the kinase encoding gene, *At1g29720* but how has yet to be determined. This example demonstrates how CAGED could be a tool in beginning to explore new regions of interest.

Chapter 3: Analysis of the relationship between gene co-expression and co-location in *Arabidopsis thaliana*

3.1 Introduction

Analysis of high-throughput data, such as microarray gene expression datasets, has shown that genes in close proximity have a tendency to be co-expressed (Cohen et al., 2000; Lercher et al., 2003; Spellman and Rubin, 2002; Williams and Bowles, 2004). Correlation in gene expression is typically measured by calculating the Pearson's Correlation co-efficient (R) using the expression values obtained from normalized high-throughput data for every gene pair. Investigations into what influences gene co-expression has lead to the following being examined: tandem duplicates, gene orientation and intergenic distance (Boutanaev et al., 2002; Fukuoka et al., 2004; Herr and Harris, 2004; Hurst et al., 2002; Lee and Sonnhammer, 2003; Ren et al., 2005; Trinklein et al., 2004; Zhan et al., 2006). For a further review see Chapter 1.2. We were interested in comparing the results of analysis of the processed data used to build CAGED for gene co-expression and the characteristics displayed by the co-expressed gene pairs to previously published results.

3.2 Materials and Methods

Highly correlated gene pairs from the CAGED data (see processing details in Chapter 2) were analysed to investigate gene pair co-expression. The gene pairs were sorted into categories: neighbouring, all adjacent, adjacent without tandem duplicates and tandem duplicates using python scripts. Genes were defined as neighbouring if they lay within 100 Kbps of each other on the same chromosome. All genes on a chromosome were numbered consecutively for each of the five chromosomes. If the difference between the numbers assigned to two genes was one then they were considered adjacent. If the expression of adjacent genes were found to be correlated then they were considered to be co-expressed. Tandem duplicates were defined as adjacent gene pairs with BLASTP e-values < 2 X 10^{1} (Lercher et al., 2003). Non-adjacent gene pairs were selected after 100 randomizations of all correlated gene pairs using the criteria of gene pairs that could be any distance apart and may lie on different chromosomes. The Mean, Median and Standard Error were calculated to obtain a statistical description of the distribution of the data. All calculations were performed with the aid of the statistical programming language R. A function was written in R to compute the Standard Error. It was the standard deviation of mean Pearson correlation coefficients divided by the square root of the population of mean Pearson correlation co-efficients.

Normal distribution of each gene pair dataset was assessed with the aid of histograms if the mean was approximately equal to the median and if the data fit a bell-shaped curve. Boxplots and Q-Q plots were plotted for a graphic comparison of the mean and median between distributions of the datasets. Student's t Test was used to determine if two datasets were drawn from populations with equal means. We used the test to assist in determining whether adjacent gene pairs would be highly co-expressed in the absence of tandem duplicates.

Gene start positions were obtained from NASC (http://arabidopsis.info). Intergenic distance was defined as the shortest distance between a gene pair. If the genes overlapped then the distance was set to zero. Scatterplots were used to observe the relationship between correlation and intergenic distance of the gene pairs. As the relationship appeared non-linear between the variables, restricted cubic splines were plotted. Restricted cubic splines model relationships by positioning control points called knots on the data through which a line is passed to form a curve. Four knots were used as this allowed flexibility of the curve without too much loss of precision.

3.3 Results and Discussion

3.3.1 Analysis of Adjacent Gene Pairs

3.3.1.1 Identification of Highly Correlated Gene Pairs

For each pair of genes we calculated the Pearson correlation coefficient of their expression across experiments as described in the Materials and Methods section of Chapter 2. Gene pairs were sorted into neighbouring gene pairs, adjacent, adjacent without tandem duplicates, adjacent tandem duplicates only using the criteria outlined in 3.2 Material and Methods. This resulted in 437274 neighbouring gene pairs, 21385 adjacent gene pairs and 1080 tandem duplicates. Subsequently, 20305 adjacent gene pairs excluding tandem duplicates were determined. A detailed description of the gene pair datasets is shown in Table 1.

Table 1: Descriptive statistics of the correlation coefficients for gene pairs in *Arabidopsis thaliana*

Gene pairs	Mean R	Std error	Median R	No of gene pairs	% gene pairs with R > 0.7
Neighbouring	0.0253	0.000399	0.012986	437274	0.636 (2779/437274)
All Adjacent	0.0731	0.001927	0.057761	21385	1.604 (343/21385)
Adjacent without td	0.0605	0.001933	0.047679	20305	1.167 (237/20305)
Adjacent td only	0.3104	0.008986	0.320718	1080	9.815 (106/1080)
Nonadjacent*	0.0164	0.001816	0.004828	20000	1.75 (350/20000)

td → tandem duplicates

* After 100 randomizations of nonadjacent data

In this study, a previously established Pearson correlation coefficient (R) > 0.7 cut-off for highly co-expressed gene pairs was used (Cohen et al., 2000; Zhan et al., 2006). The percentage of highly expressed gene pairs for each of the datasets is displayed in Table 1. Very few of the neighbouring gene pairs were highly co-expressed (2779/437274 (0.63 %)). Two percent (343/21385) of all the adjacent gene pairs were highly co-expressed. One percent (237/20305) of the adjacent gene pairs which excluded tandem duplicates were highly co-expressed. Other studies have examined co-expression domains (comprised of small sets of highly correlated genes) and putative clusters in Arabidopsis (Ren et al., 2005; Zhan et al., 2006) and other Eukaryotes (Boutanaev et al., 2002; Cohen et al., 2000) though they all agree that a small percentage of genes are highly co-expressed.

Ten percent (106/1080) of tandem duplicates were highly co-expressed. Alternatively, it could be stated that 31 % of the highly co-expressed adjacent gene pairs were tandem duplicates (106/343). Similar proportions of highly co-expressed tandem duplicates were also previously reported in Arabidopsis (Ren et al., 2005; Zhan et al., 2006) as well as in a cross species comparison study (Fukuoka et at., 2004) where highly co-expressed tandem duplicate pairs constituted between 8-30% of co-expressed gene pairs. It is should be noted that in Table 1 there is a greater percentage of gene pairs highly co-expressed for the nonadjacent genes than for adjacent without tandem duplicates. This could be explained by the presence of paralogs in the nonadjacent gene pair dataset influencing the result, since paralogs have been removed from the adjacent gene-pairs without tandem duplicates dataset.

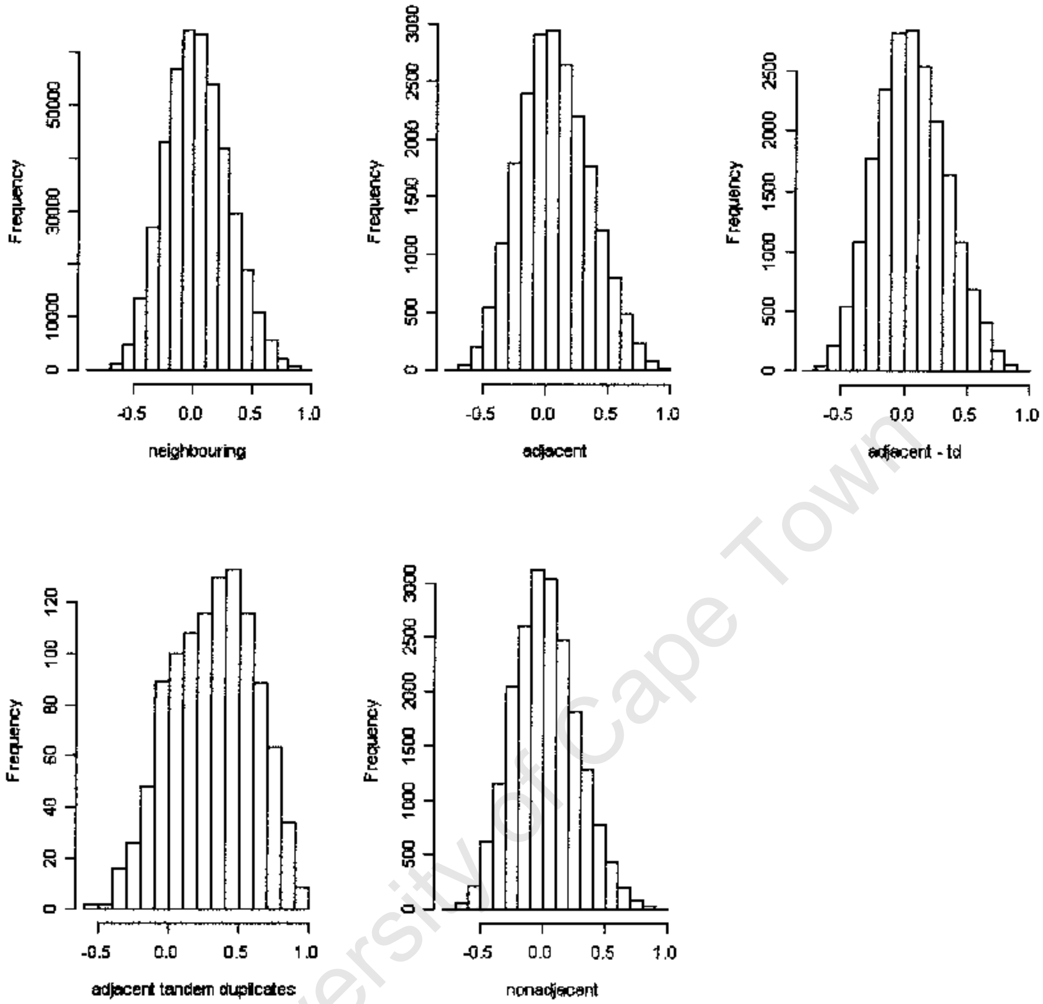
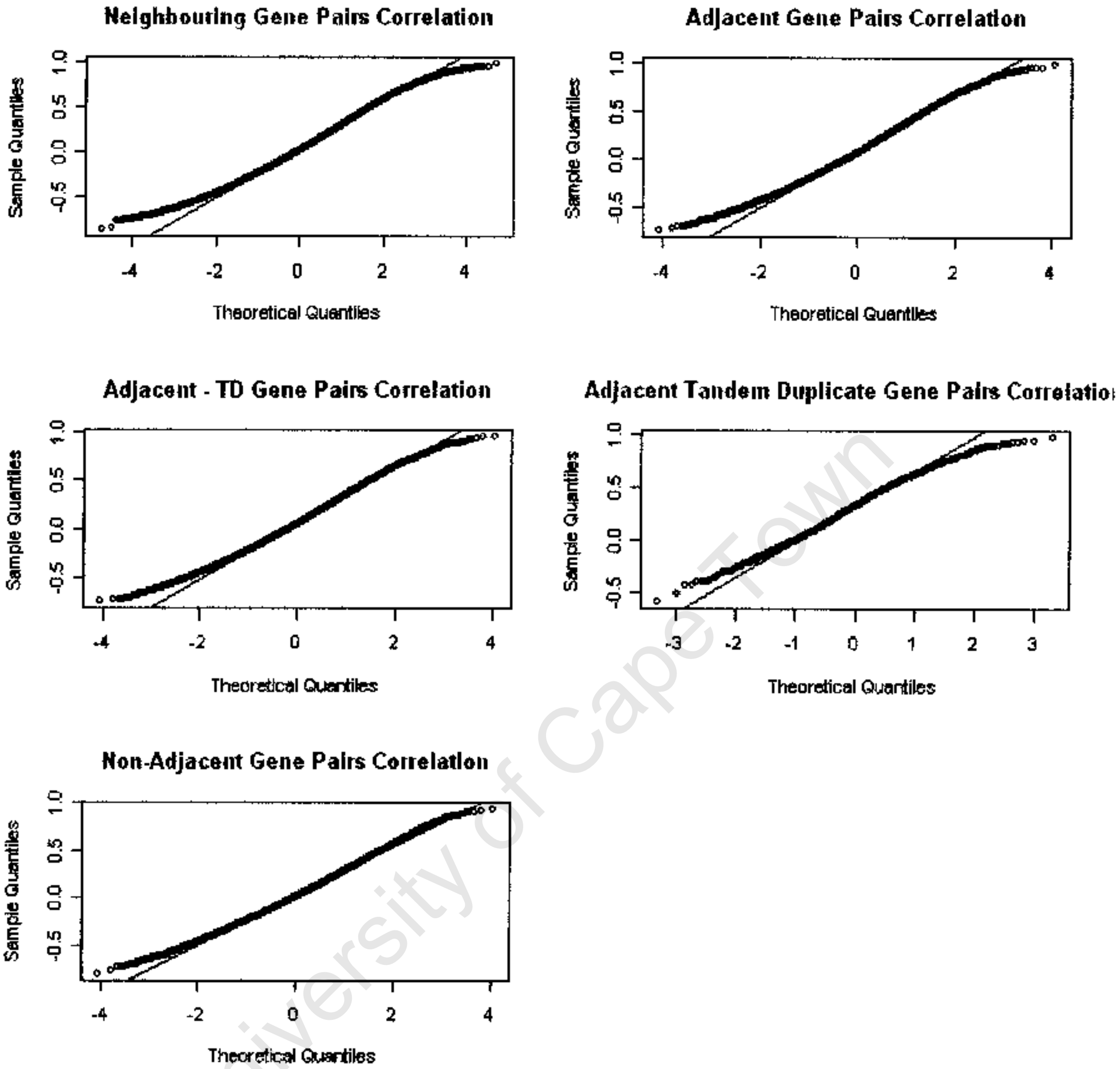


Figure 1: Histograms of the distributions for the gene pair Pearson correlation coefficients.

The distributions of the Pearson correlation coefficients were compared between the datasets and the means, medians, numbers of gene pairs and numbers of gene pairs with R > 0.7 are shown in Table 1. The histograms of Pearson correlation coefficients for each dataset followed a normal distribution (Figure 1) with positive means and medians (Table 1). The values of the mean and median of each distribution are similar, reflecting the fact that the distributions are nearly symmetric for each dataset. The Q-Q plots (Figure 2) and boxplots (Figure 3) confirm that the distributions are approximately normal. The adjacent tandem duplicates have a very positive mean which is approximately 4.5 times higher



Neighbouring Gene Pairs Correlation

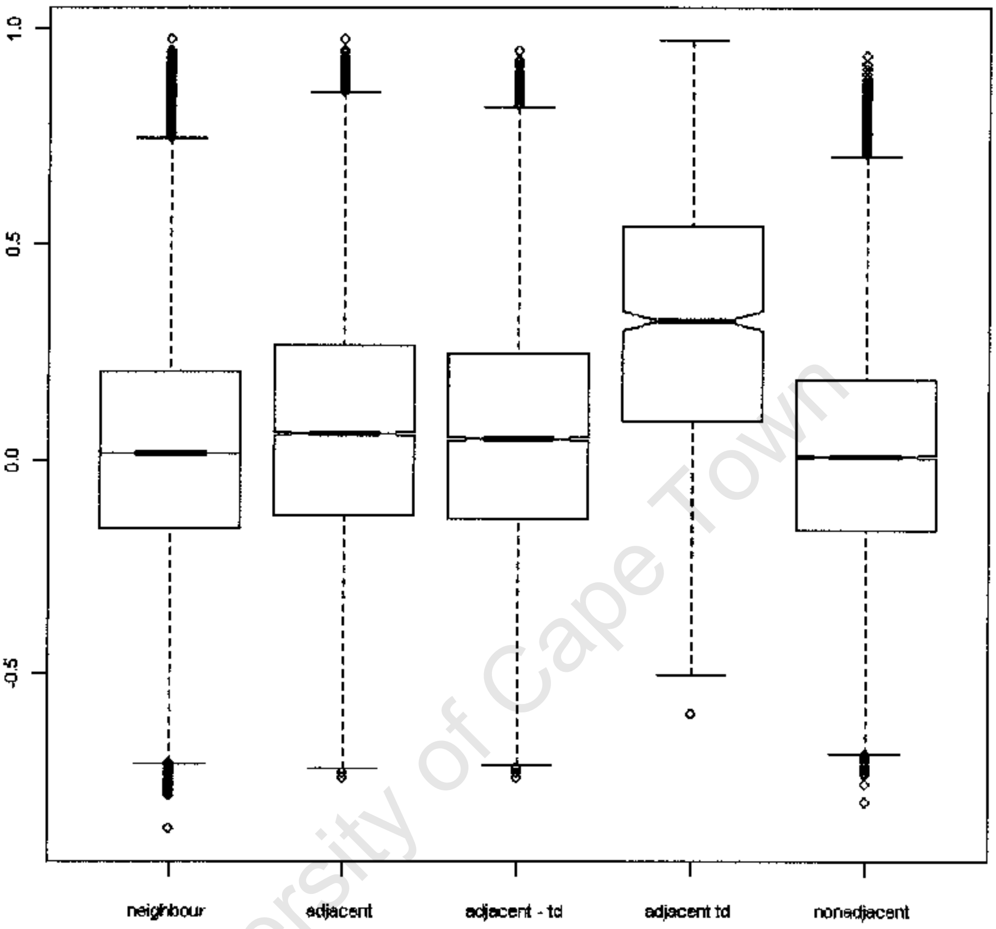
Adjacent Gene Pairs Correlation

Adjacent - TD Gene Pairs Correlation

Adjacent Tandem Duplicate Gene Pairs Correlatio

Non-Adjacent Gene Pairs Correlation

Figure 2: Q-Q plots to compare the distribution of correlation coefficients to the Normal distribution.



Note: td - tandem duplicates

Figure 3: Box plots of the distributions for the gene pair correlation coefficients.

than that of the adjacent gene pairs (Table 1). This might indicate that tandem duplicates have a major influence on co-expression of adjacent genes. When the adjacent gene pairs without the tandem duplicates dataset was compared with the all adjacent gene pairs dataset, there was a subtle difference between the values for the mean and median and distribution plots (Table 1, Figures 1, 2 and 3). Furthermore, the significant p-value from the Student's t test on the

means shows that both datasets are drawn from populations with unequal means (Table 2). This would indicate that in the absence of tandem duplicates, there must be additional factors other than tandem duplicates that drive gene co-expression.

Table 2: The diagonal contains the mean R correlation coefficient (in Bold). Student's t tests were performed on the correlation coefficients for each pair of datasets (off diagonals)

	Neighbouring	All Adjacent	Adjacent – td	Adjacent td only	Nonadjacent*
Neighbouring	**0.0253**				
All Adjacent	< 2.2e-16	**0.0731**			
Adjacent – td	< 2.2e-16	3.786e-06	**0.0605**		
Adjacent td only	< 2.2e-16	< 2.2e-16	< 2.2e-16	**0.3104**	
Nonadjacent*	3.702e-09	< 2.2e-16	< 2.2e-16	< 2.2e-16	**0.0164**

td → tandem duplicates

* After 100 randomizations of nonadjacent data

3.3.1.2 Gene pair orientation and Intergenic distance

Aside from tandem duplicates, sharing of promoter elements is another common factor known to cause co-expression of adjacent genes in *Drosophila* and humans (Herr and Harris, 2004; Trinklein et al., 2004). Potential promoter sharing is suggested by the presence of divergently orientated gene pairs that are co-expressed and subsequently confirmed by experimental assays (Trinklein et al., 2004). The datasets were divided into the three gene pair orientations, parallel, divergent and convergent. The gene pairs were expressed as the number of gene pairs in a particular orientation and as a percentage in Table 3. Gene pairs were found to be orientated in parallel for approximately half of the adjacent gene pairs and adjacent gene pairs excluding tandem duplicates (51.51 % and 49.81 % respectively, as shown in Table 3). It is likely that since parallel gene pairs are comprised of forward (—*--)) and reverse (4-4—) orientated gene pairs, there was a bias towards parallel orientation. Dividing the parallel orientated gene pairs into forward and reverse reveals that these gene pairs are nearly equal in proportion (see Supplementary Table 1). The distribution of the gene pairs into the orientations was consistent with previous published observations in *Arabidopsis* (Ren et al., 2005). Tandem duplicate gene pairs were found to occur mostly orientated in parallel (83.52 %).

Table 3: The number and percentage of gene pairs expressed in the different orientations

Gene Pairs	Parallel		Divergent		Convergent		Total Gene Pairs
	No.	(%)	No	(%)	No	(%)	
All adjacent	11016	51.51	5169	24.17	5200	24.32	21385
Adjacent - td	10114	49.81	5075	25.19	5116	25.20	20305
Adjacent td only	902	83.52	94	8.70	84	7.78	1080

Table 4: Orientations of gene pairs expressed as a number and a percentage for R > 0.7.

Gene Pairs	Parallel		Divergent		Convergent		Total
	No.	(%)	No	(%)	No	(%)	Gene Pairs
All Adjacent	231	67.35	70	20.41	42	12.24	343
Adjacent -td	143	60.34	57	24.05	37	15.61	237
Adjacent td only	88	83.02	13	12.26	5	4.72	106

However, when highly co-expressed gene pairs (i.e. R > 0.7) were considered, there was a change in the distribution of gene pairs among the different gene pair orientations (Table 4). The gene pairs are no longer distributed approximately equally across the gene pair orientations. Instead, the majority of the adjacent gene pairs are arranged in parallel (67.35 %), followed by gene pairs divergently arranged (20.41 %) and then convergently arranged (12.24 %). The distribution among the gene pair orientations was similar when tandem duplicates were removed. Approximately sixty percent of those gene pairs are arranged in parallel, twenty-four percent in divergent orientation and sixteen percent in convergent orientation. Zhan and colleagues identified a similar distribution pattern for highly co-expressed *Arabidopsis* gene pairs (Zhan et al., 2006). Highly co-expressed tandem duplicates are still predominately orientated in parallel (83.02 %). Approximately three times more tandem duplicate gene pairs are divergently arranged compared to convergently arranged (12.26 % and 4.78 %, respectively).

Recently, it was reported that the zebrafish genome contains pre-dominantly parallel orientated co-expressed gene pairs (Ng et al., 2009). The authors explain this phenonomen as being a result of limited genome annotation or a feature unique to zebrafish. The Arabidopsis genome is well annotated and since our findings concur with those from a previous investigation (Zhan et al., 2006), it is possible that the distribution pattern of gene pair orientation is not limited to the zebrafish. The lack of divergently arranged gene pairs which are highly co-expressed would indicate that bi-directional promoters are not a dominant contribution to Arabidopsis gene co-expression. In mammals and humans, there is an abundance of bi-directional promoters (Trinklein et al., 2004) while in yeast, conflicting evidence suggests a minor (Tsai et al., 2007) or dominant occurance (Cohen et at., 2000).

Table 5: Intergenic distances for adjacent gene pairs

				R > 0.7		
Gene Pairs	Mean (bp)	Std error (bp)	Median (bp)	Mean (bp)	Std error (bp)	Median (bp)
All adjacent gene pairs	2996.954	36.63944	1331	3046.14	268.0647	1346
Adjacent gene pairs – td	3032.028	38.25603	1335	3094.975	343.6426	1255
Adjacent Gene pairs td only	2337.525	92.78326	1293	2936.953	404.668	1545.5

Close proximity of adjacent genes has also been studied for its influence on gene co-expression. The mean and median intergenic distances were calculated for the gene pairs of the three datasets (Table 5). Intergenic distance was defined as the shortest distance between two adjacent genes. If the genes overlapped, then the distance was defined as zero. The datasets with and without tandem duplicates have a similar distribution (see Figure 4) . The means lie far to the right (2996.956 by for all adjacent gene pairs and 3032.028 by for adjacent gene pairs that exclude tandem duplicates) compared to their medians (1331 by and 1335 bp, respectively). This indicates that most of the gene pairs for both datasets have short intergenic distances with fewer outlying gene pairs being seperated over long distances. This is also evident from Figure 4b and 4d, where the distribution of gene pairs with intergenic distance of less than 10 kbps are plotted. The mean intergenic distance for the tandem duplicate gene pairs

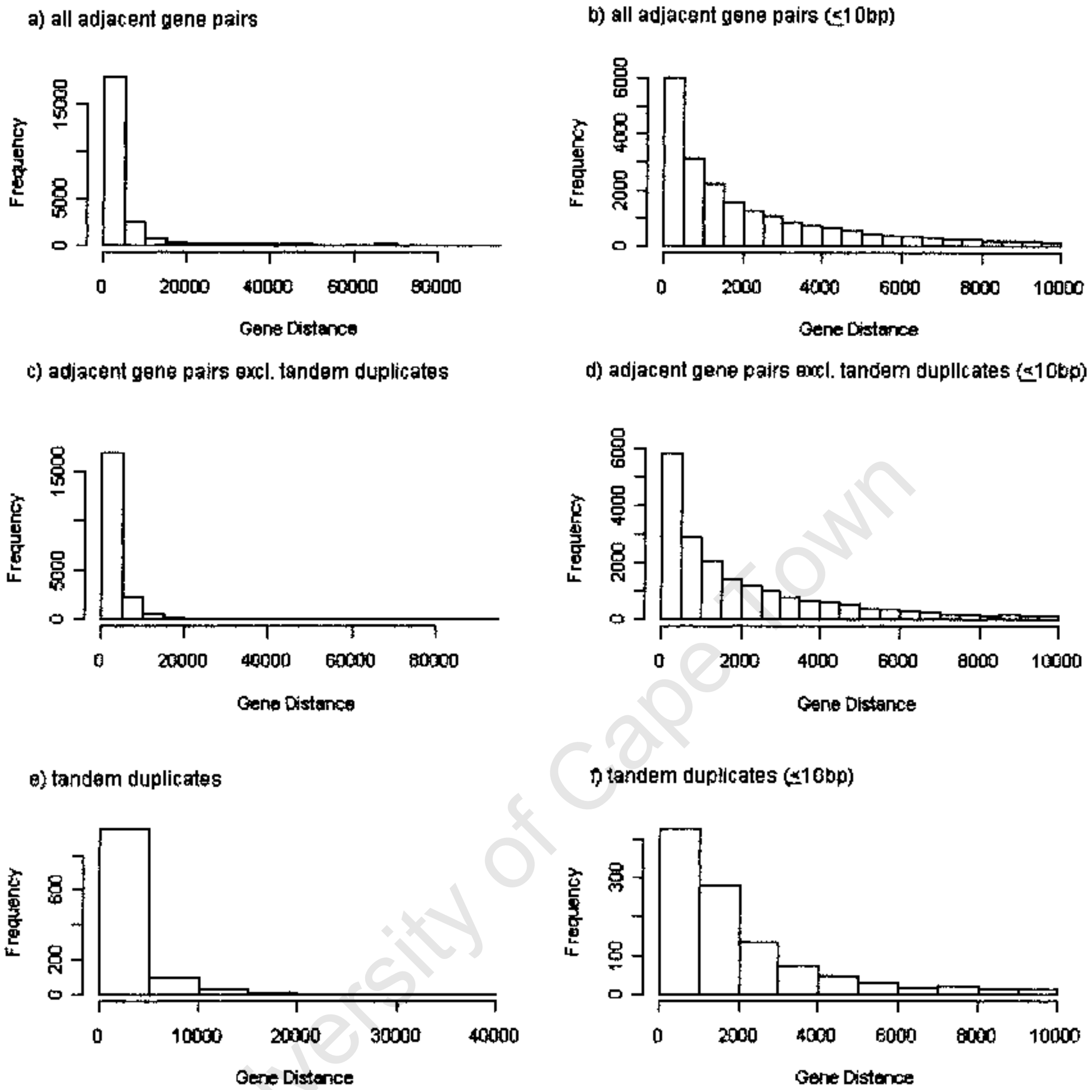


Figure 4: Histograms of the distribution of intergenic distances for gene pair datasets. a) all adjacent gene pairs, b) all adjacent gene pairs with intergenic distance s 10 kbp, c) adjacent gene pairs excluding tandem duplicates, d) adjacent gene pairs excluding tandem duplicates with intergenic distance s 10 kbp, e) adjacent tandem duplicates and f) adjacent tandem duplicates with intergenic distance s 10 kbp.

was less than for all the adjacent gene pairs or the adjacent gene pairs excluding tandem duplicates by 659 by and 694 by respectively.

In highly co-expressed gene pairs i.e where R > 0.7, the mean intergenic distance for the adjacent gene pairs excluding tandem duplicates is larger than the mean intergenic distance for all the adjacent gene pairs (see Table 5). But when comparing the median intergenic distance, adjacent gene pairs excluding tandem duplicates are, on average, closer than was the case for all the adjacent gene pairs (by 91 bp). In addition, the median intergenic distance for highly co-expressed tandem duplicates is higher than adjacent gene pairs with or without tandem duplicates. The large differences between the mean and median intergenic distances are likely to be a result of outliers in the data which have a strong influence on the mean, but much less influence over the median. Since highly co-expressed adjacent gene pairs (excluding tandem duplicates) are closer together than all adjacent pairs, the data suggest that short intergenic distance contributes to gene co-expression. Small intergenic distance has been described as a feature of gene co-expression in other eukaryotes. In yeast, fifteen percent of co-expressed adjacent pairs were found to occur within close proximity (within 1000bp) (Cohen et al., 2000). It was reported in C. *Elegans* that after excluding operons and tandem duplicates, co-expression was significant only within an intergenic distance of less than 20 kb (Lercher et al., 2003). Previous studies in Arabidopsis also indicated gene proximity as a factor of gene co-expression (Ren et al., 2005; Zhan et al., 2006).

To characterise the relationship between correlation and gene distance for highly co-expressed gene pairs, scatterplots were first plotted to provide a broad indication of the distribution (Figure 5). Co-expression of gene pairs could occur over large distances of up to 80 kbps and tandem duplicates were mostly close together with few scattered up to 35 kbps apart. Noticeably, most highly co-expressed gene pairs, positive or negative have intergenic distances of < 20 kbps.

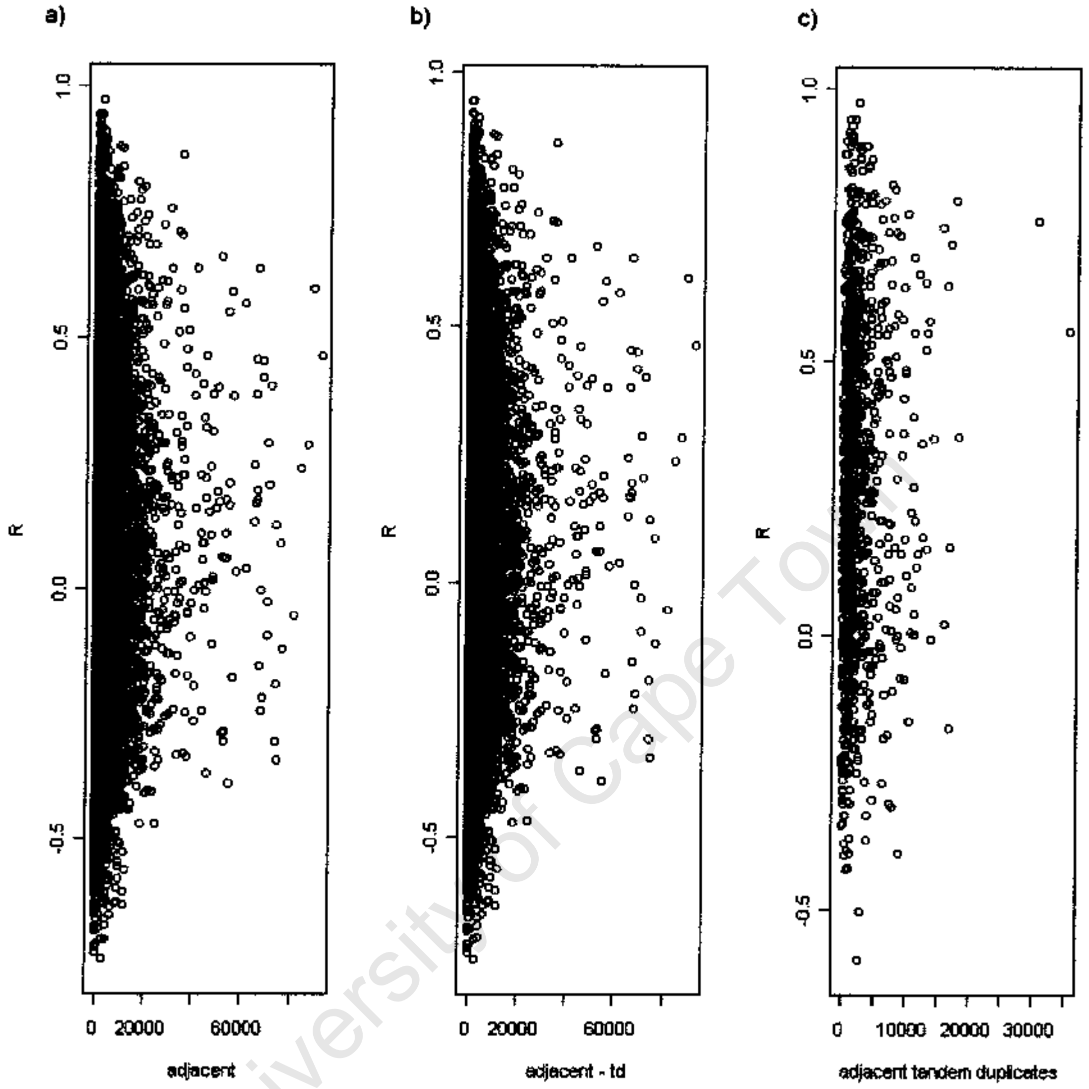


Figure 5: Scatterplots for R vs intergenic distance. a) all adjacent gene pairs, b) adjacent gene pairs excluding tandem duplicates and c) adjacent tandem duplicates.

To further investigate the impact of intergenic distance on gene pairs with regards to correlation, restricted cubic splines were plotted. Restricted cubic splines provide insights into trends that are difficult to see from the scatterplot. The models fitted to all adjacent gene pairs and adjacent excluding tandem duplicates were significant (Figure 6, p-value < 2.2e-16 and Figure 7, p-value < 8.006e-16, respectively) but not the tandem duplicate model (Figure 8, p-value < 0.069). For each gene set, the curve started as a sharp "blip" followed by gentle

slope. The sharp "blip" might suggest that for extremely close gene pairs a linear relationship with correlation and intergenic distance may exist. This would hold for all adjacent gene pairs and the adjacent gene pairs without tandem duplicates.

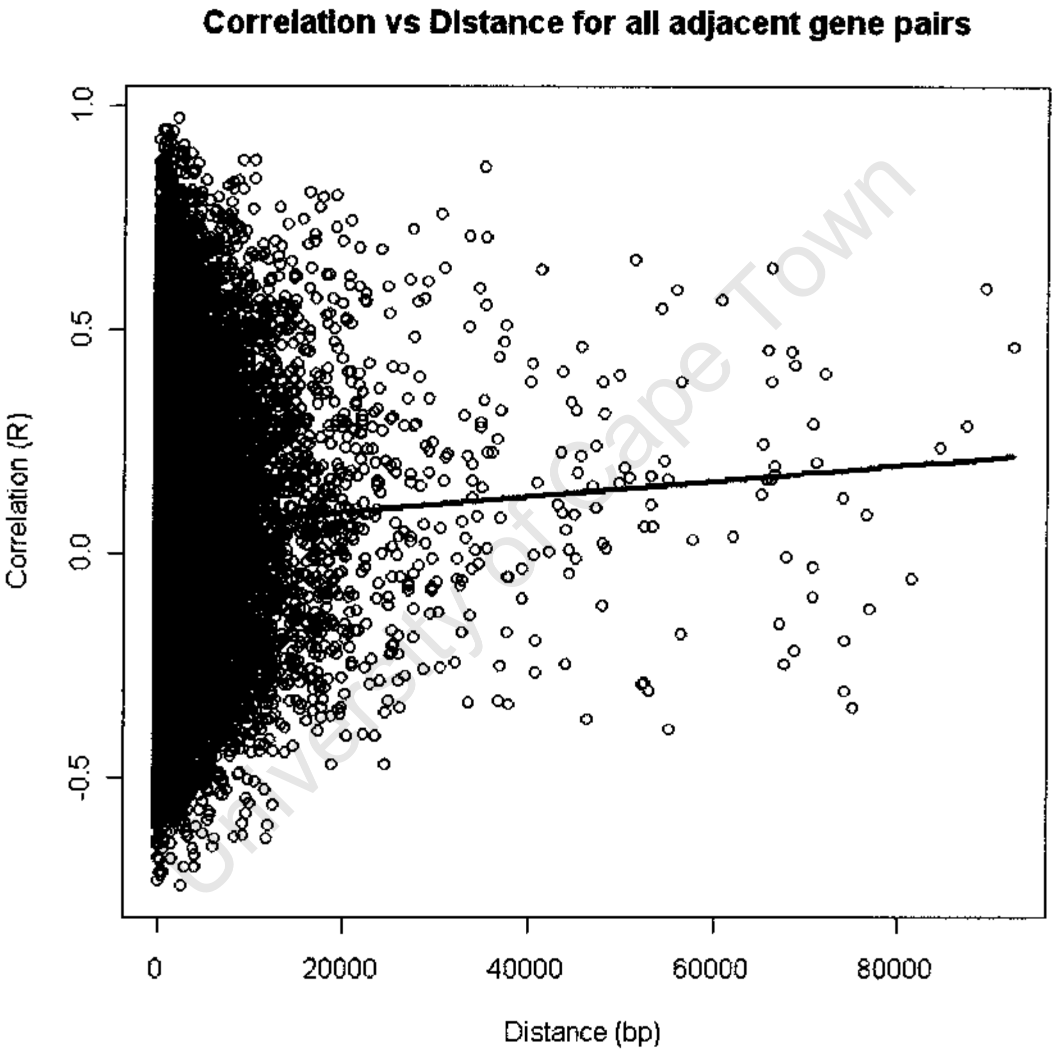


Figure 6: Restricted cubic spline plot of Correlation (R) vs Intergenic Distance for all adjacent gene pairs.

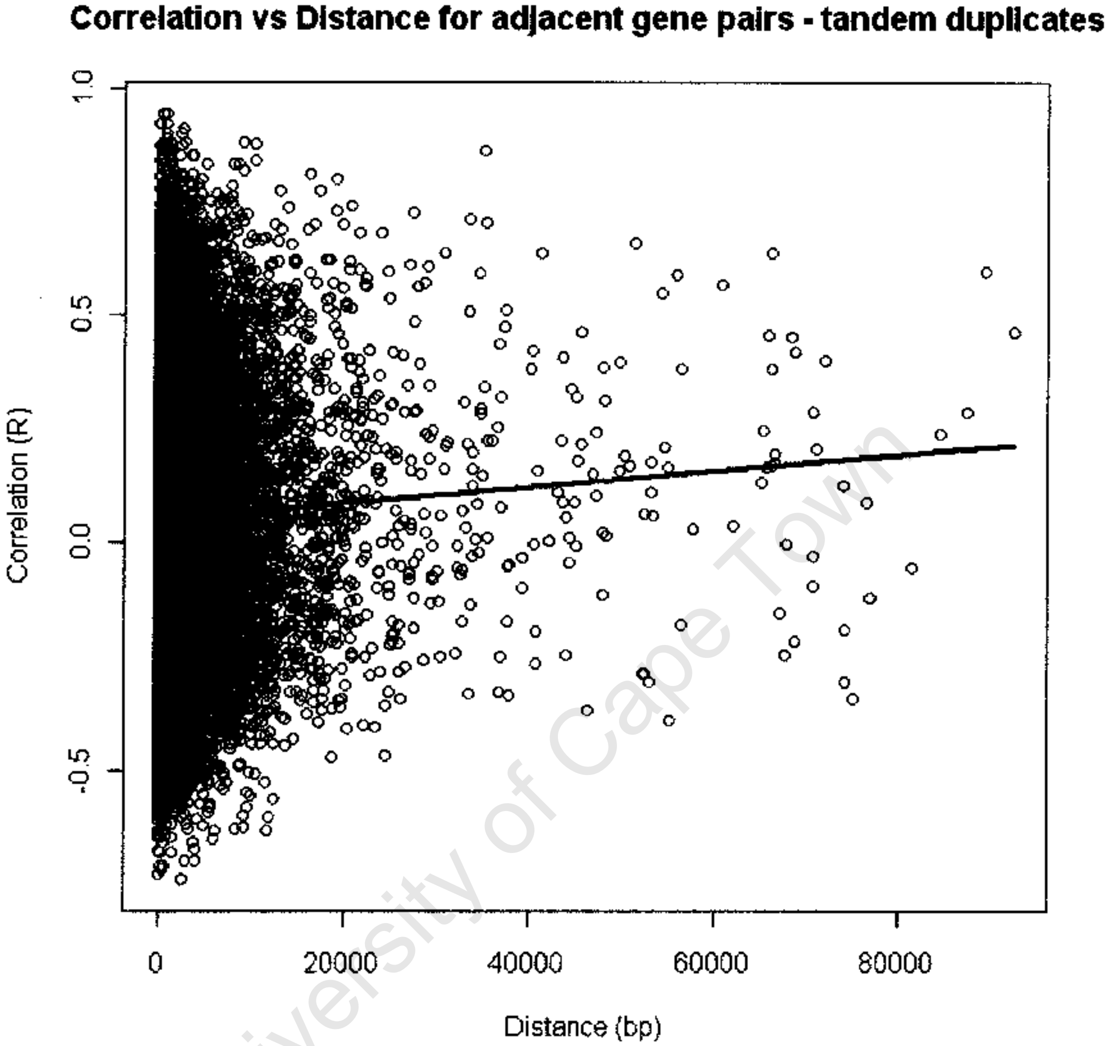


Figure 7: Restricted cubic spline plot of Correlation (R) vs Intergenic Distance for all adjacent gene pairs excluding tandem duplicates.

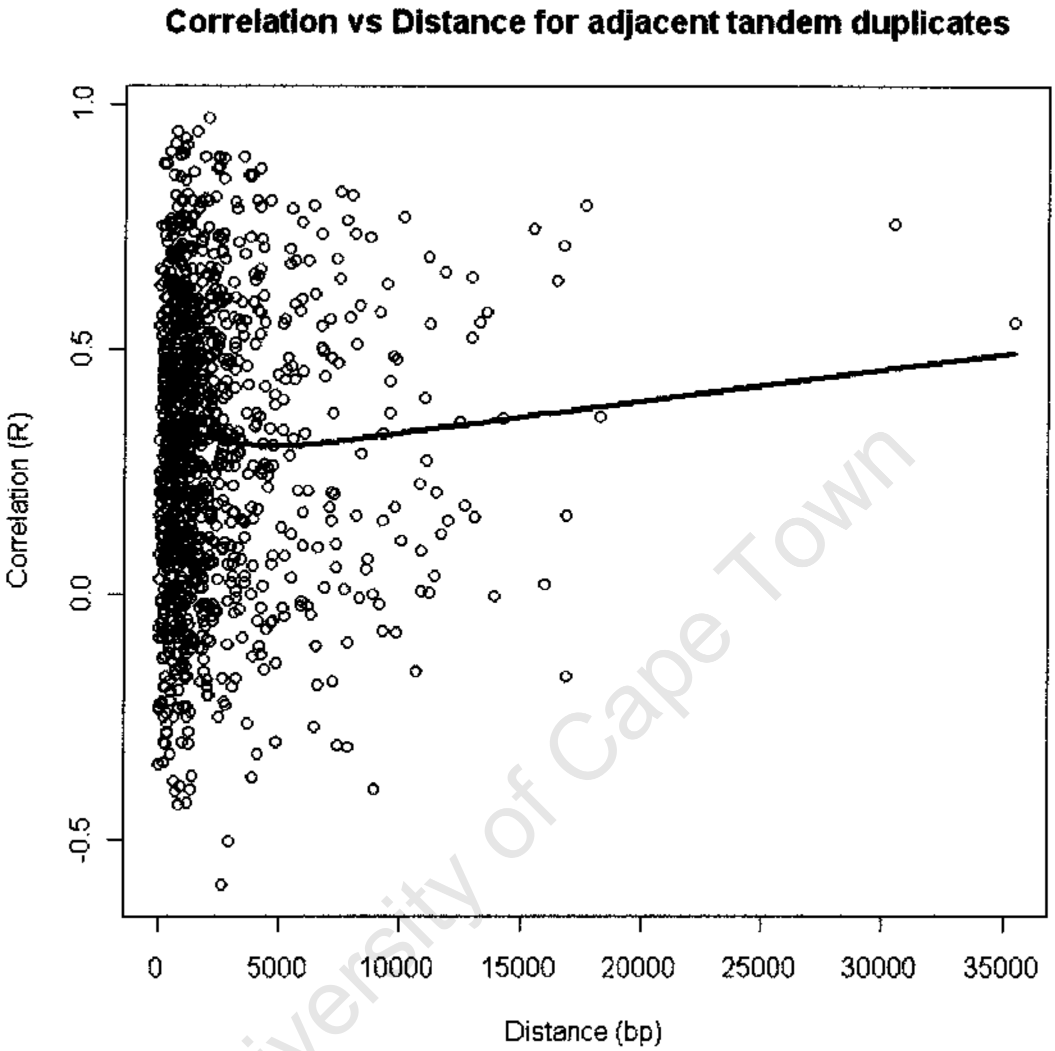


Figure 8: Restricted cubic spline plot of Correlation (R) vs Intergenic Distance for all adjacent tandem duplicates.

Restricted cubic splines were also plotted for each of the three gene pair orientations for adjacent gene pairs that excluded tandem duplicates. Both parallel (Figure 9) and convergent (Figure 10) exhibited curves that began with a "blip" and then rose into a gentle incline. The divergent curve behaved differently (Figure 11). There was a sharp drop followed by a gradual incline. This might indicate that there is an inverse relationship between distance and likelihood of co-expression for divergently arranged co-expressed gene pairs that lie close to each other. To probe further, restricted cubic splines were plotted for divergently arranged co-expressed gene pairs that were within ranges of 5000 by and 1000 by of each other. The proposed inverse relationship proved to be absent when examined at these ranges (Figure 12 and Figure 13). In contrast to our results, a previous investigation into the expression of neighbouring genes demonstrated that there is high correlation between close adjacent gene pairs but that this decreases as the intergenic distance increases (Cohen et al., 2000; Williams and Bowles, 2004). In adjacent yeast genes, it was clear that closely positioned genes were more likely to be highly co-expressed than gene pairs further apart but no linear relationship could be established (Cohen et al., 2000). While other studies supported our findings that no clear relationship could be described other than that close adjacent gene pairs are co-expressed (Fukuoka et al., 2004; Ren et al., 2005).

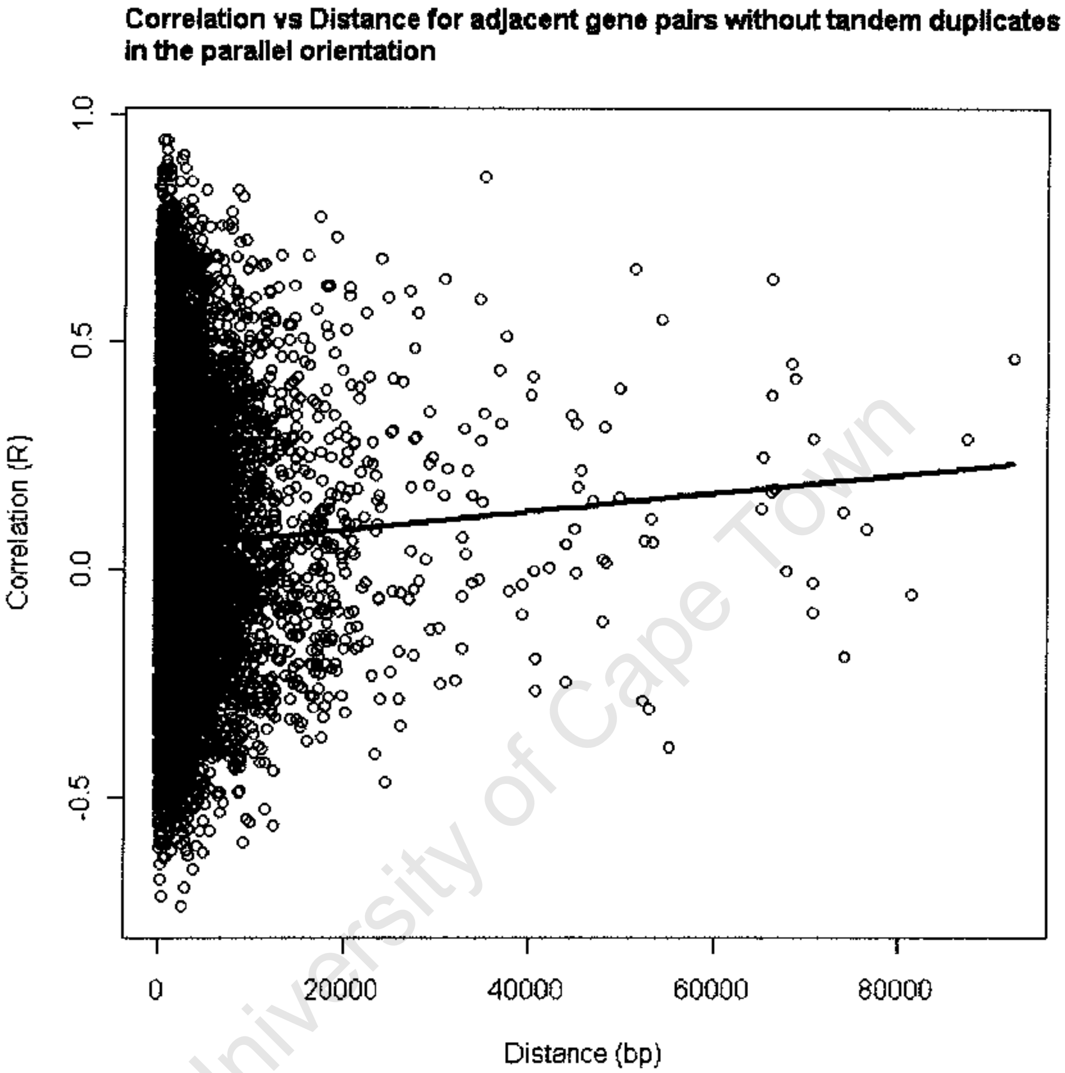


Figure 9: Restricted cubic spline plot of Correlation (R) vs Intergenic Distance for all adjacent gene pairs minus tandem duplicates that are in parallel orientation.

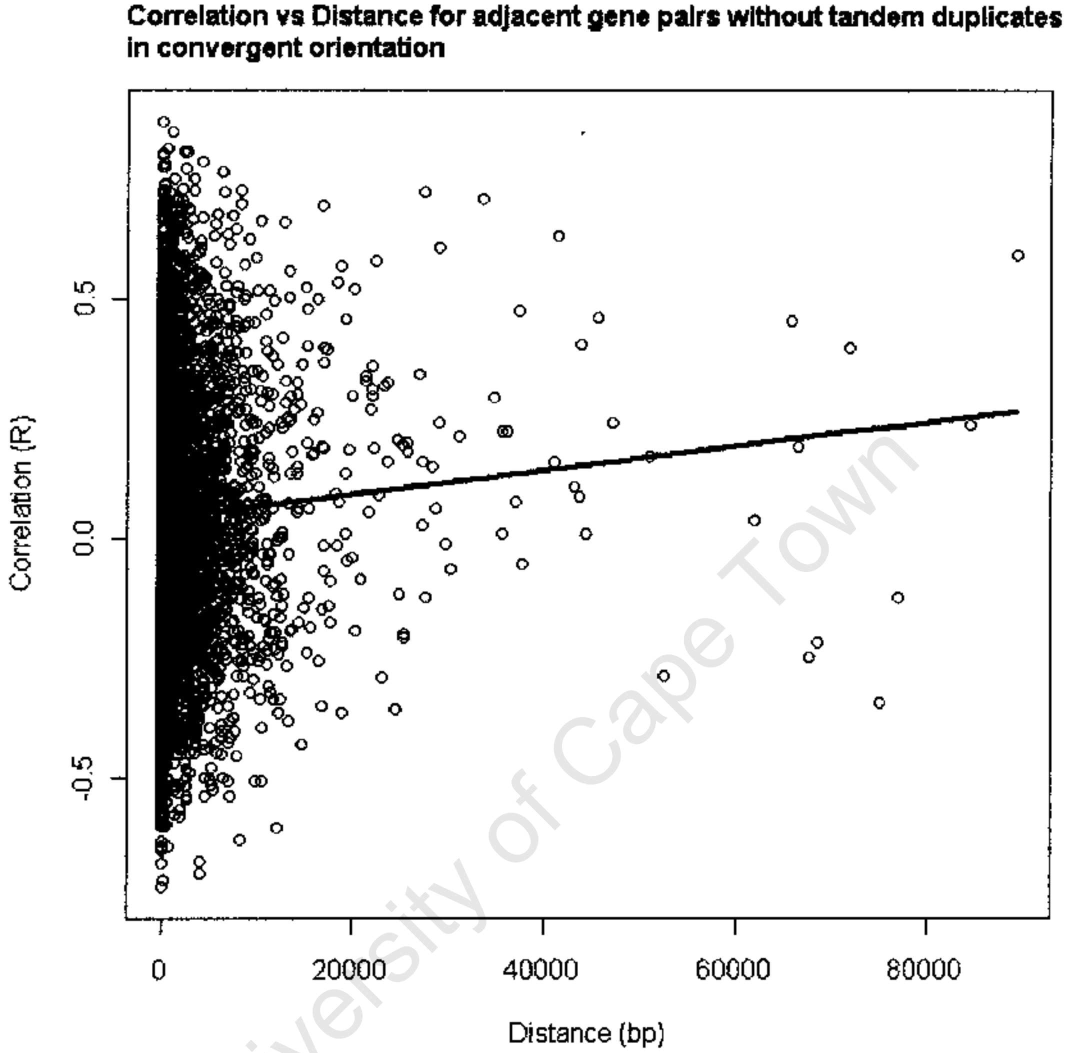


Figure 10: Restricted cubic spline plot of Correlation (R) vs Intergenic Distance for all adjacent gene pairs minus tandem duplicates that are convergently orientated.

3.3.2.2 Gene orientation and intergenic distance of chloroplast and mitochondria

Chloroplast and mitochondrial gene pair datasets were divided into the gene pair orientation categories, as shown in Table 13. The orientations of chloroplast and mitochondrial gene pairs were similar to other adjacent gene pairs. In the chloroplast, the gene pairs were distributed in approximately equal proportion for each orientation. Recall that parallel orientation is made up of two gene pair directions (-4-4+-44 Highly co-expressed chloroplast gene pairs were found to be predominantly in parallel (62.5 % in Table 15), followed by convergent (25 %) and divergent arrangement (12.5 %). Mitochondria gene pairs occur more in the parallel orientation (56.25 %) followed by convergent (31.25 %) and then divergent (12.5 %). Both of the highly co-expressed mitochondria gene pairs were arranged in parallel orientation (data not shown). The distribution of `Chloroplast' and 'Mitochondria' gene pairs into the orientations is consistent with the distribution of all the gene pairs from Table 4.

Table 13: Gene orientations of chloroplast and mitochondria gene pairs expressed as a number and a percentage.

Gene Pairs	Parallel		Divergent		Convergent		Total
	No.	(%)	No	(%)	No	(%)	Gene Pairs
Chloroplast adjacent	59	53.64	23	20.91	28	25.45	110
Mitochondria adjacent	9	56.25	2	12.50	5	31.25	16

Genes with common or related functions, such as common biological pathways, may sometimes be within close proximity to each other and form clusters (See Introduction 1.2.4). The intergenic distance of chloroplast and mitochondrial gene pairs is shown in Table 14. The medians lie closer to the left than the means. The median intergenic distance of 588.5 by and 436.5 by for chloroplast and mitochondria gene pairs, respectively, indicate that the genes in each pair are in much closer proximity to each other compared to the dataset from which

they were derived from i.e. adjacent gene pairs minus tandem duplicates (1335bp from Table 6).

Table 14: Gene distance for chloroplast and mitochondria gene pairs

Gene Pairs	Mean (bp)	Std error (bp)	Median (bp)
Chloroplast adjacent	1807.027	372.5438	588.5
Mitochondria adjacent	1691.75	515.0228	436.5

Table 15: Gene pair orientations and intergenic distance of chloroplast gene pairs expressed for R > 0.7.

	Gene pair orientation				Intergenic Distance		
Gene Pairs	Parallel No (%)	Divergent No (%)	Convergent No (%)	Total Gene Pairs	Mean (bp)	Std error (bp)	Median (bp)
Chloroplast	10 62.5	2 12.5	4 25	16	2990.313	2162.123	588.5

Table 16: Chloroplast gene pair distance for the different orientations

	All R			R > 0.7		
Gene pairs	Mean (bp)	Std error (bp)	Median (bp)	Mean (bp)	Std error (bp)	Median (bp)
Parallel	2620.932	665.5636	824	4302.2	3447.636	668.5
Divergent	1212.24	250.7592	947	1242	1103	1242
Convergent	532	176.6953	26.5	584.75	523.1605	95

Table 17: Mitochondria gene pair distance for the different orientations

Gene pairs	Mean (bp)	Std error (bp)	Median (bp)
Parallel	1856.556	716.9162	432
Divergent	2354	1256.247	1699.5
Convergent	314.3333	65.08797	277

Both chloroplast and mitochondria! gene pairs were found most close together in the convergent orientation (26.5 by and 277 by respectively in Tables 16 and 17). Some of the genes are located on opposite DNA strands and overlap, so the intergenic distance was set to zero. The adjacent gene pairs linked to chloroplast and mitochondrial functions lie closer together than other non-duplicate adjacent gene pairs (of which they are a subset) and are also closest in convergent orientation which is not the dominant orientation. This could suggest that organelle function may drive the genes to be located closer together, regardless of orientation. This would comply with a previous suggestion by Alexeyenko et. al (2006), that clustering of chloroplast and mitochondrial genes is caused by selection favouring organelle genes in close proximity.

Chapter 4: Concluding Remarks

Research into gene and genome function has lead to the development and use of many high-throughput methods to measure gene expression, among these, are microarrays. Various applications have been created to assist with the analysis of microarray data but few tools allow the analysis of gene co-expression in the genomic context. Genes are positioned in a non-random order on eukaryote genomes with some forming clusters while other genes interact over larger distances. CAGED was developed, using over 1700 Affymetrix Arabidopsis microarray experiments, to enable researchers to view gene co-expression within a genomic context. The image maps provide direct views of Arabidopsis gene pair co-expression across all five chromosomes. The examples provided demonstrate the potential applications of CAGED, such as visualization of previously established gene clusters, as in the thalianol pathway and identifying potential clusters, as in retrograde signaling.

While CAGED is simple and intuitive to use, there are a few limitations to note. Gene pairs were considered neighbouring for a distance of up to 100 000bp, which might exclude some gene pairs which are co-expressed as a result of long distance chromatin interactions. Also, CAGED was built with the then current Arabidopsis gene annotation and Affymetrix Arabidopsis microarray probe ID mapping. As further annotation updates are released, certain gene names might become obsolete. New methods of analyzing gene co-expression would also provide alternate insights.

A recent publication suggested that analyzing gene co-expression by the current method of evaluating Pearson correlation coefficients could not sufficiently characterize gene functional relationships (Kinoshita and Obayashi, 2009). A novel method was proposed in which the correlation coefficients of gene pairs are calculated using Principal Component Analysis (PCA). This provides a different perspective by grouping gene pair correlations into different components based on the degree of correlation. Then by subtracting the components and using GO annotation, the degree of a particular gene cluster co-expression, with respect to gene function can be understood. CAGED could be adapted by replacing the current Pearson correlation method with the formula based on PCA. Single linkage clustering could be added with an option to select a root mean square deviation (rmsd) threshold. The clusters could then be displayed on the image maps. To investigate biological function, selection of a cluster could produce a menu option which would allow the user to subtract components.

Factors which influence or drive gene co-expression is a topic of much discussion. Using the data underlying the CAGED tool, we investigated adjacent gene pair co-expression and the intergenic distance and gene pair orientations. Approximately two percent of the adjacent nonduplicate genes were highly co-expressed. Parallel orientated gene pairs were found be the most abundant and also the closest together. Bi-directional promoters did not occur in a high enough frequency to be the main cause of gene co-expression as divergently orientated gene pairs were not dominant. Similar results were determined among the chloroplast and mitochondrial genes but these gene pairs were found to lie closer together. This suggests that organelle function may cause genes in close proximity to be co-expressed. A comparativegenomics approach could provide further insights by comparing co-expressed orthologous gene pairs that may be functionally related (Daub and Sonnhammer, 2008). Investigation of change in co-expression in gene domains to environmental stresses and observation of influential factors may provide further insight. Finally, the recent mapping of the Arabidopsis epigenome (Lister et al., 2008) could be used to unravel the chromatin regulatory aspects of gene co-expression.

Supplementary Material

Supplementary Table 1: The number and percentage of gene pairs expressed for the parallel orientations (Forward and Reverse).

Gene pairs	Forward	Reverse	Total
All Adjacent	5476 (25.6%)	5540 (25.9%)	11016 (51.51%)
Adjacent – td	5035 (24.76%)	5079 (25.01%)	10114 (49.81%)
Tandem Duplicates	441 (40.83%)	461 (42.69%)	1080 (83.52%)

Reference List

Achaz,G., Coissac,E., Netter,P., and Rocha,E.P. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics *164,* 1279-1289.

Alexeyenko,A., Millar,A.H., Whelan,J., and Sonnhammer,E.L. (2006). Chromosomal clustering of nuclear genes encoding mitochondria! and chloroplast proteins in Arabidopsis. Trends Genet. *22,* 589-593.

Batada,N.N., Urrutia,A.O., and Hurst,L.D. (2007). Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. 23, 480-484.

Belfaiza,J., Parsot,C., Martel,A., de la Tour,C.B., Margarita,D., Cohen,G.N., and Saint-Girons,I. (1986). Evolution in biosynthetic pathways: two enzymes catalyzing consecutive steps in methionine biosynthesis originate from a common ancestor and possess a similar regulatory region. Proc. Natl. Acad. Sci. U. S. A 83, 867-871.

Ben Shahar,Y., Nannapaneni,K., Casavant,T.L., Scheetz,T.E., and Welsh,M.J. (2007). Eukaryotic operon-like transcription of functionally related genes in Drosophila. Proc. Natl. Acad. Sci. U. S. A *104,* 222-227.

Blumenthal,T., Evans,D., Link,C.D., Guffanti,A., Lawson,D., Thierry-Mieg,J., Thierry-Mieg,D., Chiu,W.L., Duke,K., Kiraly,M., and Kim,S.K. (2002). A global analysis of Caenorhabditis elegans operons. Nature *417,* 851-854.

Boivin,K., Acarkan,A., Mbulu,R.S., Clarenz,O., and Schmidt,R. (2004). The Arabidopsis genome sequence as a tool for genome analysis in Brassicaceae. A comparison of the Arabidopsis and Capsella rubella genomes. Plant Physiol *135,* 735-744.

Bolstad,B.M., Irizarry,R.A., Astrand,M., and Speed,T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. *19,* 185-193.

Boutanaev,A.M., Kalmykova,A.I., Shevelyov,Y.Y., and Nurminsky,D.I. (2002). Large clusters of co-expressed genes in the Drosophila genome. Nature *420,* 666-669.

Bowers,J.E., Chapman,B.A., Rong,J., and Paterson,A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature *422,* 433-438.

Clifton,R., Lister, R., Parker,K.L., Sappl,P.G., Elhafez,D., Millar,A.H., Day,D.A., and Whelan,J. (2005). Stress-induced co-expression of alternative respiratory chain components in Arabidopsis thaliana. Plant Mol. Biol. *58,* 193-212.

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

Coppe,A., Danieli,G.A., and Bortoluzzi,S. (2006). REEF: searching REgionally Enriched Features in genomes. BMC. Bioinformatics. 7, 453.

Dandekar,T., Snel,B., Huynen,M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23, 324-328.

Daub,C.O. and Sonnhammer,E.L. (2008). Employing conservation of co-expression to improve functional inference. BMC. Syst. Biol. *2,* 81.

Davis,R.E. and Hodgson,S. (1997). Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in Schistosoma mansoni. Mol. Biochem. Parasitol. *89,* 25-39.

Eszterhas,S.K., Bouhassira,E.E., Martin,D.I., and Fiering,S. (2002). Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. Mol. Cell Biol. *22,* 469-479.

Fernandez,A.P. and Strand,A. (2008). Retrograde signaling and plant stress: plastid signals initiate cellular stress responses. Curr. Opin. Plant Biol. *11,* 509-513.

Field, B. and Osbourn,A. E. (2008). Metabolic diversification--independent assembly of operon-like gene clusters in different plants. Science *320,* 543-547.

Fisher, R. A. The Genetical Theory of Natural Selection. 1930. Oxford, Clarendon Press.
Ref Type: Generic

Fu,H., Park,W., Yan,X., Zheng,Z., Shen,B., and Dooner,H.K. (2001). The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. Proc. Natl. Acad. Sci. U. S. A *98,* 8903-8908.

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

Ganko,E.W., Meyers,B.C., and Vision,T.J. (2007). Divergence in expression between duplicated genes in Arabidopsis. Mol. Biol. Evol. 24, 2298-2309.

Gerdes,S.Y., Scholle,M.D., Campbell,J.W., Balazsi,G., Ravasz,E., Daugherty,M.D., Somera,A.L., Kyrpides,N.C., Anderson,I., Gelfand,M.S., Bhattacharya,A., Kapatral,V., D'Souza,M., Baev,M.V., Grechkin,Y., Mseeh,F., Fonstein,M.Y., Overbeek,R., Barabasi,A.L., Oltvai,Z.N., and Osterman,A.L. (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J. Bacteriol. *185,* 5673-5684.

Gu,Z., Nicolae,D., Lu,H.H., and Li,W.H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. *18,* 609-613.

Gu,Z., Rifkin,S.A., White,K.P., and Li,W.H. (2004). Duplicate genes increase gene expression diversity within and between species. Nat. Genet. *36,* 577-579.

Herr,D.R. and Harris,G.L. (2004). Close head-to-head juxtaposition of genes favors their coordinate regulation in Drosophila melanogaster. FEBS Left. 572, 147-153.

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

HOROWITZ,N.H. and NETZENBERG,R.L. (1965). Biochemical Aspects of Genetics. Annu. Rev. Biochem. *34,* 527-564.

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

Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U., and Speed,T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 4, 249-264.

Itoh,T., Takemoto,K., Mori,H., and Gojobori,T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol. Biol. Evol. *16,* 332-346.

Jacob,F., Perrin,D., Sanchez,C., and Monod,J. (1960). [Operon: a group of genes with the expression coordinated by an operator]. C. R. Hebd. Seances Acad. Sci. *250,* 1727-1729.

Jen,C.H., Manfield,I.W., Michalopoulos,I., Pinney,J.W., Willats,W.G., Gilmartin,P.M., and Westhead,D.R. (2006). The Arabidopsis co-expression tool (ACT): a WVVW-based tool and database for microarray-based gene expression analysis. Plant J. *46,* 336-348.

Johnson,P.J., Kooter,J.M., and Borst,P. (1987). Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. Cell *51,* 273-281.

Jones,B.K., Monks,B.R., Liebhaber,S.A., and Cooke,N.E. (1995). The human growth hormone gene is regulated by a multicomponent locus control region. Mol. Cell Biol. *15,* 7010-7021.

Khaitovich,P., Weiss,G., Lachmann,M., Hellmann,I., Enard,W., Muetzel,B., Wirkner,U., Ansorge,W., and Paabo,S. (2004). A neutral model of transcriptome evolution. PLoS. Biol. *2,* E132.

Kim,J., Shiu,S.H., Thoma,S., Li,W.H., and Patterson,S.E. (2006). Patterns of expansion and expression divergence in the plant polygalacturonase gene family. Genome Biol. 7, R87.

Kinoshita,K. and Obayashi,T. (2009). Multi-dimensional correlations for gene coexpression and application to the large-scale data of Arabidopsis. Bioinformatics. 25, 2677-2684.

Kliebenstein,D.J. (2008). A role for gene duplication and natural variation of gene expression in the evolution of metabolism. PLoS. ONE. 3, e1838.

Koonin,E.V., Mushegian,A.R., and Rudd,K.E. (1996). Sequencing and analysis of bacterial genomes. Curr. Biol. *6,* 404-416.

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

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

Lawrence,J.G. and Roth,J.R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics *143,* 1843-1860.

Lee,J.M. and Sonnhammer,E.L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. *13,* 875-882.

Lee,S.J. (1991). Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. Proc. Natl. Acad. Sci. U. S. A *88,* 4250-4254.

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

Liao,B.Y. and Zhang,J. (2006). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. 23, 530-540.

Liao,B.Y. and Zhang,J. (2008). Coexpression of linked genes in Mammalian genomes is generally disadvantageous. Moi. Biol. Evol. *25,* 1555-1565.

Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H., and Ecker,J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133,* 523-536.

Liu, H., Sachidanandam,R., and Stein, L. (2001). Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order. Genome Res. *11,* 2020-2026.

Lynch,M. and Conery,J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290,* 1151-1155.

Maclean,D., Jerome,C.A., Brown,A.P., and Gray,J.C. (2008). Co-regulation of nuclear genes encoding plastid ribosomal proteins by light and plastid signals during seedling development in tobacco and Arabidopsis. Plant Mole Biol. *66,* 475-490.

Maere,S., De,B.S., Raes,J., Casneuf,T., Van,M.M., Kuiper,M., and Van de,P.Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. U. S. A *102,* 5454-5459.

Makova,K.D. and Li,W.H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. *13,* 1638-1645.

Maniatis,T., Fritsch,E.F., Lauer,J., and Lawn,R.M. (1980). The molecular genetics of human hemoglobins. Annu. Rev. Genet. *14,* 145-178.

Mayor,L.R., Fleming,K.P., Muller,A., Balding,D.J., and Sternberg,M.J. (2004). Clustering of protein domains in the human genome. J. Moi. Biol. *340,* 991-1004.

Mushegian,A.R. and Koonin,E.V. (1996). Gene order is not conserved in bacterial evolution. Trends Genet. *12,* 289-290.

Mutwil,M., Obro,J., Willats,W.G., and Persson,S. (2008). GeneCAT--novel webtools that combine BLAST and co-expression analyses. Nucleic Acids Res. *36,* W320-W326.

Ng,Y.K., Wu,W., and Zhang,L. (2009). Positive correlation between gene coexpression and positional clustering in the zebrafish genome. BMC. Genomics *10,* 42.

Pal,C. and Hurst,L.D. (2004). Evidence against the selfish operon theory. Trends Genet. *20,* 232-234.

Persson,S., Wei,H., Milne,J., Page,G.P., and Somerville,C.R. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc. Natl. Acad. Sci. U. S. A *102,* 8633-8638.

Pesaresi,P., Masiero,S., Eubel,H., Braun,H.P., Bhushan,S., Glaser,E., Salamini,F., and Leister,D. (2006). Nuclear photosynthetic gene expression is synergistically modulated by rates of protein synthesis in chloroplasts and mitochondria. Plant Cell *18,* 970-991.

Price,M.N., Dehal,P.S., and Arkin,A.P. (2008). Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol. *9,* R4.

Price,M.N., Huang,K.H., Arkin,A.P., and Alm,E.J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res. *15,* 809-819.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2008. Vienna, Austria.
Ref Type: Computer Program

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

Ren,X.Y., Stiekema,W.J., and Nap,J.P. (2007). Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains. Plant Mol. Biol. *65,* 205-217.

Riley,M.C., Clare,A., and King,R.D. (2007). Locational distribution of gene functional classes in Arabidopsis thaliana. BMC. Bioinformatics. 8, 112.

Rocha,E. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. *10,* 393-395.

Rocha,E.P. and Danchin,A. (2003). Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat. Genet. *34,* 377-378.

Rocha,E.P., Sekowska,A., and Danchin,A. (2000). Sulphur islands in the Escherichia coli genome: markers of the cell's architecture? FEBS Lett. *476,* 8-11.

Rossmann,M.G., Moras,D., and Olsen,K.W. (1974). Chemical and biological evolution of nucleotide-binding protein. Nature *250,* 194-199.

Satou,Y., Hamaguchi,M., Takeuchi,K., Hastings,K.E., and Satoh,N. (2006). Genomic overview of mRNA 5'-leader trans-splicing in the ascidian Ciona intestinalis. Nucleic Acids Res. 34, 3378-3388.

Schranz,M.E., Lysak,M.A., and Mitchell-Olds,T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. *11,* 535-542.

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

Singer,G.A., Lloyd,A.T., Huminiecki,L.B., and Wolfe,K.H. (2005). Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol. Biol. Evol. 22, 767-775.

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

Spieth,J., Brooke,G., Kuersten,S., Lea,K., and Blumenthal,T. (1993). Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell 73, 521-532.

Srinivasasainagendra,V., Page,G.P., Mehta,T., Coulibaly,I., and Loraine,A.E. (2008). CressExpress: a tool for large-scale mining of expression data from Arabidopsis. Plant Physiol *147,* 1004-1016.

Stahl,F.W. and Murray,N.E. (1966). The evolution of gene clusters and genetic circularity in microorganisms. Genetics 53, 569-576.

Takai,D. and Jones,P.A. (2004). Origins of bidirectional promoters: computational analyses of intergenic distance in the human genome. Mol. Biol. Evol. 21, 463-467.

Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E., and Koonin,E.V. (1996). Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coll. Curr. Biol. *6,* 279-291.

Trinklein,N.D., Aldred,S.F., Hartman,S.J., Schroeder,D.I., Otillar,R.P., and Myers,R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res. *14,* 62-66.

Trivedi,P., Edwards,J.W., Wang,J., Gadbury,G.L., Srinivasasainagendra,V., Zakharkin,S.O., Kim,K., Mehta,T., Brand,J.P., Patki,A., Page,G.P., and Allison,D.B. (2005). HDBStatt a platform-independent software suite for statistical analysis of high dimensional biology data. BMC. Bioinformatics. 6, 86.

Tsai,H.K., Su,C.P., Lu,M.Y., Shih,C.H., and Wang,D. (2007). Co-expression of adjacent genes in yeast cannot be simply attributed to shared regulatory system. BMC. Genomics 8, 352.

Vision,T.J., Brown,D.G., and Tanksley,S.D. (2000). The origins of genomic duplications in Arabidopsis. Science *290,* 2114-2117.

Wang,X., Shi,X., Li,Z., Zhu,Q., Kong,L., Tang,W., Ge,S., and Luo,J. (2006). Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. BMC. Bioinformatics. 7, 447.

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

Yerushalmi,U. and Teicher,M. (2007). Examining emergence of functional gene clustering in a simulated evolution. Bull. Math. Biol. *69,* 2261-2280.

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