1-1-2007

Organization and evolution of information within eukaryotic genomes.

Matthew Graham Links
University of Windsor

Organization and Evolution of Information within Eukaryotic Genomes

by

Matthew Graham Links

A Thesis
Submitted to the Faculty of Graduate Studies and Research
through Biological Sciences
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2007

Your file Votre référence
ISBN: 978-0-494-35043-0
Our file Notre référence
ISBN: 978-0-494-35043-0

ABSTRACT

Much of the bioinformatics research to date has focused on the identification and classification of gene structure, content, and function derived from a model organism *in silico* and the subsequent application of patterns *in vivo* to analyze additional species. Through the examination of gene length it is shown here that the distribution of gene length is conserved in a taxonomically meaningful fashion. It is further demonstrated that the mechanisms by which a linkage group encodes information may serve to optimize the density of information carried by fungal linkage groups. A novel sequence feature, a proposed site of DNA methylation, is shown in this work to be identified *a priori* with the subsequent feature distributions being taxonomically significant. Lastly, presented is a reverse genetics system for *Arabidopsis thaliana* which would enable the functional assessment of novel sequence features such as the putative DNA methylation targets described in this work.

DEDICATION

Soli Deo Gloria

ACKNOWLEDGEMENTS

I thank Janet Hill and Sean Hemmingsen for collaboration on the *CPN60* work. Genome Canada and Genome Prairie supported portions of the work presented here through funding of the FGAS project and thereby the development of the APED EST portal software. Luke McCarthy collaborated on the development of APED. I am indebted to Agriculture and Agri-Food Canada (Harrow Research Centre: Kangfu Yu, Margaret Haffner, and Vaino Poysa) who collaborated in the development of the Fast neutron (Fn) irradiated population of *Arabidopsis thaliana* mutants described in this work. I am grateful to Jay brogan for assistance with harvesting of the Fn population and organization of seed stocks, and to Xiaobo Lu for assistance with the organization of seed stocks, genomic DNA extractions for the FN population. Finally, I thank my supervisor Dr. William Crosby for unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

GENOME PROFILING ON THE BASIS OF CODING GENE LENGTH

Introduction

The technological success of the human genome sequencing effort was dependant upon the ability to sequence nucleic acids by high throughput methods. While it took four years to obtain the first billion bases of the human genome only four additional months were needed to double the acquired data [1]. As a collaborative project the sequencing of the human genome required the coordination of international labs in order to provide the collaborative capacity to fully sequence the human genome. The increase in the capacity which enabled the throughput necessary to complete the sequence of the human genome came as a result of technological advances and a reduction of the costs associated with DNA sequencing.

To date most publicly available DNA sequences were derived by a technique developed by Frederick Sanger in the 1970s [2,3]. Sanger's method relies on the enzymatic activity of DNA dependant DNA polymerase to incorporate inhibitors, di-deoxynucleotides (ddNTPs), which terminated newly synthesized DNA strands. It is through the use of ddNTPs in DNA replication which allows for the creation of a set of DNA fragments which are terminated in a nucleotide dependant fashion. In four reactions, one for each nucleotide moiety, a DNA polymerase was used to generate a series of termination products through the incorporation of deoxynucleotides (dNTPs) and the termination of synthesis through the incorporation of a specific P^{32}-radio-labeled ddNTP. The four reactions, each representing the termination products ending in a

distinct ddNTP, were separated by gel electrophoresis and the DNA sequence was visualized on x-ray film by autoradiography.

The major limiting factor of Sanger sequencing is its reliance on a DNA dependent DNA polymerase which in turn requires a sequence-specific primer. Thus some portion of the molecule being sequenced must be known in order to design the sequencing primer. In cases where the DNA molecule being sequenced is in a known cloning vector then part of the cloning vector sequence is commonly used for designing a sequencing primer. If no sequence data is available a restriction digest can be used to ensure a known sequence for the 5' end of the DNA template. Thus the sequence of the restriction site of the endonuclease serves as a sequence-specific primer site.

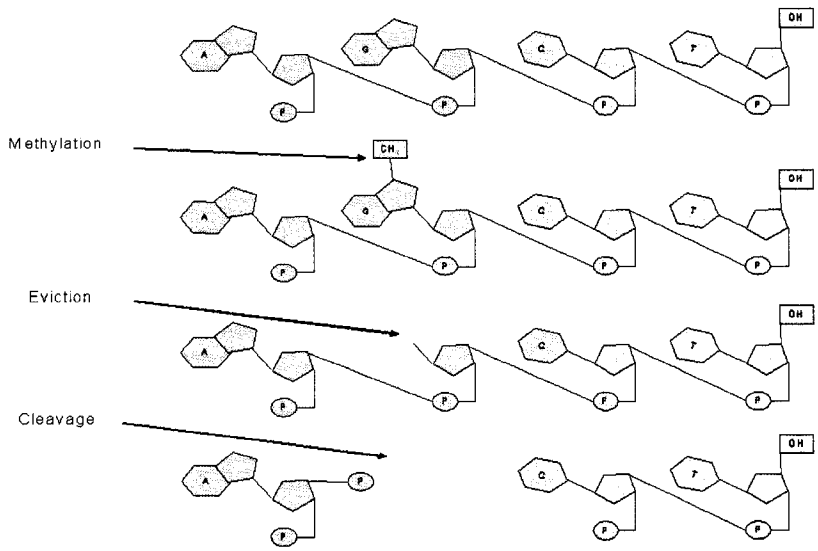

Figure 1 Schematic of Maxam-Gilbert sequencing.

In parallel with the aforementioned Sanger methodology, Maxam and Gilbert described in 1977 another method for sequencing radio-labeled DNA through nucleotide dependent chemical degradation [4,5]. Similar to Sanger sequencing, Maxam and Gilbert relied on the exploitation of four reactions, each comprised by fragmented products terminated by a unique nucleotide moiety. Chemical degradation was performed on DNA

radio-labeled with P^{32} at its 5' end, through three sequential chemical reactions: base modification, eviction, and strand cleavage of the resulting abasic site (Figure 1). In order to achieve nucleotide specific base pair modification, the nucleotide moiety of interest was selectively methylated. The methylated base was then evicted by breaking the bond to its sugar through heat and piperidine. Subsequent cleavage of the DNA strand was achieved through further heat and alkali conditions. Depending on the reagents used selective nucleotide moieties were methylated and subsequently evicted to form nucleotide specific abasic sites. For example di-methyl sulphate was used to methylate guanine residues whereas formic acid methylates both adenine and guanine; therefore treatment with di-methyl sulphate results in only guanine sites being abasic whereas formic acid caused all Purine sites to be abasic. Subsequent analysis of autoradiograph bands produced by formic acid treatment which did not correspond to bands in a di-methyl sulphate treatment were used to identify the adenine sites in the original DNA template. A similar technique was used to identify thymine terminated fragments through the comparison between hydrazine and hydrazine + sodium chloride treatments. The degradation products for each nucleotide moiety were separated by gel electrophoresis and visualized on x-ray film by autoradiography. DNA sequences were then read 5' to 3' from the bottom of the autoradiograph. The major disadvantages to the Maxam and Gilbert method of DNA sequencing were the use of hazardous chemicals, sample processing required for each nucleotide specific treatment and the subsequent low throughput. Variants of the Sanger sequencing method which use non-radioactive labels dominate modern DNA sequencing protocols. However, the Maxam and Gilbert technique still remains useful in modern labs for at least two scenarios: when no prior

sequencing knowledge exists upon which a primer could be designed, or to identify methylation patterns in DNA.

A major advance to the original Sanger sequencing protocol which enabled high throughput sequencing was the use of multiple labeled dNTP moieties. The reliance on a sole radiolabel meant that Sanger sequencing required four separate reactions, one per ddNTP moiety. In order to facilitate the automation of Sanger sequencing it has been common practice to use four fluorescent dyes, one per ddNTP moiety. In reactions with multiple labeled ddNTP moieties present, the differential fluorescence between ddNTPs allows the identification of sequential nucleotides. Modern high throughput sequencing employs a single reaction as opposed to the four reactions required by the original Sanger sequencing protocol [2,3]. At present the common dyes used in automated sequencing are rhodamine derivatives which have distinct excitation and emission peaks: dR110 - blue, dR6G - green, dTAMRA - yellow, and dROX - red. Depending on the particular application either the sequencing primer or ddNTPs can be labeled. In practice labeled ddNTPs produce weaker fluorescence when terminating longer fragments due to there being an inverse relation with product length and frequency. Therefore when fragment length is being optimized in a sequencing protocol a labeled sequencing primer is commonly employed.

Both the Sanger and Maxam and Gilbert DNA sequencing methods rely on the separation of DNA on slab gels by electrophoresis. While the early versions of automated DNA sequencers used polyacrylamide slab gels, it was the use of capillary electrophoresis (CE) which made high throughput sequencing a reality and enabled the completion of the human genome sequencing project in a timely fashion [1]. CE offers a

series of advantages over slab gel electrophoresis: a capillary pore size of < 100μm allows the use of higher electrical fields as the capillaries are better able to dissipate heat due to their high surface area to volume ratio as compared to slab gels, increased electrical fields positively affect throughput, and sample loading into the capillaries is much simpler than the preparation of slab gels [6]. Modern applications of CE technology group a series of capillaries into an array allowing multiple samples to be resolved at once.

The capacity provided through modern capillary sequencers has accelerated the acquisition large volumes of sequence data over recent years. Coincident with the explosion of DNA sequencing by public sequencing efforts, there has arisen increasing pressure to ensure that DNA sequence data derived from publicly funded programs be made publicly available. In order to facilitate the world-wide distribution of sequence data, centralized repositories were developed to store and disseminate sequence data via the Internet. Since the inception of CE sequencing methods there has been an exponential rise in the amount of publicly available sequence data as seen in the growth of GenBank (Figure 2).

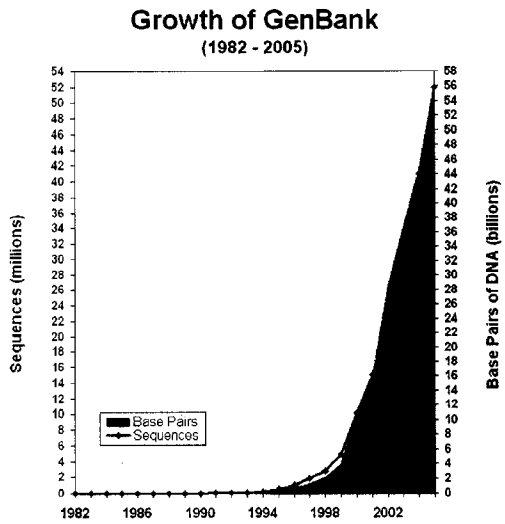

Figure 2 Growth of Genbank http://www.ncbi.nlm.nih.gov/Genbank/genbankgrowth.jpg

With the increased recognition of the utility of nucleic acid sequencing for diverse biological applications there has been a corresponding growth in the demand for sequencing capacity and new technologies are actively being developed to facility extremely high throughput sequencing. Pyrosequencing is based on three tightly coupled chemical reactions (Equation 1): the production of Pyrophosphate (PP_i) during DNA synthesis through the incorporation of a dNTP by DNA polymerase, the production of ATP through the consumption of PP_i by ATP sulphurylase, and the activity of luciferase to produce light through the consumption of ATP [7].

$$(DNA)_n + dNTP \xrightarrow{DNApolymerase} (DNA)_{n+1} + PPi$$
$$PPi + APS \xrightarrow{ATPsulphurylase} ATP + SO_4^{-2}$$
$$ATP + luciferin + O_2 \xrightarrow{luciferase} AMP + PPi + CO_2 + photon$$

Equation 1 Reactions involved in Pyrosequencing

Due to the interdependence of luciferase, ATP sulphurylase, and DNA polymerase it is possible to detect light emission as a result of the incorporation of dNTPs. Pyrosequencing is performed by providing a DNA dependent DNA polymerase

with a specific dNTP moiety and determining whether there is light emitted by luciferase. Light emission corresponds to the incorporation of the dNTP as complimentary to the next character of the DNA template. In the case of *de novo* pyrosequencing a cyclic presentation of dNTPs is employed. To date most pyrosequencing is limited to 30 nucleotides; new chemistries and algorithmic advances have increased the read length to 60 nucleotides [8]. Given the limitations on sequence read length in pyrosequencing it is not suited to *de novo* sequencing. Rather, pyrosequencing is best suited for re-sequencing protocols such as small-transcript profiling as well as for the metagenomic profiling of complex microbial communities [9]. Massively parallel signature sequencing (MPSS) has recently been developed as another high throughput method for re-sequencing on the basis of sequence signatures [10].

By exploiting the specificity of type II restriction endonucleases MPSS is capable of sequencing transcripts through successive rounds of adapter ligation and subsequent cleavage. Type II restriction endonucleases contain recognition and restriction sites separated by a known distance of nucleotides. In MPSS type II restriction endonucleases are used to attach a known adaptor and subsequently restrict a known distance into a transcript. Thus through successive rounds of adaptor ligation and subsequent restriction the type II restriction endonucleases dissect the 3' end of a transcript. Given that MPSS employs fluorescently labeled adapters the target transcripts can be sequenced by using a regime of known adapters and restriction endonucleases. While the data acquisition rate of MPSS is currently the fastest (80 million nucleotides per run) limitations on the length of sequencing reads are generally considered too short to facilitate *de novo* sequencing programs and hence MPSS is best suited for high throughput re-sequencing applications.

Since the inception of high throughput DNA sequencing methods, the rate of sequence acquisition has enabled the complete sequencing of numerous genomes. While the available genome sequences provide concrete insight into the genome composition of select model organisms, organisms which do not serve as a model for other species or are low social or economic impact are not likely to benefit from sufficient scientific or public pressure to justify the cost associated with deriving a complete genome sequence. Thus, there has been and will continue to be numerous non-model organisms with little or no prospect for determining a complete genome sequence. For organisms without the prospect of a sequenced genome, the expressed repertoire of genes is commonly investigated through the analysis of cDNA sequences as expressed sequence tags (ESTs).

ESTs are commonly derived through the sequencing of cDNAs prepared from a population of mRNAs extracted from specific tissues of an organism under a specific set of experimental conditions. The resulting mRNA population is enzymatically converted to cDNA *in vitro* through the used of an RNA dependent DNA polymerase, resulting in the generation of duplex DNA molecules which are complimentary to the captured RNA species. By exploiting the use of oligo-dT primers in the preparation of first-strand cDNA, the duplex DNA population can be enriched for molecules containing the 3' un-translated (UTR) component of the corresponding RNA species. Once cloned, individual cDNAs from the population can be readily sequenced using existing high throughput methods.

Protocols which enrich the mRNA populations on the basis of their 5' cap are commonly employed when it is desirable to derive a predominately full length cDNA library. The Cap-Trap technique was developed as a high throughput protocol for the

enrichment of predominately full length mRNAs through the use of RNaseI to degrade single stranded RNA molecules while leaving cDNA-RNA molecules intact [11,12]. In the CAP-Trap technique mRNA molecules are biotinylated at diol groups in their 5' cap and 3' UTR. Subsequently, first strand cDNA synthesis served to create cDNA-mRNA hybrid molecules. Treatment with RNaseI enriched samples for cDNA-mRNA hybrid molecules which were subsequently captured on streptavidin coated magnetic beads. RNaseH actively degraded the mRNA portion of the cDNA-mRNA duplex. Use of a terminal nucleotide transferase allowed the addition of an oligo(dG) at the 5' end of the single stranded cDNA molecule. A second strand cDNA synthesis using an oligo(dC) primer finally served to create predominately full length duplex cDNA molecules which were cloned and subsequently sequenced. With the development of EST sequence datasets a primary biological question of interest is to derive the non redundant set of transcripts actively transcribed in a tissue under the specific biological conditions. Appendix A: Wheat EST resources for functional genomics of abiotic stress and Appendix B: Large scale EST analysis; provide details of the analytical steps required to analyze a collection of ESTs.

For organisms with a sequenced genome, EST collections provide a useful basis for validating computationally derived gene models. Typically, as a genome is being sequenced a series of gene finding algorithms are used to predict the location and boundaries of genes within linkage groups. While computational methods for gene finding are under continual development, it is only with examination against biological evidence that the computational predictions/models can be validated. Through the alignment of EST sequences with genome sequence data it is possible to validate

predicted gene models with the observed mRNA species which is of critical import for predictions at the 5' and 3' most exons [13].

Taxonomic profiling is the process by which samples from disparate organisms are organized according to their evolutionary relationship to one another. By profiling the rate of sequence divergence of specific loci it is possible to correlate the rate of mutation with the evolutionary distance between organisms. Prior to the use of sequence based phylogenies, taxonomic assignments relied primarily on phenotypic measures such as morphological, physiological, and chemical properties of the organism. With advances in DNA sequencing protocols it has become possible to acquire large amounts of nucleic acid sequence data in reasonable time-frames and at reasonable costs; thus the use of DNA sequence based techniques is now warranted to taxonomically classify organisms [14-17]. By drawing upon large collections of sequence data for target loci, molecular phylogenies are built in a locus specific manner.

In particular, by measuring the sequence divergence between two organisms for orthologous genes it is possible to generate a relative measure of the evolutionary distance between the two organisms [18]. To date the most widely exploited locus for molecular phylogenies is the gene encoding the *16S* ribosomal RNA (*16S rRNA* [14-18]). The principle characteristics which affect the efficacy of a phylogenetic marker are universality and highly constrained function [18]. In order for a phylogenetic marker to be informative it must be present in all organisms under study (universal). For inference to be made using a sequence based marker there must also be some connection between sequence divergences and speciation (constrained). Therefore in practice phylogenetic markers typically are involved in pathways essential for survival. The gene encoding the

16s rRNA has been a major focus of interest for prokaryotic phylogenetics because of its high degree of conservation, which presumably is due to its role as the small ribosomal subunit of prokaryotes and its involvement in mRNA translation. Given the evolutionary pressures to constrain protein synthesis the pressure to reduce mutation at the *16S rRNA* locus is very high. Therefore the gene encoding the *16S rRNA* has been the standard target for molecular phylogenies due to its conservation throughout prokaryotes.

Chaperonins represent a specific class of molecular chaperones originally identified in plants and *Escherichia coli* [19]. As molecular chaperones, chaperonins are involved in protein folding, targeting and assembly of proteins [19]. Chaperonins fall into two groups: Type I chaperonins (*CPN60/HSP60/GroEL*) are molecular chaperones of approximately 60kDa which function in organelles (plastids, mitochondria and chloroplasts) while Type II chaperonins (*HSP70/CPN70/CCT*) function in the cytoplasm. Type I chaperonins are present in practically all organisms ranging from bacteria through archea to the organelles of eukaryotes (plastids and mitochondria). The only taxonomic class to be lacking *CPN60* are the microsporidia [20]. Recently *HSP70* sequences more similar to mitochondrial vs. eukaryotic have been identified in the nuclear genome of microsporidians [21].

Molecular phylogenies built on highly conserved domains of chaperonin sequences (e.g. *CPN60*) tend to have greater discerning power relative to *16S rRNA* phylogenies, especially when discriminating between closely related species [22]. Being involved in protein synthesis *16S rRNA* has been strongly conserved due to its central role in mRNA translation. Similarly, *CPN60* genes are thought to be conserved because of their chaperone roles in protein folding [23] and their involvement in cellular signalling

[24]. The likely basis for *CPN60* to be more discerning with respect to molecular phylogenies likely arises from the involvement of *CPN60* in protein folding as opposed to protein synthesis, together with proposals that evolutionary pressure on functional conservation is more acute with respect to protein folding as opposed to protein synthesis.

The availability of large numbers of *CPN60* gene sequences from a wide taxonomic sampling of organisms lead Goh and colleagues to develop a set of universal degenerate PCR primers [25]. Focusing on a large conserved portion of the *CPN60* sequence these workers identified a "universal target" which served to define the conserved domain of the *CPN60* locus amongst the study organisms. Subsequent applications of *CPN60* profiling focused on the distribution and composition of microbial communities based on the taxonomic classification of organisms via the sequence of their *CPN60* locus. Identifying and classifying microbial communities has implications to industry [26], livestock [22], as well as human health [27]. Through the taxonomic profiling of the microbial flora of a complex microbial community it is possible to correlate variation in microbial community composition with various responses of the host organism.

Microbial profiling may be a strong correlation indicator of patient outcomes in response to pathogenic infections. Based on samples of vaginal microflora collected from healthy human females it has been demonstrated that the application of *CPN60* profiling to human health can, in a culture-independent fashion, detect the presence of Chlamydia species. In particular, microbial profiling was capable of identifying *Chlamydia psittaci* on the basis of similarity to a reference database which other existing diagnostic tests were unable to detect ([28] and Appendix C: Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequenced-based methods). Research is

continuing to establish the complexity of the microbial flora present in the human reproductive tract and to determine the effect that the microbial community has on pre-mature births as well as susceptibility to other gynaecological diseases.

With the increase in the number of publicly available genome sequences, there has been a shift to examine the sequence conservation rates across multiple genes when determining taxonomic relationships. Fisher *et al.* demonstrated that by examining the sequence divergence of 25 orthologous genes it was possible to calculate the phylogenetic relationship among several hemiascomycetous yeast [29]. While single gene phylogenies were capable of resolving taxonomic relationships among related species, an analysis of multiple genes enabled the detection of fine relationships such as conserved gene order (*sensu stricto*) and higher levels of syntenic relation.

Molecular phylogenies to date have focused on the sequence divergence between orthologous genes particularly in terms of point mutations. By focusing on sequence substitutions and applying a metric to the rate at which substitutions happen, it is possible to quantify the evolutionary distance between two organisms on the basis of their sequence substitution rates. While sequence-based phylogenies provide directed insight into the evolutionary distance between closely related species, sequence based phylogenies breakdown when evaluating the evolutionary distance between distant organisms. This is particularly relevant when the genes being used for molecular phylogenies are organellar, as opposed to nuclear in origin. While all chromosomes as linkage groups are under selective pressure with respect to their gene compliment, the degree to which evolutionary pressures, such as compaction, impact their evolution will vary with the organellar location of the linkage group. In the case of organellar linkage

groups (mitochondrial and chloroplastic) the effect of sequence compaction will be different than that of linkage groups of nuclear origin. Therefore the rates of sequence divergence will vary between orthologous genes on the basis of their sub-cellular origin.

When using a single gene to determine the relatedness between multiple organisms, the possibility exists that the evolutionary pressures which caused the divergence between two species may have nothing to do with the specific gene used for phylogenetic analysis. For example, inversions, deletions, and translocations among and within chromosomes could occur such that a subset of genes was unaltered. If the orthologous gene(s) being use for phylogenetic analysis were within the unaffected set, it would be difficult to detect the genomic differences using models of sequence divergence. Conversely, if the measure of relationship between two organisms was based in part on a genome property which included chromosomal aberrations such as indels, duplications, rearrangements and translocations then such abnormalities would provide a potentially more effective approach through which to evaluate the evolutionary relationship amongst organisms.

Proposed here is a methodology to compare organisms on the basis of the distribution of their gene lengths. By comparing the distribution of gene length observed in a given organism it is hypothesized that it will be possible to taxonomically distinguish organisms based on the similarity of the distribution of their gene lengths. It is further hypothesized that if organisms can be distinguished on the basis of the distribution of their gene lengths such comparison would be able to detect gross chromosomal aberrations and thereby better classify large evolutionary distances between organisms when compared to molecular phylogenetics.

It has been demonstrated that the conservation of gene order can provide additional information when describing the evolutionary relationship between closely related species [29]. Existing evidence has also demonstrated that coding sequence composition is indicative of evolutionary relationships [22,26-33]. Thus if gene order and gene sequence are indicative of close order taxonomic relationships, then it would be reasonable to predict that organisms of similar evolutionary stage would exhibit corresponding similarity in the distribution of their gene lengths such that the similarity between distributions of gene length would correlate proportional with the relatedness of the species.

Materials and Methods

To date there have been a number of hemiascomycetous yeast genomes sequenced. The evolutionary distance spanned by the hemiascomycetous yeast is estimated to be equivalent to that of the phylum *Chordata* [29] and therefore as a data set offer a relatively complete and tractable set of model eukaryotic genome sequences for use in comparative genomic studies.

The genome sequences used in this study were drawn from the complete eukaryotic genome data sets publicly available via GenBank. The Hemiascomycetous yeast data set used in this study included: *Aspergillus fumigatus*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica*. Vertebrate genomes used were *Homo sapiens, Mus musculus, Rattus norvegicus, Pan troglodytes, Canis familiaris,* and *Danio rerio.* Finally, invertebrates examined included *Drosophila melanogaster* and *Apis*

mellifera together with the lone plant genome of *Arabidopsis thaliana.* While other genome sequences were available at the time of this study many were and still are in varying stages of completion [34]. Annotation consistency was also a factor in the selection of the test data sets, for example, the genome sequence of *Oryza sativa* was not considered in this study because its annotation was not held in GenBank and as such may have contained features which were inconsistently defined vis-à-vis with the other organisms annotated and available from GenBank.

For each genome selected, the distribution of gene lengths was calculated by identifying all coding sequence (CDS) features as published by GenBank (September 2005). CDS sequence information is limited to the nucleotide sequence which corresponds to the translated protein sequence [35]. Thus, analyses of CDS features were based on the protein coding complement of genes for an organism and did not include RNA-based genes. CDS features were identified using BioPerl scripts and a distribution was generated by grouping the CDS features into 100 base pair bins. In order to exploit the functionality of Genespring GX™ the distribution of gene lengths was used to create text files analogous to a series of microarray experiments such that there was an experiment data file for each 100bp bin, within which each organism was represented as a probe set. A $\log_{10}$ transform was performed on the frequency counts prior to analysis in Genespring GX™.

Analysis of the CDS length distributions was done by loading the distributions into Genespring GX™ and performing 2 types of clustering. In one approach, K-means clustering was used with varying numbers of clusters to classify the distributions into related clusters. Cluster sizes up to 5 were preformed based on a number of clusters equal

to 1 more than the predicted number of taxonomic divisions within the data (4 = basidiomycetes, archiascomycetes, ascomycetes and microsporidia). In a second approach, hierarchical clustering (genetrees) was used to identify the interrelatedness between organisms on the basis of their CDS feature distributions. For all hierarchical clustering the similarity measure was based on the Pearson correlation.

Results

A general assessment of the distribution of gene length was performed by plotting the number of occurrences of CDS that occurred for a given length class (100bps). As can be seen in Figure 3, Figure 4, Figure 5, and Figure 6, while most genomes appear to abide a log-normal distribution in their CDS feature distributions, there was a distinct bias seen in the frequency of CDS features in mammalian genomes. Shown in Figure 6 is a distinct bias in the mammalian genomes for CDS features with a length of 1kbps. For subsequent analyses involving mammalian genomes the distributions were normalized to the median of the CDS distribution to account for the bias seen in the CDS distribution of the mammalian genomes.

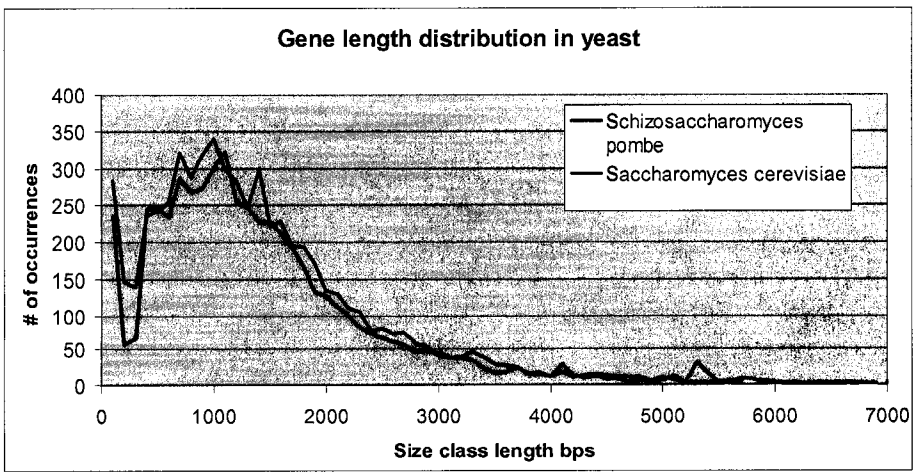

Figure 3 CDS length distribution for *S. pombe* and *S. cerevisiae*. The frequency of CDS feature length is plotted in 100bp size classes.

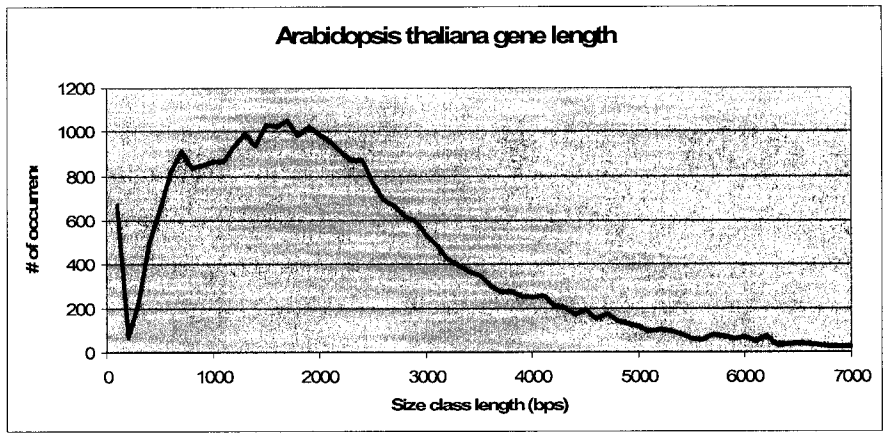

Figure 4 CDS length distribution of *A. thaliana*. The frequency of CDS feature length is plotted in 100bp size classes.

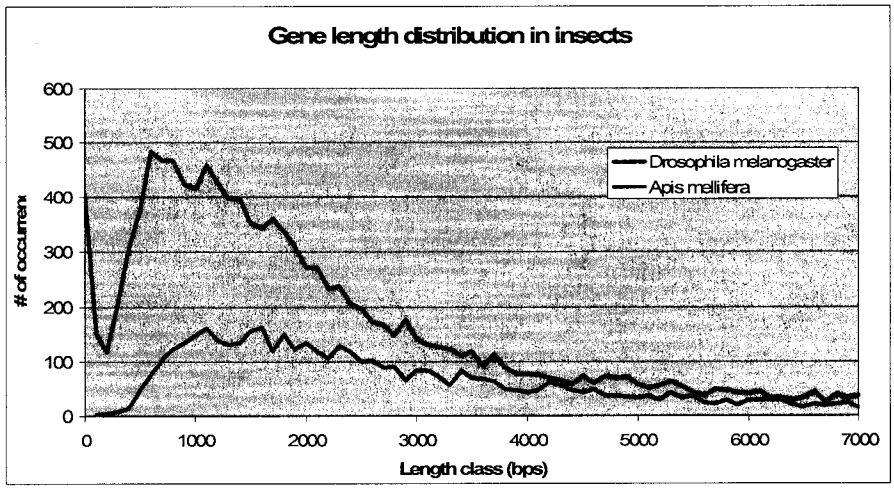

Figure 5 CDS length distribution in insects. The frequency of CDS feature length is plotted in 100bp size classes.

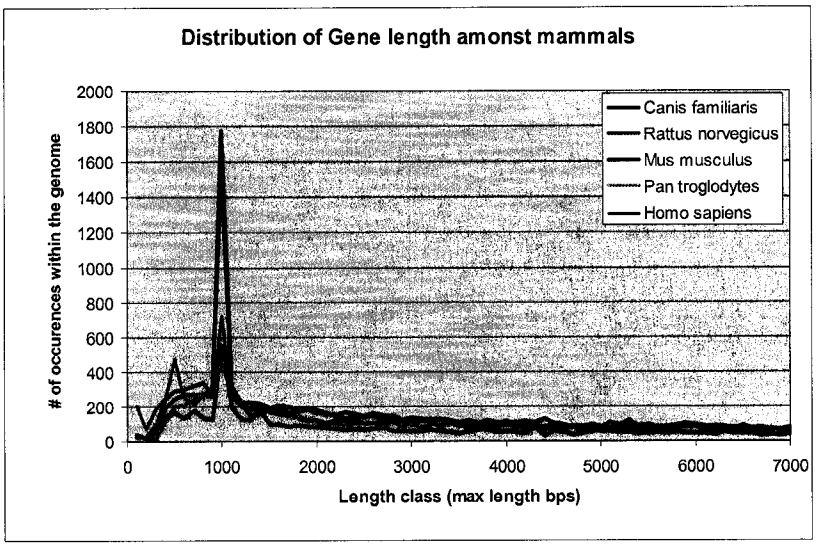

Figure 6 CDS length distribution in mammals. The frequency of CDS feature length is plotted in 100bp size classes.

Through the use of K-means clustering it was possible to identify gross taxonomic relationships between organisms. There were two taxonomic groups which were relatively over-represented in the dataset: mammals and the hemiascomycetous yeast. Therefore it was proposed that there would be three distinct clusters seen in the CDS distributions present in this study: mammals, yeast and others. As shown in Figure 7, the mammalian genomes were found to cluster closely. Similarly, the hemiascomycetous yeasts were clustered together with the exception of *C. neoformans* and the inclusion of *D. melanogaster*.

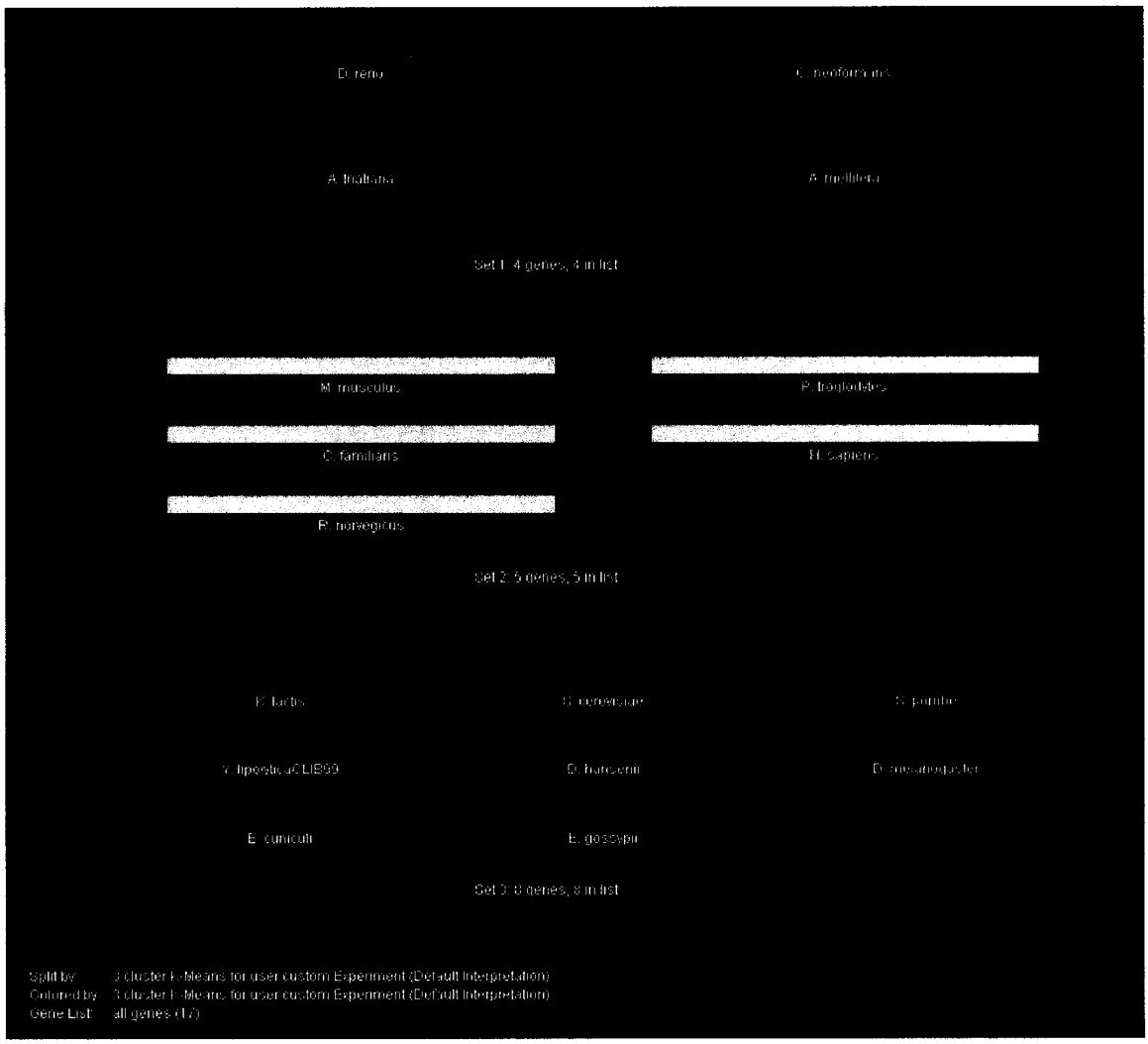

Figure 7 Kmeans (3) clustering of the CDS feature distributions of all genomes. Genomes were analyzed according the distribution of their CDS lengths. Clusters of genomes with similar distributions are shown as similar colors. In the figure the use of the term "gene" is a reference by Genespring-GX™ which in this analysis is equivalent to "genome"

In order to further examine the ability for CDS distributions to provide a mechanism for the phylogenetic analysis of organisms, hierarchical clustering was performed on the hemiascomycetous yeast genomes as shown in Figure 8.

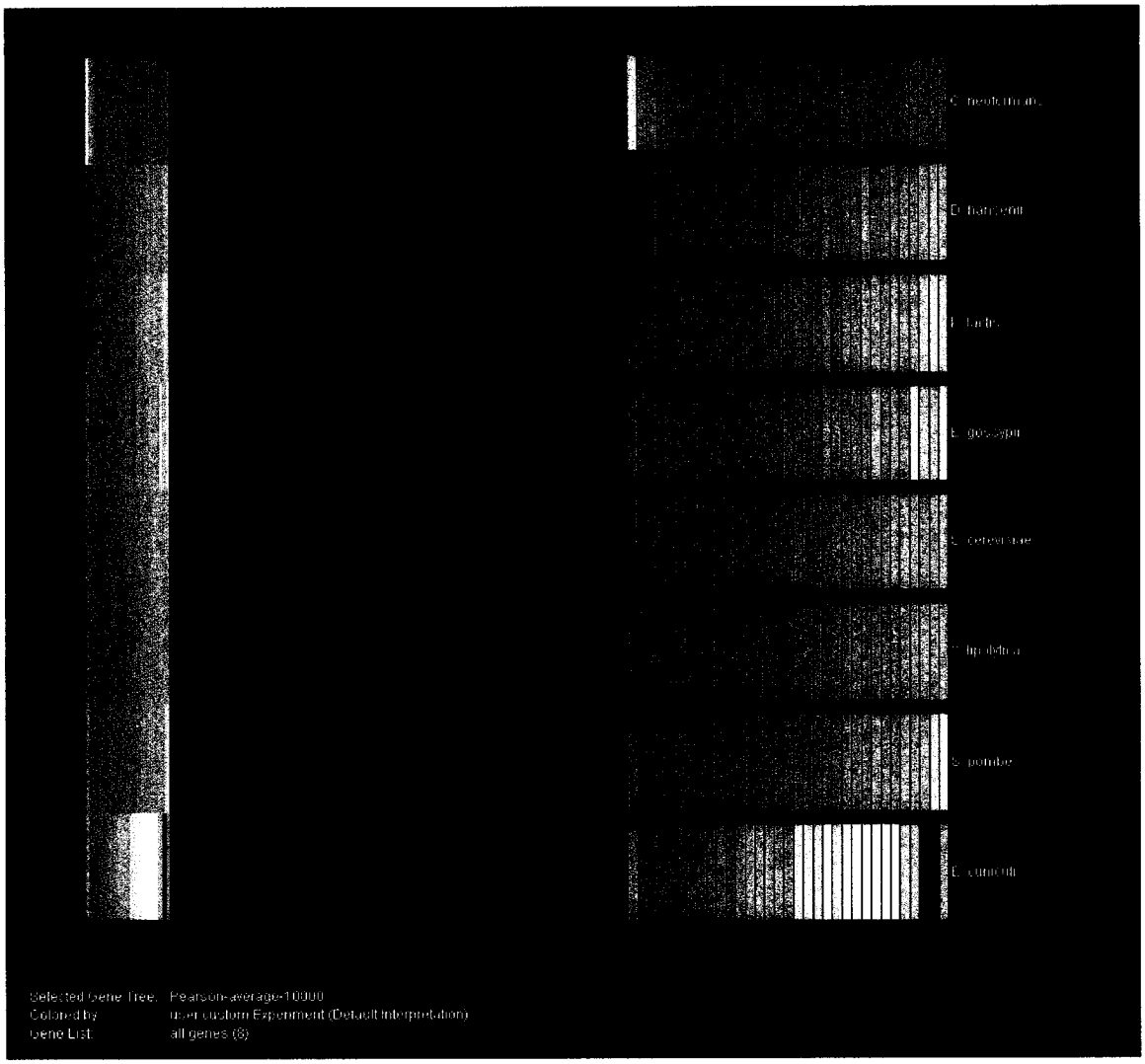

Figure 8 Hierarchical clustering of hemiascomycetous yeast CDS distributions using the average Pearson correlation. CDS length distributions are plotted as bars where each bin size increases by 1Kbp increments from left to right. Coloration of CDS distributions is done using a heat map.

Given that the overall relationship between CDS distribution paralleled closely the distribution observed using molecular phylogenies, K-means clustering was used to color the genetree. Since the genetree for the hemiascomycetous yeast, effectively classified the organisms into four clades (basidiomycetes, microsporidian and 2 ascomycete clusters), K-means clustering was applied to find the four most similar clusters as summarized in Figure 9.

Figure 9 Hierarchical clustering of hemiascomycetous yeast CDS distributions using the average Pearson correlation. CDS length distributions coloured according to a K-means clustering of 4 clusters.

An important observation from this analysis was that the correlation between the four clades identified in the Pearson based hierarchical clustering were confirmed using a K-means clustering of 4 clusters. Indeed, the resulting clusters were found to delineate the basidiomycetes from the microsporidian, in turn from the ascomycetes. In order to further explore the relationship amongst the ascomycetes, the K-means clustering was taken to five clusters and the resulting segregation was evaluated.

Figure 10 Hierarchical clustering of hemiascomycetous yeast CDS distributions using the average Pearson correlation. CDS length distributions coloured according to a K-means clustering of 5 clusters.

Increasing the number of clusters in the K-means analysis to 5 revealed that *Y. lipolytica* was less similar to *E. gossypii*, *S. cerevisiae*, and *S. pombe* as summarized in Figure 10.

Discussion

Using K-means clustering it was possible to classify the study organisms into 3 clusters based on the distribution of their CDS features: mammals, yeast and others

(Figure 7). The segregation of the mammalian organisms was discrete with no additional organisms clustering with them, and all of the mammals were grouped into a single cluster. Similarly the CDS length distributions derived from the model fungi used in this study were well separated into a distinct cluster. Only the basidiomycete *C. neoformans* clustered apart from the fungi, clustering instead with *A. thaliana, D. rerio, and A. mellifera*. The only other organism with a spurious result was *D. melanogaster* which clustered with the ascomycetes and the microsporidian *E. cuniculi*. Analysis of CDS feature distributions faithfully classified organisms into taxonomically informative groups. Further, the taxonomic distinctions of disparate organisms were detectable within the distribution of gene lengths.

Applying hierarchical clustering to the hemiascomycetous yeast genomes revealed evolutionary relationships that were highly similar to those derived using molecular phylogenetics [29]. Among the fungal genomes used here, three distinct clusters were formed representing the basidiomycete, the microsporidian and the ascomycete genomes (Figure 9). As could be predicted, the two most distant genomes were defined by non-ascomycete yeasts; the basidiomycete *C. neoformans* and the microsporidian *E. cuniculi*.

Analyzing the relationships amongst the ascomycetes with K-means and hierarchical clustering revealed that there are three distinct sub-clades arising within the ascomycetes (Figure 10). The ascomycetes cluster containing *D. hansenii* and *K. lactis* were lodged in a clade distinct from *S. pombe*, *S. cerevisiae* and *E. gossypii* while *Y. lipolytica* fell into its own clade. Comparing these findings to those generated from a phylogenetic analysis involving 25 distinct loci, there was a high degree of correlation [29].

The relationships between *S. cerevisiae*, *A. gossypii* (synonymous with *E. gossypii*), *K. lactis* and *D. hansenii* were all supported by both existing molecular phylogenies as well as the clustering of their CDS distributions. The one major difference between this study and the molecular phylogeny of Fisher's work was the proximity of *Y. lipolytica* with *S. pombe* and *S. cerevisiae*. This was likely due to the inclusion of the archiascomycete *S. pombe*, which Fisher did not examine.

Investigation of the taxonomic relationships amongst fungi to date has relied on varied techniques. Distinct characteristics between fungi based on their sexual compatibility, morphology, and physiological specialization have been largely surpassed through the use of molecular characteristics [36]. In order to assess speciation amongst fungi it is desirable to determine whether there is interbreeding between populations and thereby determine whether two populations represent distinct species. Unfortunately, as a broad approach breeding studies are complicated amongst diverse fungi. Culture conditions may bias the mating efficacy and cause difficulties in assessing the sterility of offspring. Further there are some fungi which are strictly asexual (some dermatophytes). Therefore for determining the taxonomic relationship amongst fungi, sexual mating studies are not as effective (broadly speaking) as other techniques. Microscopy has been extensively used to study the morphological characteristics of various fungal species. While morphological characteristics are consistent with taxonomic classification morphological techniques are complicated in their efficacy. In cases of pathogenic fungi it is common to only examine a distinct phase of the fungal life cycle (e.g. vegetative). Further, the methods employed in culturing the organism may affect the visibility of the morphological characteristics necessary for taxonomic assessment.

As demonstrated here, a comparison of the distribution of gene length amongst organisms allows for the classification into taxonomically meaningful groups on the basis of a genome property. The information represented in the distribution of an organism's CDS lengths is also shown to be constructive when determining the relationship between highly similar organisms and also when comparing large taxonomic divisions. In future the possibility exists to analyze organisms which are without a complete genome sequence through the examination of EST libraries. With the use of the CAP-Trap protocol libraries can be substantially enriched in their representation of full length mRNA sequences [11,12]. With a CAP-Trap based cDNA library which was representative of the gene composition of an organism, future analysis could use the distribution of CAP-Trap cDNAs as a reasonable surrogate for the CDS features used in this study.

The advent of new and high throughput sequencing techniques such as pyrosequencing and MPSS provide for DNA acquisition at rates orders of magnitude faster than those employed by the human genome project. By dramatically increasing the acquisition rate of sequence data, these new technologies will enable deeper biological inquiry into expressed repertoires of genes chiefly by reducing the costs and labour associated with DNA sequencing. Application of MPSS and pyrosequencing to the sequencing of EST libraries stands to dramatically increase the depth of expressed gene information available for many organisms for which there is little or no genome sequence data available. While MPSS and pyrosequencing are not suited to *de novo* genome sequencing both are well suited to high throughput screening of complex communities (metagenomics). Existing techniques for microbial profiling are limited, largely by labour and cost associated in the cloning and subsequent sequencing of distinct samples.

Presently, Sanger sequencing is 10-30 times the cost of pyrosequencing and MPSS [37]. In order to survey communities at a depth which allows for statistically significant conclusions to be drawn orders of magnitude improvement are needed in the rate of data acquisition and cost relative to that of Sanger based sequencing [27].

The identification of *C. psittaci* in human vaginal flora was first identified in previous work ([27] and Appendix C: Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequenced-based methods). While other Chlamydia species have been detected in human vaginal flora (e.g. *C. trachomatis and C. abortus*) the extent to which individual Chlamydia species are present in healthy and non-healthy women is not know. Given the correlation between *C. abortus* and premature births, there exists an important justification to better understand the role of Chlamydia species in human reproductive health [38]. Further, culture-independent techniques provide an unbiased approach to interrogating the diversity and composition of microbial communities. Thus the automated system presented in Appendix C: Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequenced-based methods, tied with new high throughput sequencing of large libraries represents a focal point for future investigation into how microbial communities may correlate with and affect human reproductive health.

As shown here, there exist significant differences in the distribution of coding gene length among taxonomically distant eukaryotes. Specifically, there are detectable differences between mammalian and fungal gene length distributions such that exploitation of these differences may offer a novel approach to DNA-based molecular diagnostics of complex samples. In the case of fungal infection in mammals, it may be

possible to analyze complex EST libraries in the absence of a complete genome sequence. By exploiting such high-throughput techniques as Pyrosequencing, it may be possible to dramatically increase the number of EST sequences acquired in a reasonable time. If the EST resources were generated such that they represent predominately full length transcripts then the library should be a reasonable alternative to the CDS distributions used in this study. With a dataset representative of CDS length, it may be possible to detect multiple statistically significant distributions in the length of mRNAs sampled. Thus it could be possible to provide an initial assessment of the taxonomic classification of the pathogen in the absence of genome sequence data. Similar work is already being done for investigating pathogenic infections such as the oomycete *Albugo candida* on *Brassica juncea*. By isolating complex mRNA samples from *B. juncea* challenged with *A. candida* it is possible to identify ESTs derived from either organism (host and pathogen together) based on sequence similarity to reference databases. New techniques based on the distribution of gene length would enable assessment of pathogenic infections *a priori*. By identifying pathogens *a priori,* such techniques would provide methodology to interdict new and complex diseases in organisms without complete genome sequences.

The analysis of CDS length as described in this study provides a mechanism for investigating the taxonomic relationship amongst eukaryotes. As shown in this work CDS length distributions are informative with respect to taxonomic relationships at both large (mammal versus fungi) and small (amongst fungi) scales. With advances in the rate of acquisition of DNA sequence data, the ability now exists to perform large scale EST based surveys of many organisms. By combining high throughput sequencing techniques

with protocols which enrich for full length transcripts it is possible to rapidly derive a representative set of predominately full length cDNA sequences for an arbitrary set of study organisms. While the work in this chapter focuses on taxonomic identity in the absence of genome sequence data there likely exists a relationship between taxonomy and the organization of genes within chromosomes. In Chapter II (**Recasting the C-value paradox in terms of chromosomal coding capacity**) I examine the gene encoding strategies of a set of model organisms in order to address whether there is a correlation between the chromosomal organization of genic information and taxonomic identity.

CHAPTER II

RECASTING THE C-VALUE PARADOX IN TERMS OF CHROMOSOMAL CODING CAPACITY

Introduction

For a given genome the C-value refers to the number of nucleotides required to encode the set of structural and coding features heritably transmitted by it. In the 1960s there was increasing evidence that the DNA content of organisms with similar morphological complexity varied by orders of magnitude [39]. The C-value paradox arose from the apparent contradictory finding that multiple organisms of similar complexity, in terms of protein coding genes contained within their genomes, exhibited wide ranges of genomic DNA content [39]. Of specific focus to the paradox is the observation that there is no direct correlation between the number of genes and the amount of DNA comprising a given eukaryotic genome [40]. The resolution of the paradox came through the realization that DNA exhibits many more possible functions than just to encode for RNA transcripts and thereby proteins [41]. Whether genomic DNA encodes protein, acts in a physical sense as a spacer element (introns or intergenic) or functional RNA species (tRNA, rRNA or microRNAs), the DNA still represents functional information organized into one or more distinct linkage groups. While this initially seems to represent a contradiction to the paradox, it is in fact a realization that a gene is a *functional unit of DNA* regardless of its mode of action. The specific function of the DNA is not critical when one examines the relationship between DNA content and the number of genes in the context of a specific genome. While the C-value paradox has focused on the seeming discrepancy between gene number and genome size, I suggest here that the paradox needs to be re-examined in

a manner which accounts for the variation of gene length with respect to the length of the chromosome into which the genes are organized.

Claude Shannon's concept of information theory provides a methodology to examine how efficiently a medium can transmit a message [42]. Applications of information theory to computational biology typically consider an organism as recording messages (genes) about its environment into its genome (medium) and then transmitting the information on to subsequent progeny [43]. At the root of the C-value paradox is the notion that there is an average message length (gene length), which when amortized against the length of the medium (genome size), should result in a direct correlation between the total gene number and the genome size. By assuming that the entire genome is capable of storing information the paradox assumes that there is no evolutionary pressure on the length of a gene.

Examining the gene content (information capacity) in terms of a genome is not strictly speaking a useful approach because a genome is a virtual collection of, generally, multiple independent linkage groups [40]. Therefore it is not the genome *per se* but rather its constituent independent linkage groups that present certain characteristics. While the C-value paradox relies on the null hypothesis that genes have a consistent length, I propose here that the null hypothesis needs to be recast in the context of how information is organized into linkage groups.

Assuming that a linkage group can organize information throughout its length (which for the organisms under study is presently accepted), it would be reasonable to calculate an expected mean gene length in terms of the number of genes spaced equally throughout a linkage group. Inherent to this assumption is that all genes would require the

same amount of genomic DNA to represent them. Further, if a linkage group contains a known number of genes then the mean expected gene length is simply expressed as the length of the chromosome divided by the number of genes it contains. Therefore the expected mean gene length represents the average amount of a linkage group required to faithfully represent an *average* gene.

In reality the observed mean gene length will likely differ from the expected due to various constraints on how genes are organized into linkage groups. I propose here that the amount of information that a linkage group of a given size can store is the critical factor, and not simply genome size. By quantifying the difference between the observed and expected mean gene lengths for a linkage group it is possible to define the variance from the C-value paradox for a given chromosome. In an idealized situation a chromosome which was optimally dense would be one in which the number of nucleotides required to encode each gene was the same. Hence, if a linkage group were to wholly abide the C-value paradox then the observed and expected mean gene lengths would be the same and their difference would be therefore zero. Conversely as a linkage group deviates from the C-value paradox the difference between observed and expected mean gene lengths will increase. While there will be a variation between the observed versus expected mean gene lengths the variance would be predicted to approach zero as a chromosome more effectively encodes information. Thus when the optimal spacing of genes is achieved in a given linkage group, that linkage group is predicted to more faithfully abide the C-value paradox.

When a single gene is stored in genomic DNA it will be present on one of the two anti-parallel strands. If the other strand contains no other genic information the nucleotide pairs underlying the gene are discreetly storing information (Figure 11).

```
Sense      5'----        ---ATGTCTGCGAAGAGATTGTGTTGAAGAGTTCCGATTGTGAATCTTTAA----3'
Antisense 3'----ATATTTT---TACAGACGCTTCTCTAACACAACTTCTCAAGGGTAACACTTAGAAATT----5'
Score        00000000000000011111111111111111111111111111111111111111111110000000
```

Figure 11 Schematic of a gene which is discreetly encoded. Shown is a single intronless gene encoded in the sense strand. The scoring array is 1 for each nucleotide pair within the coding region.

Conversely if both anti-parallel strands store information then the nucleotide pairs are storing information in a retrograde fashion (Figure 12).

```
Sense      5'----        ---ATGTCTGCTTAGAAGATTGTGTTTGTGAATCTTTAACAT---TTTTATA---3'
Antisense 3'----ATATTTT---TACAGACGAATCTTCTAACACAAACACTTAGAAATTGTA---       ---5'
Score        00000000000000111111111112222222222222222222222211111100000000000000
```

Figure 12 Schematic of two genes in retrograde orientation. The scoring array is 1 where each nucleotide pair is required for a discreet portion of a gene and 2 where the nucleotide is required for a gene on either strand.

In cases where nucleotide pairs represent a single gene the nucleotide pair is under evolutionary pressure based on the conservation of one nucleotide in the pair. By contrast, nucleotide pairs belonging to genes encoded in retrograde fashion are subjected to twice the evolutionary constraint of discreet pairs since retrograde pairs are bound by pressures from both DNA strands. Therefore in order to evaluate the linkage group with respect to its evolutionary pressure, it is necessary to examine each nucleotide pair for discreet versus retrograde genic information.

By examining the variance between observed mean gene length and the portion of a linkage group under retrograde encoding it should be possible to taxonomically classify linkage groups if the way in which information is encoded into a chromosome is

indicative of adaptation. In order to assess the density of information stored in a linkage group, a system to quantify the capacity of a chromosome in terms of its coverage with coding features is derived and presented here. The coding capacity of a chromosome is calculated proportional to its length (Equation 2 and Equation 3); specifically, the capacity is the number of nucleotide pairs dedicated to represent information as a percentage of the linkage group's length. The maximum capacity for a linkage group to store information (100%) would arise when a linkage group only stores information on a single strand (discreet) for each nucleotide pair and is subjected to positive selective pressure throughout its length.

The capacity of a linkage group, which stores information partially in a retrograde fashion, can be calculated relative to a chromosome of same length which organizes information solely in a discreet manner. For chromosomal regions where information is stored in a retrograde fashion, the additive effect on capacity would be twice that of a discreet region (Figure 11 vs. Figure 12). By weighting the impact of retrograde regions more heavily the calculation accounts for information within both anti-parallel DNA strands and, more specifically, that these regions would be subjected to greater evolutionary pressure. In this way it is possible to quantify the effect of information storage schemes for a given linkage group in terms of the chromosome's maximum capacity. By examining the correlation between the encoding strategy of chromosomes and the optimization of mean gene length I demonstrate here that it is the additive effect of discreet and retrograde storage schemes which best correlate ($r > 0.97$) with optimally dense linkage groups.

The publicly available set of fungal genome sequences deposited in GenBank represent one set of Eukaryotic genomes through which one can demonstrate significant diversity in evolutionary terms, while being tractable for comparative genomics. The value of fungal genomes as a model for other eukaryotes and for comparative genomic studies derives from the fact that multiple fungal genomes have been completely sequenced to a high quality. Thus the resulting data sets are complete, reasonably small and publicly available. Proposed here is a methodology to examine the efficacy of linkage groups to encode protein coding genes (CDS features). Demonstrated is that by storing information in retrograde as well as discreet fashion, linkage groups are capable of approaching 100% capacity. Further I show that the only genome among the currently available fungi which exceeds 100% capacity in its linkage groups is that of the microsporidian, *Encephalitozoon cuniculi.*

E. cuniculi is an obligate intracellular organism which is unusual amongst eukaryotes in that it lacks mitochondria and at the same time presents the smallest known eukaryotic genome at 2.5 megabases [44]. There has been great controversy over the taxonomic classification of *E. cuniculi* [45-48]. Due to its lack of typical eukaryotic organelles and general prokaryotic characteristics *E. cuniculi* was originally classified as a member of the Archezoa [49]. Thus *E. cuniculi* was proposed as an ancient eukaryote dating back before the gain of the endosymbiont which is postulated to have given rise to the mitochondrion. Subsequent use of molecular phylogenetics challenged the original Archezoa classification of *E. cuniculi* and has suggested that the microsporidia are more closely related to yeast [46].

Recent work has discovered that a number of microsporidians including *E. cuniculi* actually encode within their nuclear genomes chaperonin genes which appear prokaryotic-like based on sequence similarity [21,50-52]. Thus the current literature points to an evolutionary chain of events where the microsporidians likely participated in horizontal gene transfer event(s) prior to shedding the mitochondrial-endosymbiont during their evolution. In divesting itself of mitochondria during evolution, the presumed effect on the genome of *E. cuniculi* would be to require the encoding of more information than that of a typical eukaryote (> 100% capacity). As is shown here, the finding that the *E. cuniculi* chromosomes display ~105% coding capacity is consistent with a scenario of the microsporidian nuclear genome acquiring additional gene functions from the mitochondrial endosymbiont prior to its loss during evolution.

Materials and Methods

DNA sequence data sets and corresponding annotation files were obtained from Genbank (September 2006) for publicly available annotated fungal genomes, including *Aspergillus fumigatus*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica*. Two chromosomes, one each, from *S. cerevisiae* (NC_001224) and *S. pombe* (NC_001326) were excluded from the study due to a lack of any retrograde CDS features. All of the linkage groups analyzed were nuclear in origin and therefore the study did not include the respective mitochondrial genome sequences.

Table 1 Summary of Eukaryotic linkage groups analyzed.

Organism	Accession	Length (bps)	Number of CDS	Observed Mean CDS Length (bps)	Expected Mean CDS length (bps)
Aspergillus fumigatus	NC_007194	4918979	1660	1662	2963
	NC_007195	4844472	1696	1688	2856
	NC_007196	4079167	1402	1602	2910
	NC_007197	3923705	1270	1606	3090
	NC_007198	3948441	1375	1643	2872
	NC_007199	3778736	1251	1674	3021
	NC_007200	2058334	647	1600	3181
	NC_007201	1833124	622	1545	2947
Candida glabrata	NC_005967	485192	202	1465	2402
	NC_005968	502101	209	1526	2402
	NC_006026	558804	231	1510	2419
	NC_006027	651701	278	1510	2344
	NC_006028	687501	284	1546	2421
	NC_006029	927101	389	1496	2383
	NC_006030	992211	435	1503	2281
	NC_006031	1050361	455	1517	2308
	NC_006032	1089401	458	1489	2379
	NC_006033	1192501	514	1543	2320
	NC_006034	1302002	551	1496	2363
	NC_006035	1440588	566	1584	2545
	NC_006036	1400893	609	1543	2300
Debaryomyces hansenii	NC_006043	1249565	678	1376	1843
	NC_006044	1349926	750	1344	1800
	NC_006045	1592360	853	1389	1867
	NC_006046	1602771	952	1276	1684
	NC_006047	2037969	1150	1355	1772
	NC_006048	2336804	1309	1368	1785
	NC_006049	2051428	1201	1337	1708
Eremothecium gossypii	NC_005782	691920	381	1434	1816
	NC_005783	867699	462	1506	1878
	NC_005784	907057	497	1469	1825
	NC_005785	1466886	819	1410	1791
	NC_005786	1519138	800	1509	1899
	NC_005787	1813164	982	1483	1846
	NC_005788	1476513	777	1530	1900
Kluyveromyces lactis	NC_006037	1062590	530	1409	2005
	NC_006038	1320834	666	1407	1983
	NC_006039	1753957	877	1437	2000
	NC_006040	1715506	878	1364	1954
	NC_006041	2234072	1131	1408	1975
	NC_006042	2602197	1249	1467	2083
Saccharomyces cerevisiae	NC_001133	230208	94	1509	2449
	NC_001134	813178	406	1504	2003

	NC_001135	316617	159	1374	1991
	NC_001136	1531918	753	1512	2034
	NC_001137	576869	273	1444	2113
	NC_001138	270148	126	1468	2144
	NC_001139	1090946	525	1503	2078
	NC_001140	562643	281	1449	2002
	NC_001141	439885	207	1512	2125
	NC_001142	745667	356	1577	2095
	NC_001143	666454	312	1545	2136
	NC_001144	1078175	508	1555	2122
	NC_001145	924429	460	1521	2010
	NC_001146	784333	393	1480	1996
	NC_001147	1091289	536	1472	2036
	NC_001148	948062	461	1505	2057
Schizosaccharomyces pombe	NC_003421	2452883	908	1471	2701
	NC_003423	4509021	1888	1437	2388
	NC_003424	5572983	2287	1513	2437
Yarrowia lipolytica	NC_006067	2303261	731	1377	3151
	NC_006068	3066374	936	1560	3276
	NC_006069	3272609	956	1438	3423
	NC_006070	3633272	1133	1498	3207
	NC_006071	4224103	1445	1480	2923
	NC_006072	4003362	1319	1475	3035
Cryptococcus neoformans	NC_006670	2300533	811	1942	2837
	NC_006679	1085720	325	1992	3341
	NC_006680	1019846	337	2037	3026
	NC_006681	906719	314	1906	2888
	NC_006682	787999	254	1940	3102
	NC_006683	762694	231	1989	3302
	NC_006684	1632307	567	1963	2879
	NC_006685	2105742	715	1900	2945
	NC_006686	1783081	629	1995	2835
	NC_006687	1507550	530	1992	2844
	NC_006691	1438950	474	2037	3036
	NC_006692	1347793	467	1960	2886
	NC_006693	1194300	385	2011	3102
	NC_006694	1178688	436	1912	2703
Encephalitozoon cuniculi	NC_003229	197426	157	1103	1257
	NC_003230	194439	158	1062	1231
	NC_003231	218329	172	1068	1269
	NC_003232	211018	172	1068	1227
	NC_003233	220294	172	1113	1281
	NC_003234	226576	188	1078	1205
	NC_003235	238147	211	975	1129
	NC_003236	262797	190	1245	1383
	NC_003237	267509	211	1095	1268
	NC_003238	251002	206	1075	1218
	NC_003242	209982	159	1004	1321

In order to assess the capacity of each sample genome to store CDS information the Genbank annotation files for each linkage group were examined using scripts developed in Perl exploiting the use of BioPerl modules [53]. By recording the position and overlap between CDS features [35] it was possible to calculate the percentage of each linkage group which exhibited a given organizational strategy (discreet vs. retrograde). By comparing the observed coding coverage to the length of the linkage group, it was further possible to assess the effect of coding strategies on the overall capacity of each linkage group.

Both discreet and retrograde strategies were examined as part of this study. Discreet regions of a linkage group were defined by those regions where CDS features were present within one strand of the linkage group, whereas, CDS features encoded on both of the anti-parallel strands the nucleotide pairs were recorded as retrograde. Assessment of the coding capacity of each linkage group was carried out by creating an array for both the sense and antisense strand of the linkage group. Traversing the Genbank annotation, each element of the array corresponding to a specific strand was incremented according to the occurrence of a CDS feature. Thus nucleotide positions corresponding to discreet regions of the linkage group would score 1 (sense[i] = 1 Xor antisense[i]). In cases where retrograde CDS features were observed both arrays were incremented (sense[i] > 0 and antisense[i] >0).

The percentage capacity of a linkage group covered by discreet CDS features was calculated by the summation of all nucleotide positions which were subsumed by a lone CDS feature on either strand as a percentage of the linkage group's length.

$$\left(\frac{\sum_{1}^{length(linkagegroup)} Xor(sense[i], antisense[i])}{length(linkagegroup)}\right)*100$$

Equation 2 Capacity of a chromosome based on discreet CDS features.

The percentage of a linkage group which displayed retrograde CDS features was calculated as the summation of nucleotide pairs where there was a CDS feature observed on both antiparallel strands divided by the length of the linkage group.

$$\left(\frac{\sum_{1}^{length(linkagegroup)} And(sense[i], antisense[i])}{length(linkagegroup)}\right)*100$$

Equation 3 Capacity of a chromosome based on retrograde encoding.

In order to assess the effect of retrograde organization of CDS features on the capacity of a given linkage group a relative measure was made by calculating the equivalent percentage of a linkage group of same length but displaying no retrograde encoding. The effective capacity for a linkage group was further calculated as the sum of discreet nucleotide pairs plus twice the sum of the retrograde nucleotide pairs.

$$\left(\frac{\left(\sum_{1}^{length(linkagegroup)} Xor(sense[i], antisense[i])\right)+\left(2*\sum_{1}^{length(linkagegroup)} And(sense[i], antisense[i])\right)}{length(linkagegroup)}\right)*100$$

Equation 4 Calculation of the effective capacity of a linkage group.

As an indicator of the capacity of a linkage group, two measures of mean CDS length were calculated. The Observed mean CDS length for a given linkage group was calculated as the average length of a CDS for a given linkage group.

$$ObservedMeanCDS = \frac{\sum_{1}^{n} length(CDS)}{n}$$

Equation 5 Calculation of Observed Mean CDS length for a linkage group. Where *n* is the number of genes encoded within a chromosome.

In addition, the expected mean CDS length was calculated as the length of the linkage group divided by the number of genes (CDSs) encoded by the linkage group.

$$ExpectedMeanCDS = \frac{length(linkagegroup)}{n}$$

Equation 6 Calculation of Expected mean CDS length for a linkage group. Where *n* is the number of genes encoded within a chromosome.

To assess the efficacy of CDS organization for a given linkage group the difference between the Observed mean CDS length and the Expected mean CDS length was calculated as follows.

$$VarianceCDS = \left(\frac{\sum_{1}^{n} length(CDS)}{n} \right) - \left(\frac{length(linkagegroup)}{n} \right)$$

Equation 7 Calculation of the variance between observed and expected mean CDS length. Where *n* is the number of genes encoded within a chromosome.

Assessment of the effect of CDS organization strategies on the variance in mean gene length where made by performing quintic regressions between the variance of mean gene lengths and the two organization schemes under analysis (discreet and retrograde).

Results

By plotting the Observed mean CDS length of a linkage group versus the percentage of a linkage group, which contains retrograde CDS features, 3 distinct classes of linkage groups were identified as shown in Figure 13. Further the fungal linkage

groups clustered into 3 classes which correspond with the phylogenetic relationship between the organisms based on molecular phylogenetics [46]. All linkage groups of the Ascomycetes fungi fell into Cluster II which was distinct from the linkage groups of the microsporidian *E. cuniculi* and the basidiomycete *C. neoformans* (Figure 13).

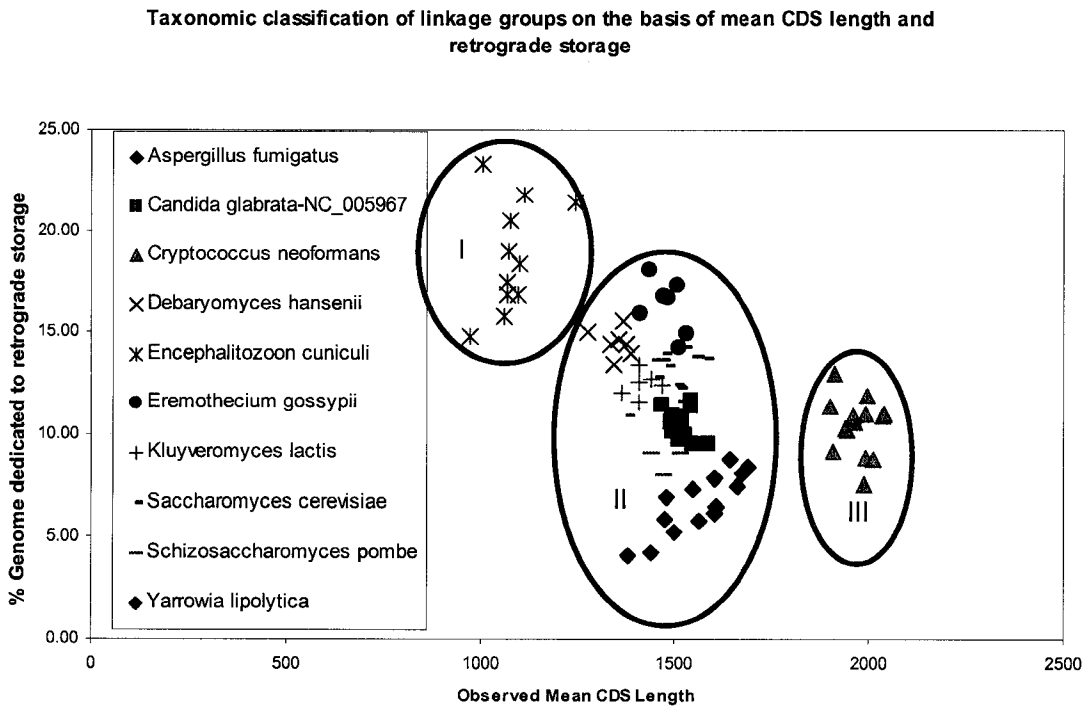

Figure 13 Observed mean CDS length and percentage of linkage group retrograde encoding taxonomically classify fungal chromosomes. Cluster I is comprised solely of the chromosomes from the microsporidian *E. cuniculi*. Cluster II comprises all the Ascomycota chromosomes. Cluster III encompasses the chromosomes of the basidiomycete *C. neoformans*.

If the encoding strategy employed by a linkage group had an affect on increasing the coding capacity towards the optimal limit, one might expect a correlation between the percentage of the linkage group coding in a particular fashion and a reduction in the difference between the Observed and Expected mean CDS length. Investigating the correlation between encoding strategy and total capacity revealed that the percentage of a

linkage group discreetly encoded correlated strongly (r >0.95) with the approach of the Observed mean CDS length to the expected value for 100% capacity (Figure 14).

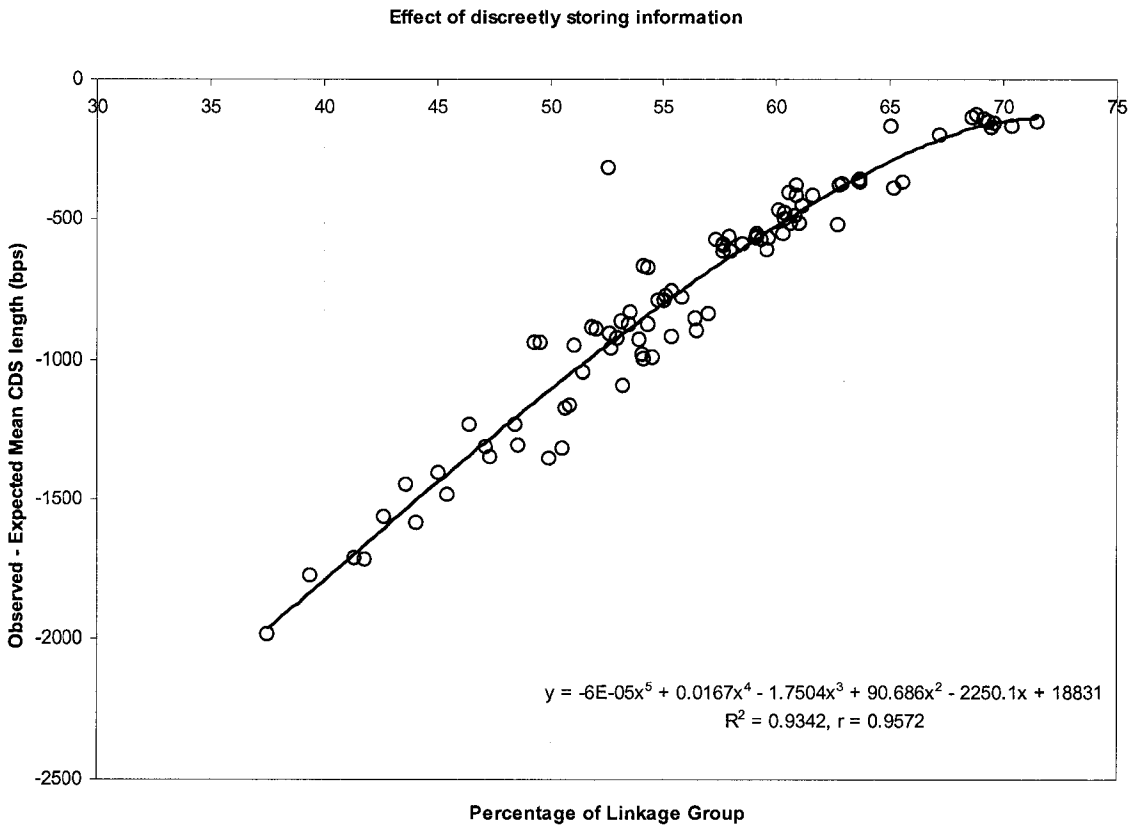

Figure 14 Correlation between the optimization of CDS feature length and discreetly encoded CDS features. Fungal chromosomes are plotted according to the percentage of each linkage group dedicated to discreet encoding versus the variance between observed and mean CDS length.

Similarly the organization of CDS features in a retrograde fashion also correlated strongly (r > 0.91) with the approach of the mean CDS length to the expected limit as illustrated in Figure 15.

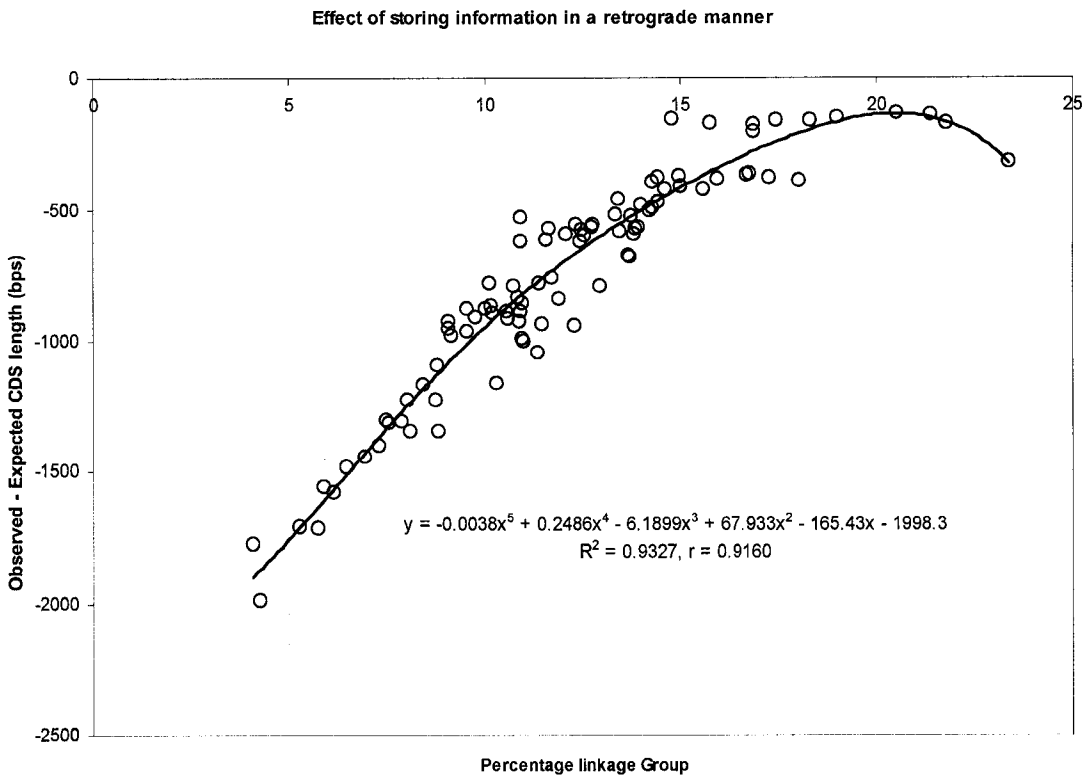

Figure 15 Correlation between the optimization of CDS feature length and retrograde encoded CDS features. Fungal chromosomes are plotted according to the percentage of each linkage group dedicated to retrograde encoding versus the variance between observed and mean CDS length.

The strongest affect on the mean CDS length ($r > 0.97$) was revealed by the regression of the coding capacity accounting for both discreet and retrograde strategies against the mean CDS length (Figure 16).

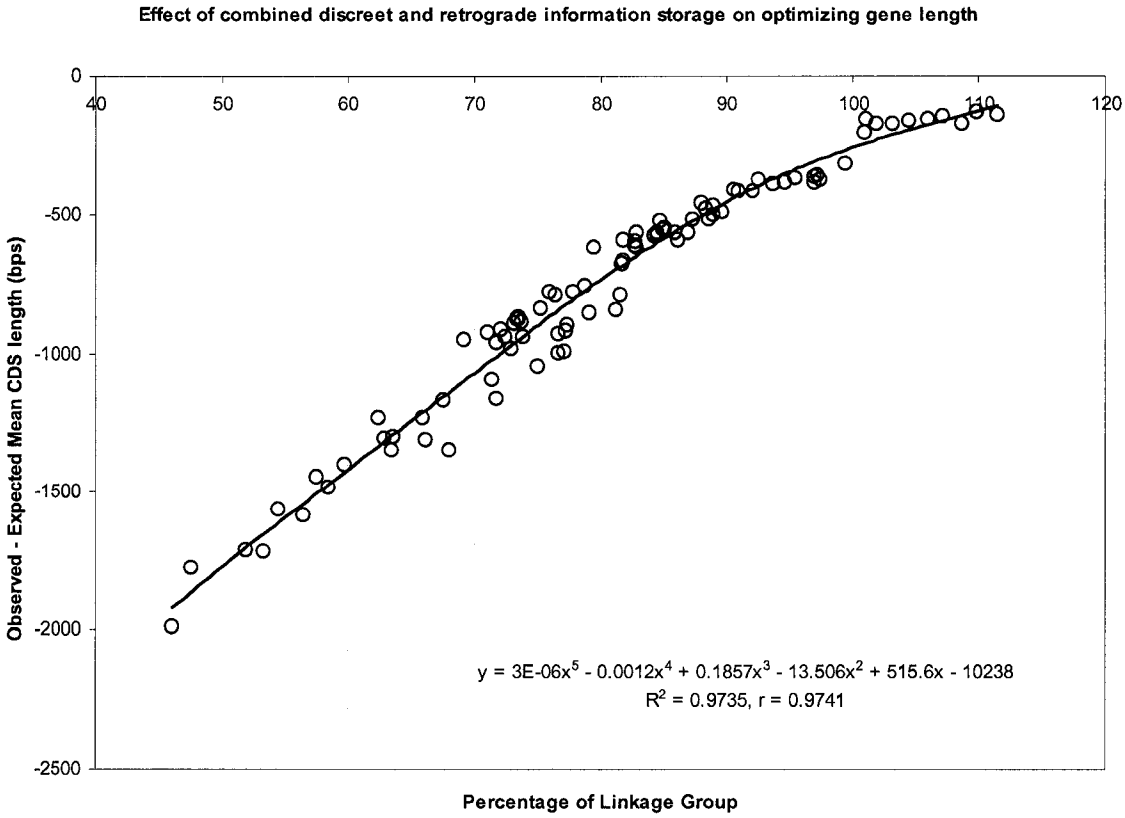

Figure 16 Correlation between the optimization of CDS feature length and both discreet- and retrograde-encoded CDS features. Fungal chromosomes are plotted according to the percentage of each linkage group encoding CDS information (accounting for both retrograde and discreet encoding) versus the variance between observed and mean CDS length.

In order to assess whether a 100% capacity limit was encountered *in vivo*, the coding capacity for each linkage group was calculated so as to account for both discreet and retrograde encoding (Figure 17). Graphing the percentage of the linkage group dedicated to discreet versus retrograde CDS features it was observed that only those linkage groups belonging to *E. cuniculi* exceeded the 100% capacity (Figure 17). Indeed, only one linkage group from *E. cuniculi* was found to fall below an effective capacity of 100%.

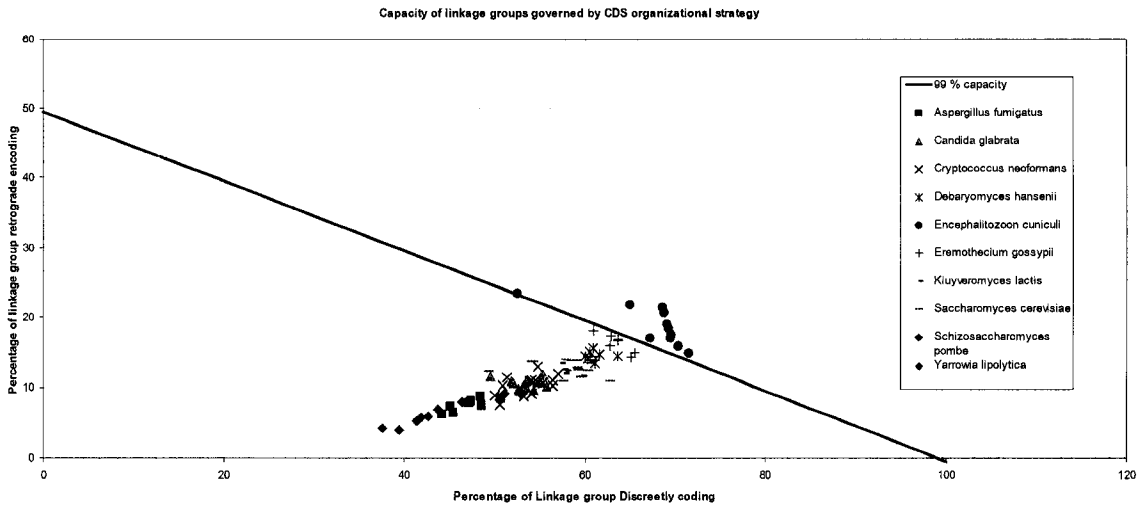

Figure 17 Effect of discreet and retrograde coding on the capacity of fungal linkage groups. Fungal chromosomes are plotted according to the percentage of each linkage group dedicated to discreet encoding versus the percentage of the chromosome dedicated to retrograde encoding. A theoretical 100% capacity is plotted.

Examining the characteristics of fungal linkage groups by organism (Table 2), it was found that *E. cuniculi* presented the smallest genome (2.5 mbps) while having the highest approximate coding capacity (~105% per linkage) and the observed mean CDS length closest to the expected. Further, the basidiomycete *C. neoformans* had an average capacity of 74.5% while the Ascomycota ranged from 51.8%-96.1% (extremes were *Y. lipolytica* and *E. gossypii* respectively). Therefore this work demonstrated that there is a connection between the coding capacity of fungal linkage groups and their taxonomic lineage.

Table 2 Average linkage group characteristics by organism.

				Mean percentage across linkage groups		
Organism	Genome size (mb)	Mean linkage length (mb)	Observed - Expected mean CDS length	Discreetly coding	Retrograde coding	Mean Coding Capacity
Aspergillus fumigatus	29.38	3.67	-1352	47.1	7.6	62.3
Candida glabrata	12.28	0.94	-857	53.4	10.5	74.5

Cryptococcus neoformans	19.05	1.36	-1011	53.8	10.4	74.5
Debaryomyces hansenii	12.22	1.75	-431	61.2	14.5	90.2
Encephalitozoon cuniculi	2.50	0.23	-173	67.4	18.7	104.9
Eremothecium gossypii	8.74	1.25	-374	63.5	16.3	96.1
Kluyveromyces lactis	10.69	1.78	-585	58.2	12.5	83.1
Saccharomyces cerevisiae	12.07	0.75	-591	58.4	13.0	84.4
Schizosaccharomyces pombe	12.53	4.18	-1035	50.2	8.7	67.6
Yarrowia lipolytica	20.50	3.42	-1698	41.1	5.4	51.8

Discussion

The C-value paradox was found to break down in response to the genome complexity exhibited by eukaryotes. In particular, the C value paradox was contradicted by the ability of prokaryotes and eukaryotes to organize genes in a retrograde fashion. Shown here is the finding that the combined impact of discreet and retrograde organization of genic information can serve to optimize the capacity of linkage groups. Upon examining the effect of retrograde and discreet organization of CDS features it was found that both schemes correlated strongly ($r > 0.91$ and $r > 0.94$ respectively) with the trend of the observed mean gene length versus the optimal expected length. This result is indicative of a situation where the exploitation of discreet and retrograde organization serves to drive the capacity of a linkage group to its upper limit. In the case of eukaryotic linkage groups there are localized regions where genes are present at disproportionately lowed densities. Heterochromatic regions and in particular centromeres typically contain fewer genes than the rest of the linkage group. Thus the findings of this study begin to explain how a linkage group can effectively store information (protein coding genes) in complex schemes and thereby compensate for regions of low gene density such as centromeres.

E. cuniculi represents a case where linkage groups can exceed 100% capacity through the exploitation of both discreet and retrograde organization of CDS features. *E. cuniculi* is also unusual in that it is an obligate intracellular organism which lacks a mitochondrion and whose linkage groups organize information such as to suggest an unusually high selective pressure to maintain small linkage groups. Shown here is the finding that *E. cuniculi*, through the use of complex organization of CDS features, is able to pack information at a density beyond 100% capacity relative to its length. Of the ten organisms included in this study, *E. cuniculi* demonstrated the closest approach to the optimal mean CDS length, suggesting that *E. cuniculi* has a collection of optimally organized linkage groups. These data imply that the exploitation of both discreet and retrograde organization is important in evolutionary situations where linkage groups must maximize their capacity.

As a microsporidian, *E. cuniculi* presents the smallest known eukaryotic genome sequenced to date and demonstrates a coding capacity of 105% per linkage group. *E. cuniculi* does not harbour a mitochondrial genome and yet encodes a chaperonin gene (*HSP70*) of mitochondrial origin on chromosome XI [21]. These findings suggest that the added coding capacity exhibited by *E. cuniculi* has been driven by an evolutionary need to incorporate additional organellar functions into its nuclear linkage groups while at the same time being under significant evolutionary pressure to maintain the small genome exemplified by this intracellular organism.

As sequencing technologies continue to achieve greater throughput at lower cost it may become feasible to profile unknown microbial samples and to taxonomically classify the organism composition on the basis of coding capacity seen in the sample. If

the capacity of a linkage group is evenly distributed throughout its length then any sub-sequence of the linkage group should present the same characteristics as the whole. Thus it may be possible to sequence genomic DNA from an unknown organism and thereby taxonomically identify the class to which the organism belongs. Such a technique would provide mechanism for taxonomic classification which is not dependant on morphological or physiological characteristics and may therefore be less prone to error.

The opportunity may also exist to profile complex samples (meta-genomics). If the capacity of a linkage group is taxonomically unambiguous, it may be possible to identify the presence of specific organisms within the sample. Further, if the ploidy of the organisms present is known then it may be possible to quantify the presence of each organism based on mean number of times each linkage group is sampled. By incorporating information about linkage group capacity it may also be possible to cluster sequences on the basis of their capacity, and thereby distinguish similar genes (paralogs) which reside on different linkage groups. Lastly, if the capacities of linkage groups differ significantly between organisms sampled it may be tractable to exploit the difference and aid in the process of genome assembly. Current advances in DNA sequencing enable the rapid acquisition of short DNA sequences (Pyrosequencing and MPSS). By applying techniques such as those in this study it may be possible to cluster partial sequence data into clusters reflective of the linkage groups to which they belong. Subsequent assembly for each linkage group may allow for the assembly of multiple genome sequences from meta-genomic data.

Measuring the capacity of linkage groups to transmit and organize CDS features serves to discriminate between microsporidia, ascomycetes, and basidiomycetes (Figure

13). This finding is a potentially powerful tool because it advocates for the divergence between these families on the basis of how their linkage groups organize information. In the case of *E. cuniculi,* molecular phylogenetics have proven contentious due to inherent statistical problems in the analyses [54]. The use of specific markers (loci) for molecular phylogenetic analysis is a common practice since these markers can be exploited prior to the derivation of a complete genome sequence. Presented here is a methodology for the assessment of the linkage group as a whole. By classifying linkage groups on the basis of their capacity, it is possible to taxonomically classify linkage groups belonging to microsporidians as distinct from ascomycetes and basidiomycetes.

Close scrutiny of the C-value paradox as shown in this study has revealed a break-down in diversity of gene structure in the context of chromosomal organization. In particular the C-value paradox is confounded due to the high variability in gene length when considering the diversity of the information encoded into genes including: structural, transcriptional, translational or possibly other, as yet unclassified features of DNA for which a structural basis in DNA is important. By recasting the focus of the C-value paradox in terms of the organization of genes and measuring the affect on the optimization of gene length it is possible to reveal predictive and strong correlations among known eukaryotes. Thus future work to evaluate these trends in higher eukaryotes will likely be possible with new advances in automated DNA sequencing and the subsequent public availability of their genome sequences.

Presently there are 38 fungal genomes actively being sequenced. Future work to incorporate the sequence data from the 38 fungi (32 Ascomycetes, 4 Basidiomycetes and 2 others) with the data presented here would provide a useful expansion of the data and

further evaluate the trends presented in this study [55]. Analysis of higher eukaryotes is warranted and with many genomes actively being sequenced the ability to perform analyses, as presented in this study, with other taxonomic divisions will be possible in the near future (Table 3).

Table 3 Summary of organisms for which a genome sequence is completed or in progress [55]

z	Group	Name	Status	Haploid chromosomes	Genome size Mb
Metazoans	Fish	Oryzias latipes	Draft	24	1000
Metazoans	Fish	Takifugu rubripes	Draft	N/A	N/A
Metazoans	Fish	Tetraodon nigroviridis	Draft	21	380
Metazoans	Insect	Drosophila melanogaster	Complete	4	180
Metazoans	Insect	Aedes aegypti	Draft	3	800
Metazoans	Insect	Anopheles gambiae	Draft	3	220
Metazoans	Insect	Apis mellifera	Draft	16	200
Metazoans	Insect	Bombyx mori	Draft	28	530
Metazoans	Insect	Drosophila ananassae	Draft	4	150
Metazoans	Insect	Drosophila erecta	Draft	4	150
Metazoans	Insect	Drosophila grimshawi	Draft	4	150
Metazoans	Insect	Drosophila mojavensis	Draft	4	150
Metazoans	Insect	Drosophila pseudoobscura	Draft	4	120
Metazoans	Insect	Drosophila simulans	Draft	4	150
Metazoans	Insect	Drosophila virilis	Draft	4	150
Metazoans	Insect	Drosophila willistoni	Draft	4	150
Metazoans	Insect	Drosophila yakuba	Draft	4	180
Metazoans	Insect	Pediculus humanus corporis	Draft	N/A	110
Metazoans	Insect	Tribolium castaneum	Draft	10	200
Metazoans	Mammals	Homo sapiens	Complete	23	3000
Metazoans	Mammals	Mus musculus	Complete	20	2500
Metazoans	Mammals	Bos Taurus	Draft	30	3000
Metazoans	Mammals	Canis familiaris	Draft	39	2400
Metazoans	Mammals	Cavia porcellus	Draft	31	3400
Metazoans	Mammals	Dasypus novemcinctus	Draft	32	3000
Metazoans	Mammals	Equus caballus	Draft	32	N/A
Metazoans	Mammals	Felis catus	Draft	19	3000
Metazoans	Mammals	Gallus gallus	Draft	39	1200
Metazoans	Mammals	Gorilla gorilla	Draft	N/A	N/A
Metazoans	Mammals	Macaca mulatta	Draft	22	N/A
Metazoans	Mammals	Monodelphis domestica	Draft	9	N/A
Metazoans	Mammals	Myotis lucifugus	Draft	N/A	N/A
Metazoans	Mammals	Oryctolagus cuniculus	Draft	22	3500

Metazoans	Mammals	Otolemur garnettii	Draft	31	N/A
Metazoans	Mammals	Pan troglodytes	Draft	24	3100
Metazoans	Mammals	Pongo pygmaeus	Draft	24	3000
Metazoans	Mammals	Rattus norvegicus	Draft	21	2800
Metazoans	Mammals	Sorex araneus	Draft	N/A	3000
Metazoans	roundworms	Caenorhabditis elegans	Complete	6	97
Metazoans	roundworms	Caenorhabditis briggsae	Draft	6	100
Metazoans	roundworms	Caenorhabditis remanei	Draft	N/A	N/A
Metazoans	Worms	Schmidtea mediterranea	Draft	4	480
Metazoans		Aplysia californica	Draft	17	1800
Metazoans		Ciona intestinalis	Draft	14	160
Metazoans		Ciona savignyi	Draft	N/A	180
Metazoans		Strongylocentrotus purpuratus	Draft	N/A	800
Plants	Green Algae	Oltmannsiellopsis viridis	Complete	1	N/A
Plants	Green Algae	Ostreococcus lucimarinus	Complete	21	13
Plants	Green Algae	Pseudendoclonium akinetum	Complete	1	N/A
Plants	Land Plants	Arabidopsis thaliana	Complete	5	120
Plants	Land Plants	Oryza sativa	Complete	12	390
Plants	Land Plants	Populus trichocarpa	Draft	19	480
Plants	Land Plants	Vitis vinifera	Draft	19	500

As can be seen in Table 4 it will be possible to evaluate the mechanisms exploited by higher eukaryotes and whether they abide the trends presented here in the analysis of the model fungal genomes. Of particular interest is whether mammals, insects and land plants organize there CDS features evenly throughout the length of their linkage groups as seen in the fungi.

Table 4 Summary of number of genome sequences in progress by Group among higher eukaryotes.

# of organisms	Group
19	Mammals
16	Insect
4	Land Plants
3	Green Algae
3	Roundworms
3	Fish
1	Worms

It is likely that examination of genome sequences derived from higher eukaryotes will present a greater degree of bias in how information is encoded within linkage groups as based on the occurrence of structural chromosomal features (E.g. peri-centromeric regions). While the fungal data presented in this study serves as a basis for examining chromosomes as a whole, future work in higher eukaryotes should explore whether there are local biases in encoding strategy. Through *in silico* fragmentation of chromosome sequences it is possible to perform an analysis similar to that presented in this study which would evaluate the localized occurrence of encoding bias. By examining correlations in local encoding bias which correlate with structural components such as telomeres, centromeres and methylation levels it may be possible to determine whether structural chromosomal features are connected with how chromosomes encode information.

The strategies exhibited by the eukaryotic chromosomes analyzed in this study suggest that significant relationships exist among organisms with respect to *how* they encode information into chromosomes. In conjunction with the findings of Chapter I **Genome profiling on the basis of coding gene length**; these studies provide insight into the taxonomic significance of *what* information is encoded by an organism (gene length) and *how* information is encoded (encoding strategy). For protein coding information to affect the adaptation of an organism the underlying genic information must be transcribed, subsequently translated and folded into a functional protein. While selective pressures exerted on an organism likely effect *what* information is encoded by a genome there are potentially other factors which affect *how* information is optimally organized within a chromosome. In Chapter III (**Wavelet analysis of Eukaryotic Linkage groups**)

I focus on the use of a novel technique to derive putative targets of DNA methylation as yet another example of chromosomal information which may be involved in the suppression of gene transcription. If DNA methylation provides a signal for the suppression of transcription and the subsequent reduction in the ability of chromosomal regions to affect adaptation then the distribution of DNA methylation sites, akin to CDS length distributions, may provide an insight into the adaptive strategy of an organism and thereby the taxonomic relationship amongst organisms.

CHAPTER III

WAVELET ANALYSIS OF EUKARYOTIC LINKAGE GROUPS

Introduction

Eukaryotic chromosomes serve to transmit genic information through cell divisions. Chromosomes are not static structures which simply carry information but rather are dynamic sub-cellular structures which function to, themselves; regulate how genetic material is accessed. The formation of chromatin as the collection of chromosomal DNA, proteins and RNA serves to package DNA in order to condense chromosomes. For most eukaryotes the re-modeling of chromatin, from loosely packaged euchromatin into densely packaged heterochromatin, serves to silence the expression of the underlying genes [56]. Of particular scientific interest is the observation that the packaging of chromatin is stably inherited. For chromosomes to exhibit structural functions which are heritable there should be some form of signal, local to the heterochromatic regions which mark them as inactivated with respect to gene expression. In most eukaryotes methylation of Cytosine (Figure 18 and Figure 19) residues in heterochromatic regions, commonly centromeres and telomeres, serves to epigenetically silence the underlying genes.

The epigenetic silencing of chromosomal regions serves significant functions in the reduction of transposon proliferation, imprinting, and targeted regulation of gene expression [57]. Genes involved in the regulation and maintenance of DNA methylation have been shown to be requisite for normal development for both mammals and the higher plant, Arabidopsis [58,59]. In mammals the *de novo* methylation of Cytosine is regulated by *DNMT3a* and *DNMT3b* while maintenance of methylation is maintained

through the action of *DNMT1*. Arabidopsis has a homolog for *DNMT1* in the locus *MET1*. While the function of *DNMT1/MET1* is the preferential methylation of Cytosine bases in CpG di-nucleotides there are a set of plant specific loci, Chromomethylases (e.g. *CMT3*) which methylate Cytosine residues at non CpG locations.

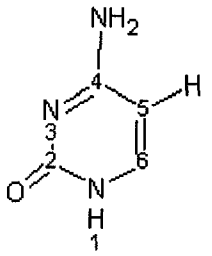

Figure 18 Molecular structure of Cytosine.

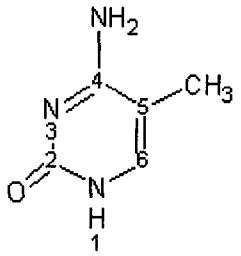

Figure 19 Molecular structure of 5-Methyl-cytosine

The process of DNA methylation, particularly on CpG di-nucleotides is of wide taxonomic importance due to DNA imprinting [60-67]. By creating an epigenetic signal Cytosine methylation can serve to delineate local regions where the function of a chromosome is different. Nucleotide composition varies between genomic feature types (protein coding genes, RNA genes, spacers, and regulatory regions) [68-70]. Isochores, in particular are identified by their relative enrichment in CpG di-nucleotide composition. Given that high levels of methylation at CpG di-nucleotides is associated with gene silencing the ability to locate and identify genomic regions where CpG di-nucleotides occur and where they are likely to be methylated is of importance to most eukaryotes.

Such analysis of methylation signals within a chromosome can serve to identify gene-rich regions without *a priori* knowledge of the underlying gene models.

Signal analysis is a process of analysing numerical signals through the identification of what component frequencies are present and where in a signal they are resident. By identifying the component frequencies of a signal and where they change within a signal it is possible to identify the component behaviours giving rise to a complex signal. For example in the analysis of music, signal analysis would focus on the identification of instrumentation (component frequencies) and rhythm (how frequencies are changing with respect to time). Fourier analysis is the most commonly applied technique for signal analysis. By decomposing a signal into a set of sine waves of varying frequencies, Fourier analysis is able to identify what frequencies are present in the signal based on the component sine waves. The major limitation to the use of Fourier analysis is that Fourier does not directly localize where in a signal a sine wave of a given frequency is present. Rather, Fourier provides insight into the spectra of frequencies observed in a signal (e.g. instrumentation). By decomposing a signal into sine waves Fourier determines what given frequencies are present anywhere within the signal and is therefore capable of determining relative intensities of the component sine waves but not the direct localization of where a given frequency exists in a signal.

Signals can be broken into windows with the subsequent windows being analyzed by Fourier analysis. A resulting windowed Fourier analysis would be capable of detecting component frequencies and would provide initial insight into the locality of a component frequency. However, the precision with which Fourier analysis can localize a component frequency would be across the entire window in which it was identified. By examining

DNA sequences using signal analysis it is possible to identify specific periodicities within genomic DNA sequences of biological relevance. Tsonis *et. al* demonstrated that through the use of Fourier analysis it was possible to determine that there was a markedly non-random periodicity in the spectrum related to coding sequences as opposed to non-coding sequences [71-73].

Wavelet analysis is a localized technique for signal analysis where signals are decomposed into frequency components with respect to time. Where Fourier analysis relies on the deconstruction of a signal in terms of component sine waves, Wavelet analysis employs Wavelets which are scaled and translated through the length of a signal to decompose a given signal. A Wavelet is a mathematical function which is square-integrable with respect to an orthonormal set of basis functions. By decomposing function $f(t)$ in terms of a set of coefficients c which given a wavelet ψ at a scale of j and a translation of k in time (Equation 8 and Equation 9) it is possible to decompose the signal into a set of low and high frequency components.

$$\psi_{j,k}(t) = \psi(2^j t - k)$$

Equation 8 Wavelet basis as derived from a wavelet at scale j and translation k within time t.

$$f(t) = \sum c_{j,j} \psi_{j,k}(t)$$

Equation 9 Wavelet based definition of the function f(t).

By identifying the locations corresponding to a coefficient of 1, it is possible to identify regions in a signal where locally the signal is behaving analogous to the analyzing Wavelet. In the case of Wavelets which represent significant transitions from high to low amplitudes (e.g. the Daubechies Wavelets) it is possible to identify positions where the component frequencies within a signal change. While Fourier analysis can only

rationalize global signal properties such as frequencies, Wavelet analysis is capable of identifying the location of specific frequencies within a signal. By decomposing signals into frequency components localized to specific locations within a signal (e.g. time) it is possible to exploit Wavelet analysis to identify where, in a signal, frequencies are changing [72,74].

Wavelets have the further advantage over Fourier analysis, in that the use of wavelets in signal analysis results in a level of detail which is scaled to the region of signal being analyzed. The act of scaling a Wavelet by an integer j is analogous to compressing the Wavelet when $j < 1$ and expansion of the Wavelet when $j > 1$. Thus when signal analysis is performed using Wavelets, the ability exists to identify periodicities across a range of scales and thereby derive inferences regarding the underlying signal behaviour which are both local as well as global.

In order to analyze DNA sequences as signals it is necessary to encode a nucleotide sequence of characters into a numerical sequence. DNA walks are numerical encodings of DNA sequences where a counter is incremented based on the composition of the DNA sequence. Mono-nucleotide walks are designed such that a single nucleotide is scored +1 (e.g. Adenine) while all others are scored -1 (e.g. Thymine, Guanine, Cytosine). A Purine walk (di-nucleotide) would be one where any occurrence of Adenine or Guanine would result in a +1 and occurrences of the Pyrimidines (Thymine and Cytosine) would score -1. As numerical sequences DNA walks are typically analyzed using multi-scale wavelet decomposition. By decomposing the DNA walk across a series of scales it is possible to asses whether there are periodic features present in the signal.

Work to date has focused on the use of mono-nucleotide and di-nucleotide walks of genomic DNA sequences [75,76]. Audit et al. showed that there were two significant periodicities which showed long range correlations: in the one instance the period was less than 200 bps, while in the other periods were defined as greater than 200 but less than 1000 bps. It has been proposed that a Purine walk would be modified from a random state by the elevation / depression of Purine levels and that these levels would correlate with recombination efficacy. Previous work applying wavelet analysis to Purine walks has identified a long range correlation which exhibits a 10bp periodicity. Further, it has been demonstrated that a significant correlation existed between the 10bp periodicity identified through wavelet analysis and the occurrence of nucleosome motifs and, further, that the long range correlation seen in the 10bp periodicity were more dependent on di-nucleotide encoding than mono-nucleotide encoding.

Previous work has demonstrated that long range correlations ($w < 200$, where w is the length of the periodic information identified through wavelet analysis) identified through wavelet analysis were detectable among a wide range of organisms known to form nucleosomes, which included some viruses [76]. Examination of eubacterial organisms revealed that the organisms did not demonstrate the same long range correlations. It has therefore been hypothesized that the long range correlations identified ($w < 200$) were related to nucleosome function as they were only detected in organisms known to form nucleosomes. By contrast long range correlations for larger feature classes ($200 \text{ bps} < w < 1000\text{bps}$) were identified in all organisms studied.

While work has identified component periodicities within genome sequences existing work has not focused on the localization of these Wavelet-based features.

Conversely, examination of Isochore boundaries has shown that wavelet analysis can be used to identify discreet features within genomic DNA [68,77]. Research has shown that once having identified Isochores through Wavelet analysis it was possible to estimate the organism of origin of a given Isochore on the basis of its GC content [68]. Using a DNA walk which favored G|C vs. A|T (+1 vs. -1 respectively) it was demonstrated that 'islands' of GC content could be identified. Further, through the identification of Isochores which were significantly different in GC profile it was demonstrated that there is a correlation between GC content and proposed gene transfer events.

Proposed in this study is a methodology for examining genomic DNA sequences using Wavelet analysis. Of particular focus in this work is the identification of CpG di-nucleotides and their presence at varying scales. By biasing the numerical encoding of DNA in order to preferentially weight CpG di-nucleotides it is hypothesized that Wavelet analysis can be exploited to detect biologically relevant patterns in the occurrence of CpG di-nucleotides. Further, through the use of a numerical encoding for DNA which favours CpG di-nucleotides it is hypothesized that the features identified would represent a set of putative features which were the targets for methylation of their Cytosine residues. The DNA sequence encoding used in this study is from A,G,C,T to 1,2,-2,-1 respectively [78]. By providing discrete numerical encodings on a per nucleotide basis, this encoding serves to maintain potential DNA sequence motif dependent information. The encoding of Adenine and Guanine as distinct (positive) from Thymine and Cytosine (negative) maintains, in the numerical encoding, the molecular distinction between Purines and Pyrimidines. By biasing the largest numerical change to the CpG di-nucleotide this encoding will make the detection of CpG di-nucleotides possible. Being symmetric about

the origin this encoding serves to be symmetric from A to T (1, -1) and G to C (2, -2). Therefore this encoding serves to mimic the semi-conservative nature of DNA replication. Further by equating the symmetry of C and G this encoding serves to equate the significance of CpG and GpC di-nucleotides which is analogous to examining the di-nucleotides as if they both occurred on the same strand. As such the wavelet analysis of either the Watson or Crick strand will bear the same result. The Daubechies' family of wavelets are commonly used to identify locations within a signal where there are breakpoints in the component frequencies. In particular Daubechies' wavelets are best suited for identifying locations in a signal where there are abrupt transitions in a signal (short duration and high frequency). Therefore the exploitation of the numerical encoding used here in conjunction with Daubechies' family of wavelets will provide a mechanism to identify features within Chromosomal DNA sequences based CpG di-nucleotides.

Materials and Methods

The genome sequences used in this study were comprised from the completed eukaryotic genomes publicly available via GenBank. In particular, the set of Hemiascomycetous yeast included: *Aspergillus fumigatus*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica*. Vertebrate genomes used included the genome sequences for: *Homo sapiens, Mus musculus, Rattus norvegicus, Canis familiaris,* and *Danio rerio.* Invertebrates examined included: *Drosophila melanogaster* and *Apis mellifera.* Lastly a single plant genome was used, *Arabidopsis thaliana.*

For each genome the corresponding linkage groups were analyzed using a custom software package Biolet (http://sourceforge.net/projects/biolet). Biolet is a compiled Matlab™ program designed to perform multi-scale wavelet decomposition on DNA sequences using the Daubechies family of wavelets (db2 – db12). Each linkage group was fragmented into 500,000bp sequences based on the need for a reasonable size fragment for computation on a host computer equipped with 2 Gbs of memory. Corresponding fragments of a linkage group were encoded from DNA sequences to numerical vectors. Encoding to a numerical vector is done by substituting the nucleotides A, G, C, T for 1, 2,-2,-1 respectively. The encoded sequence is subjected to wavelet decomposition for the Daubechies' wavelets (db2-db12) at scales 1-12. The resulting coefficients were de-noised as per [72] and then analyzed to find regions where the coefficients are indicative of changing frequency, regions where the coefficients were contiguous stretches of 1s. By identifying the contiguous coefficients Biolet is able to derive the set of genomic locations where there are CpG di-nucleotides. The resulting Wavelet-based features are exported as Generic Feature Format (GFF) files (Figure 20).

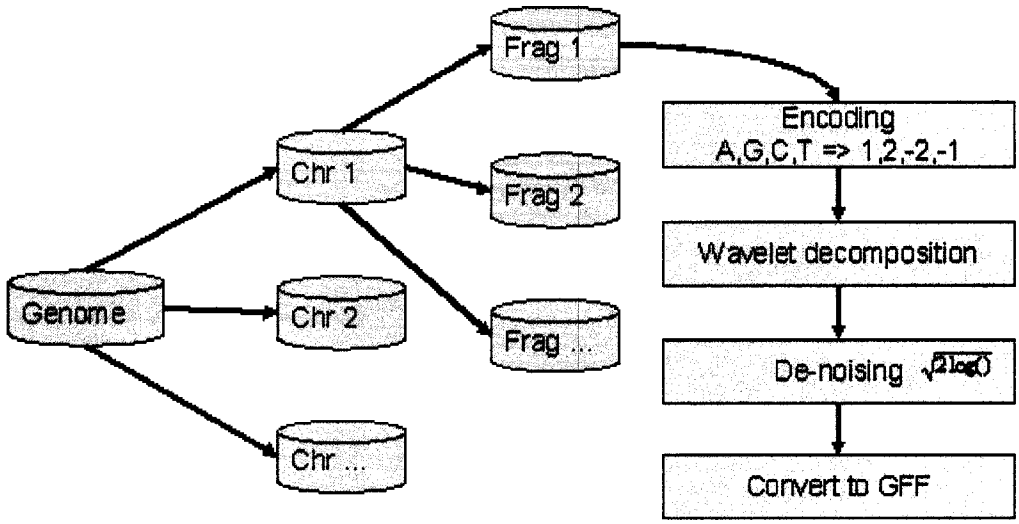

Figure 20 Wavelet analysis pipeline as performed by Biolet. Genome sequence was acquired as a set of independent chromosomes. Individual chromosome sequences were fragmented into 500Kbp sub-sequences which were then encoded as numeric signals that were subsequently analyzed using wavelet analysis. De-noising allowed the identification of significant wavelet features. The output of the analysis pipeline was a set of Gene Feature Files which were used in further analyses.

For comparative genomic analysis the distribution of Wavelet-based features for each organism was calculated. Wavelet-based features were classified into 100bp bins. Clustering and analysis of Wavelet-based feature distributions was performed using GenespringGX™ analogous to the analysis of CDS features (Chapter I: **Genome profiling on the basis of coding gene length**). Clustering was performed using K-means and hierarchical (gene tree) clustering.

A set of 235 random DNA sequences 500,000bp in length were generated and subsequently analyzed with Biolet in order to identify whether there was a bias in the wavelet decomposition related to the ends of the sequence (locations of high frequency changes in signal). In order to remove any bias of Biolet to the analysis of sequence termini, the distribution of random sequences was used to normalize each organism's

distribution based on the random feature distribution scaled to the size of the specific genome.

Results

An assessment of whether the wavelet analysis was detecting non-random features was performed by comparing the distribution of Wavelet-based features from all eukaryotes against data acquired from randomly generated sequences. Figure 21 shows the gene tree resulting from the clustering of all genomes according to the distribution of their Wavelet-based features. It is immediately apparent from comparison to random sequence data that the Wavelet-based features are non-random. In random sequence data the Wavelet-based features are limited to distinct length classes which, while present in many organisms, are not the major Wavelet-based feature classes. As such the distribution of Wavelet-based features spans a greater number of size classes in eukaryotes than is seen in random sequences. Further, there appears a distinct difference in the Wavelet-based feature distributions for mammals and honeybee as opposed the other organisms under study.

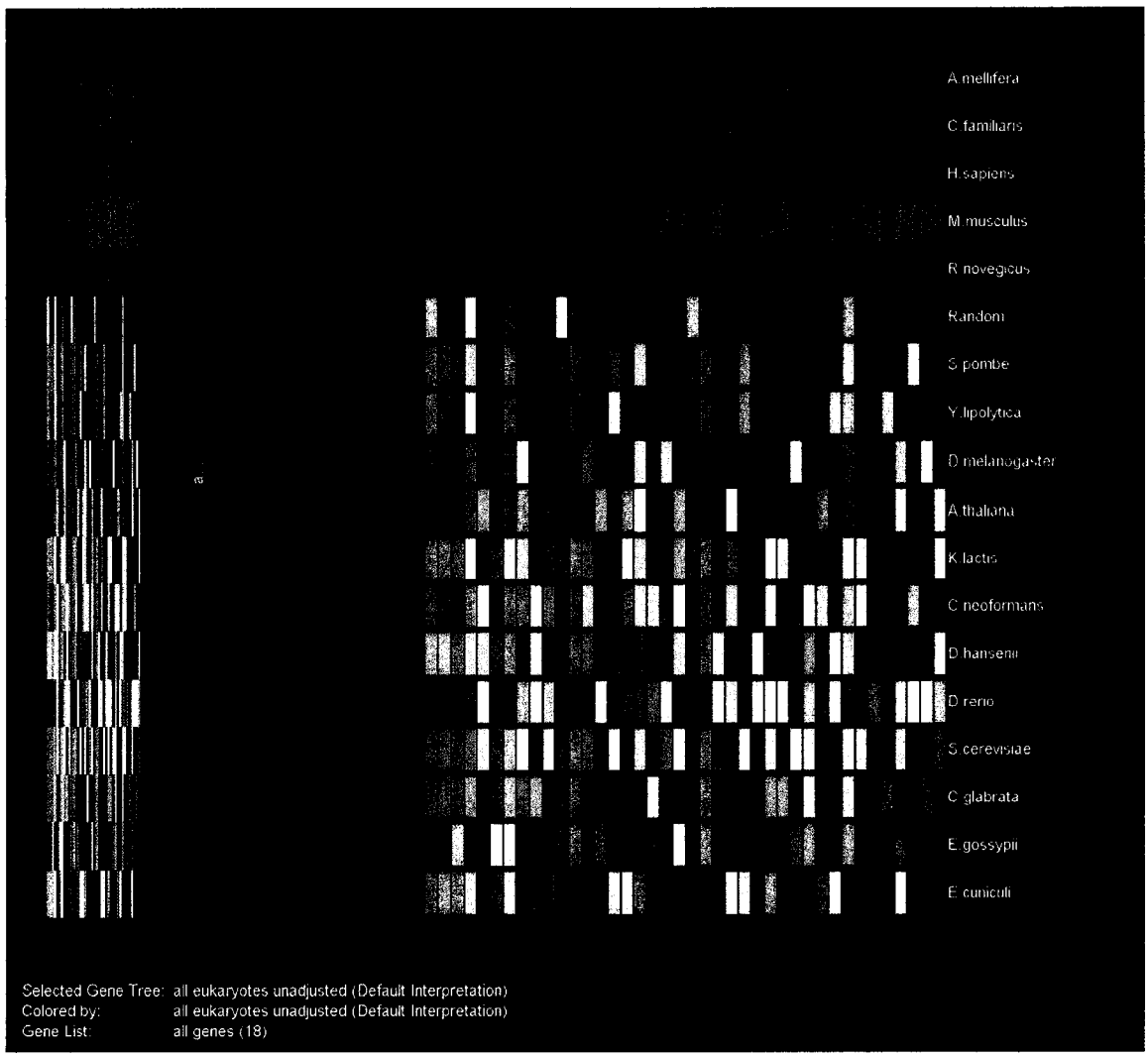

Figure 21 Analysis of eukaryotic genomes with DB wavelets. Hierarchical clustering was performed and wavelet length distributions are plotted as 'heatmap' pseudo-coloured boxes increasing in size from left to right in 1Kb increments. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome".

While the study organisms differed significantly from random it was still desirable to normalize the data against the results obtained from random sequences. The distribution of random features was used to clean up (normalize) the distribution for each organism. The random data was scaled to the genome size of each study organism and the

resulting number and size of randomly occurring Wavelet-based features were removed from further study.

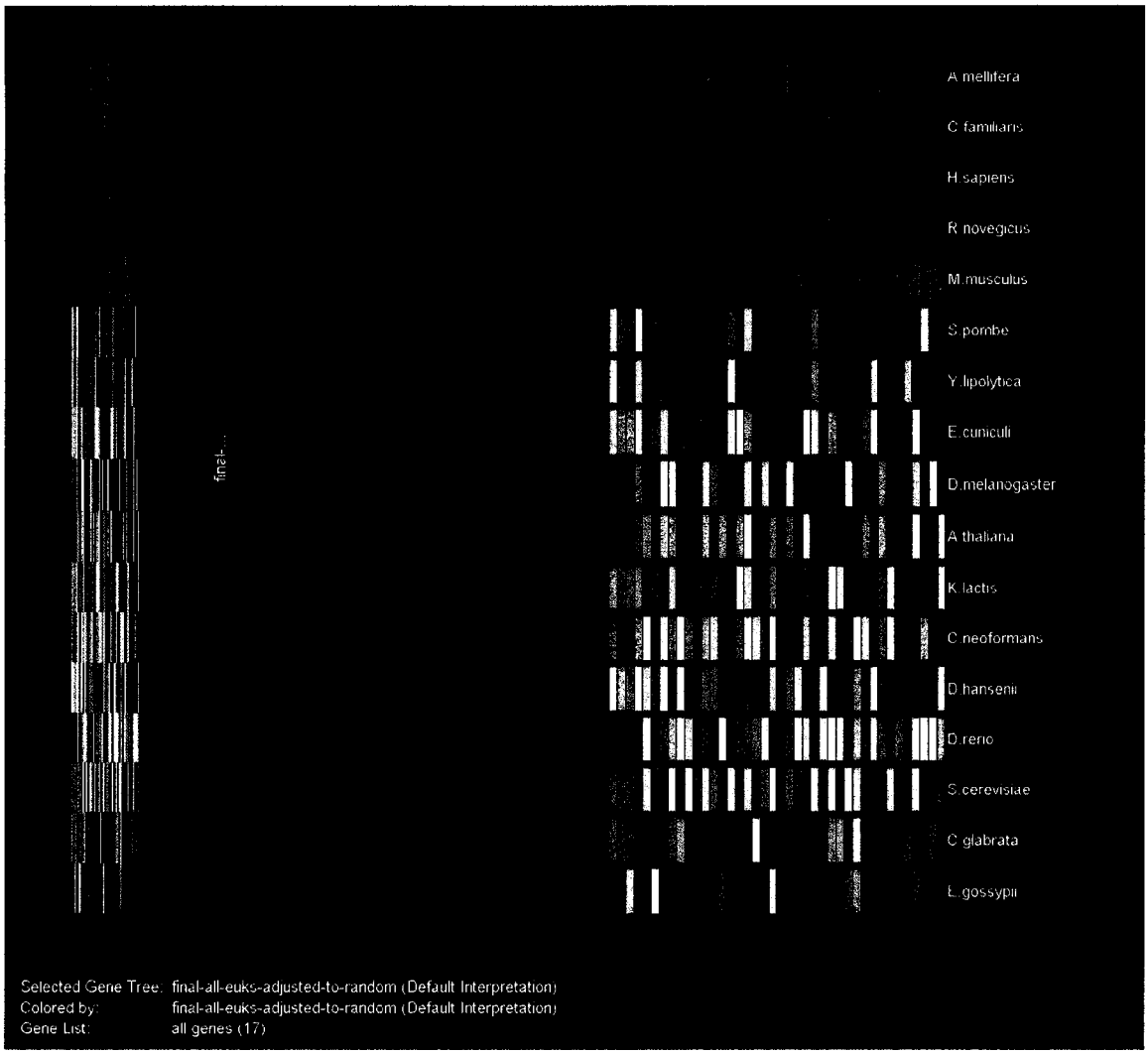

Figure 22 Analysis of eukaryotic genomes using DB wavelets. Data was normalized against randomized DNA sequences prior to performing hierarchical clustering. Hierarchical clustering was performed and wavelet length distributions are plotted as 'heatmap' pseudo-coloured boxes increasing in size from left to right in 1Kb increments. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome".

After removing the effect of random sequence bias it is further apparent that there is a high degree of correlation amongst mammals and honeybee as opposed to the fungal organisms under study (Figure 22). It is further evident that the distribution of Wavelet-

based features by size is more diverse in mammals and honeybee as opposed to fungi, Arabidopsis and Drosophila. Further investigation into the subset of fungal organisms revealed that while there is a large amount of correlation amongst mammals a similar degree of conservation is not seen amongst the fungi (Figure 23). This is due to the distribution of Wavelet-based features across multiple size classes and in particular the observation that the diversity of Wavelet-based feature size classes is much reduced in fungal organisms studied.

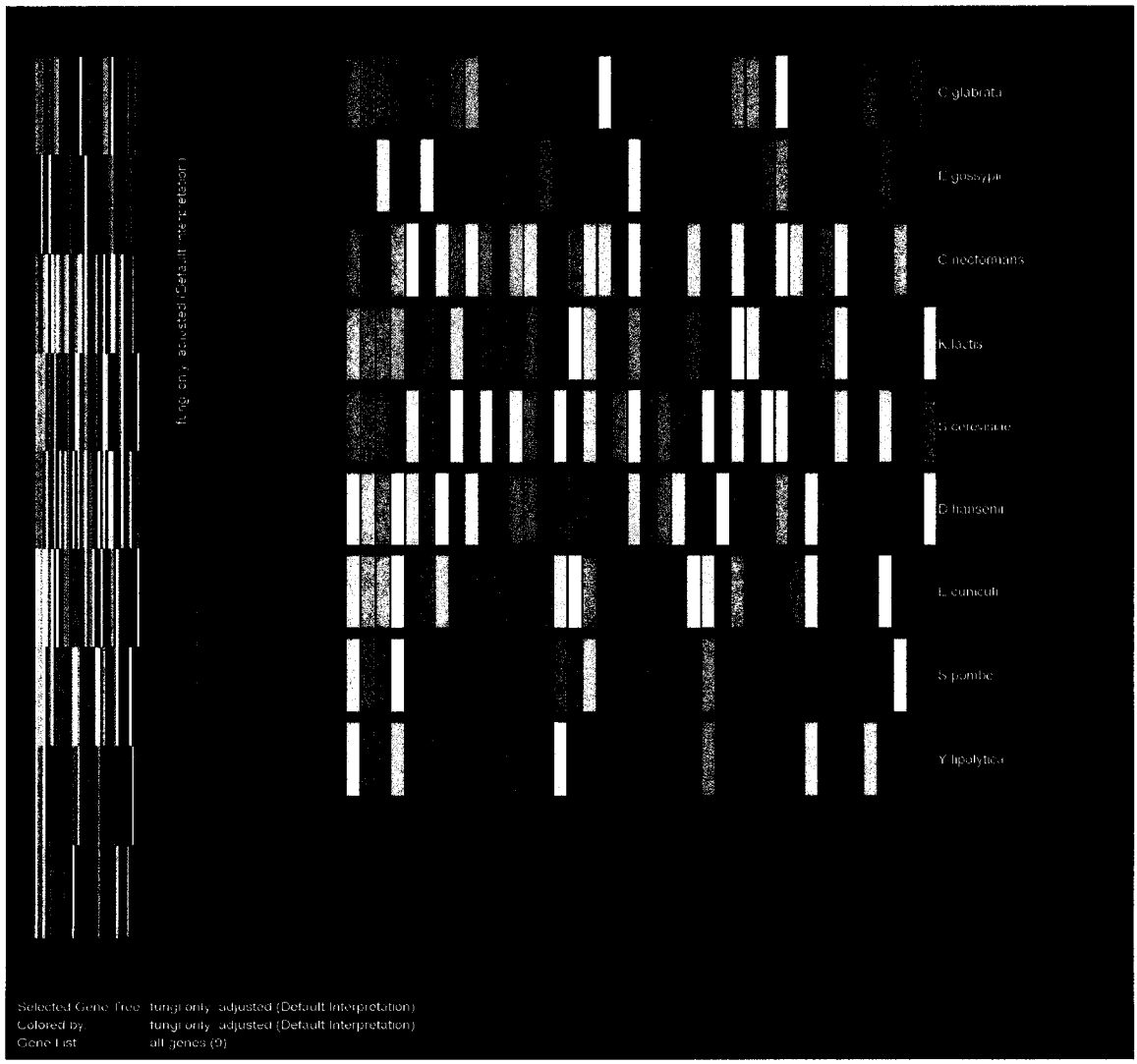

Figure 23 Wavelet-based feature distributions for fungal genomes. Hierarchical clustering is based on Pearson correlation measure with 10,000 bootstrapped replicates. Data was normalized against

randomized DNA sequences prior to performing hierarchical clustering. Hierarchical clustering was performed and wavelet length distributions are plotted as 'heatmap' pseudo-coloured boxes increasing in size from left to right in 1Kb increments. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome".

Examination of the non-fungal organisms (Figure 24) reveals that the high level of conservation amongst mammals does extend to include honeybee but surprisingly it does not extend to include Arabidopsis, Drosophila or D. rerio.

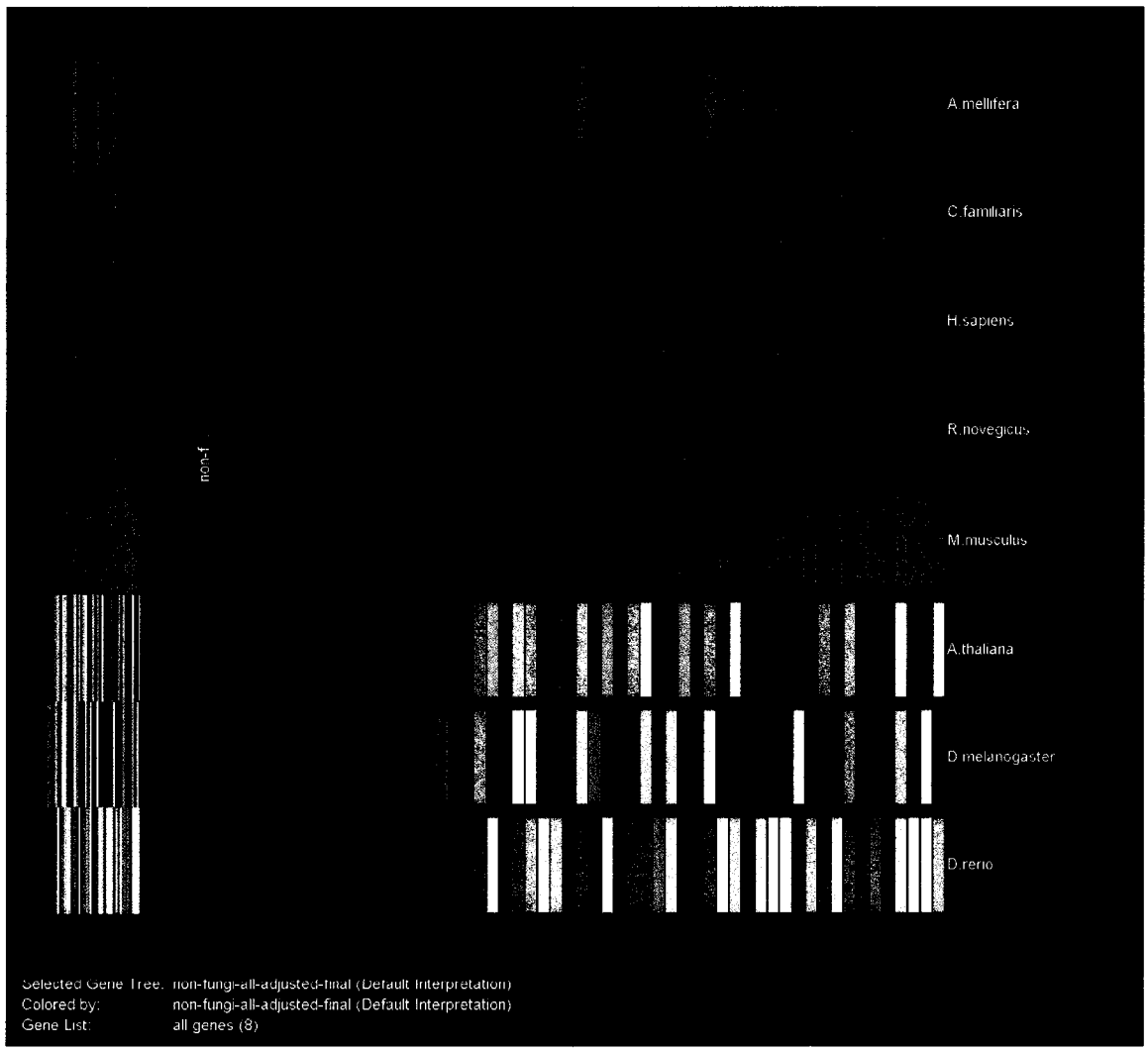

Figure 24 Analysis of higher eukaryotic genomes using DB wavelets. Data was normalized against randomized DNA sequences prior to performing hierarchical clustering. Hierarchical clustering was performed and wavelet length distributions are plotted as 'heatmap' pseudo-coloured boxes

increasing in size from left to right in 1Kb increments. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome"..

Figure 25 K-means (2) clustering of the distribution of wavelet feature lengths in higher Eukaryotes. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome".

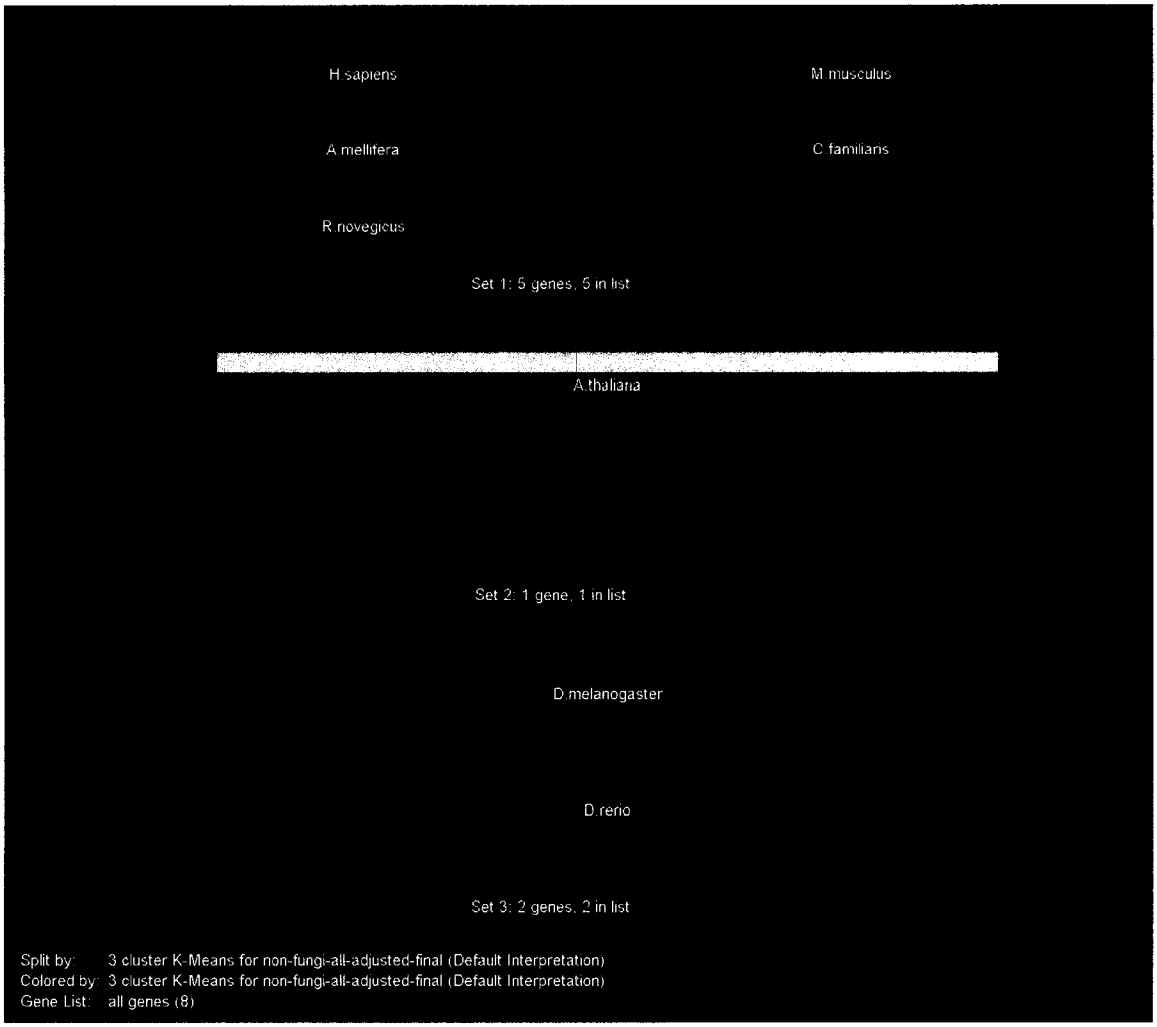

Figure 26 K-means (3) clustering of the distribution of wavelet feature lengths in higher Eukaryotes. Organisms are grouped and coloured by cluster. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome".

Figure 27 K-means (4) clustering of the distribution of wavelet feature lengths in higher Eukaryotes. Organisms are grouped and coloured by cluster. The term "gene" in the figure is a GeneSpring-GX™ reference which is analogous in this case to "genome".

Figure 28 K-means (5) clustering of the distribution of wavelet feature lengths in higher Eukaryotes. Organisms are grouped and coloured by cluster. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome".

In order to better understand the relationship amongst the Wavelet-based feature distribution of higher eukaryotes K-means clustering was performed (Figure 25, Figure 26, Figure 27, and Figure 28). Through increasing the numbers of clusters K-means clustering demonstrates that the relationship amongst mammals is preserved and the close association of honeybee to mammals is reinforced. Based on their Wavelet-based feature distributions the more distant eukaryotes were A. thaliana then D. melanogaster and finally D. rerio as the most distant.

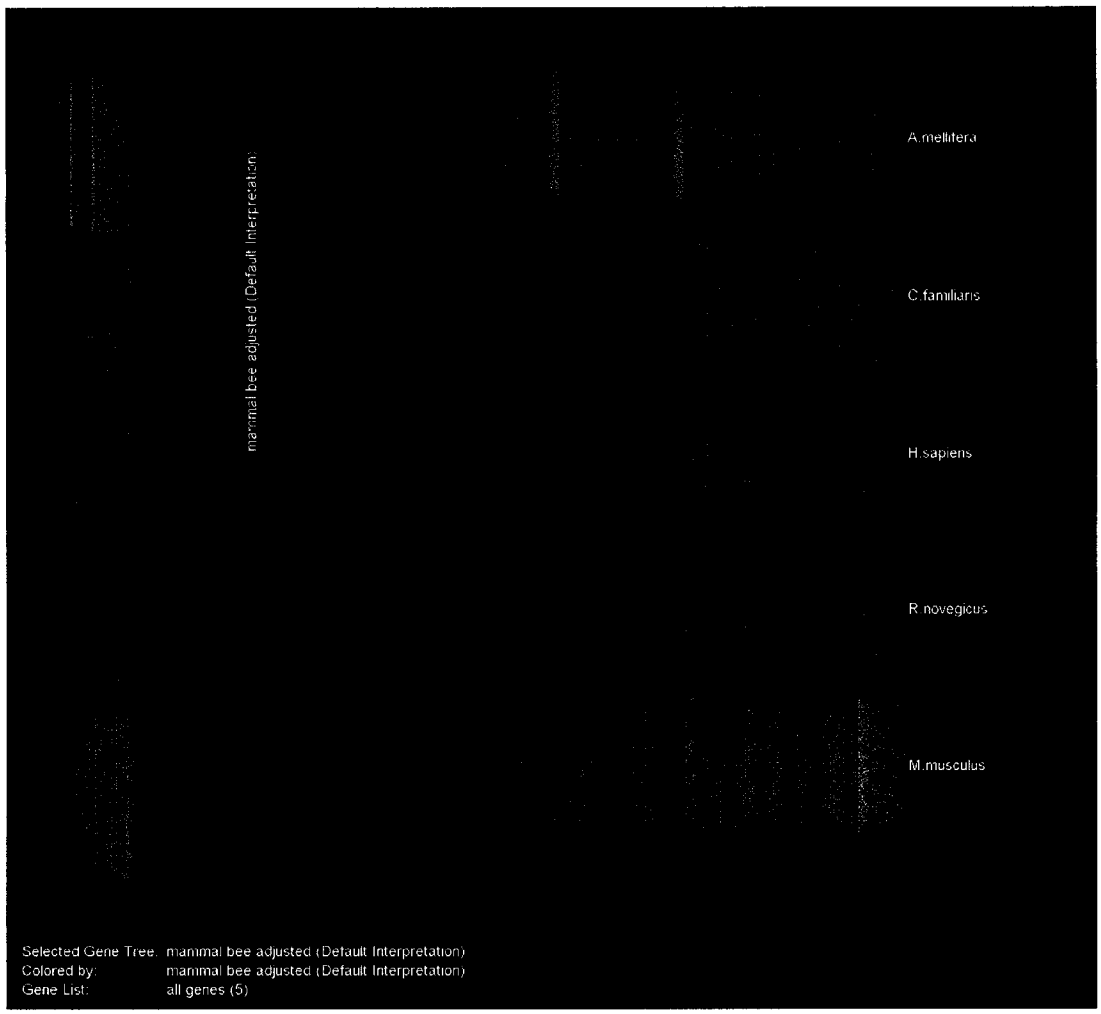

Figure 29 Genomes of honeybee and mammals with the DB wavelets. Data was normalized against randomized DNA sequences prior to performing hierarchical clustering. Hierarchical clustering was performed and wavelet length distributions are plotted as 'heatmap' pseudo-coloured boxes increasing in size from left to right in 1Kb increments. The term "gene" in the figure is a Genespring-GX™ reference which is analogous in this case to "genome"..

By performing a bootstrapped gene tree on the mammalian data in addition to that from honeybee it is possible see an initial taxonomic distinction amongst the mammals in this study. In particular Figure 29 demonstrates that mouse and rat cluster together more closely than they do with human and dog. Honeybee, while related, proves to be most distinct from the mammals analyzed.

Discussion

The data presented in this study support the ability of wavelet analysis to identify non-random features of genomic DNA. By focusing the numerical representation of a DNA sequence to bias CpG and GpC di-nucleotides (2,-2) or (-2, 2) the encoding serves to allow the preferential detection of CpG and GpC di-nucleotides over ApT and TpA. Thus the Wavelet-based features identified in this work are hypothesized to have importance in chromosomal function as targets for DNA methylation as cytosine within CpG islands is the primary target for DNA methyl-transferases [63].

The analyses presented here treat the genomes as a single entity and in doing so these data support observations at the organism level. Future work should focus on the examination of local Wavelet-based feature detection. The possibility exists to examine the genomes in this study in terms of the results by chromosome. It would be desirable to know whether the distribution of Wavelet-based features is consistent across multiple Chromosomes from the same organism. It would also be possible to represent the data in a fragmented way such that local biases within a Chromosome could be examined. There is existing data on the difference in methylation state throughout a Chromosome. In Arabidopsis it has been shown that the centromeres and peri-centromeric regions of chromosomes demonstrate very high levels of DNA methylation [57]. It would therefore be intriguing to know whether the Wavelet-based features are localized to specific regions within Chromosomes. To that end future analyses should be carried out which fragment Chromosomes and examine whether there is a bias in the detection of Wavelet-based features which corresponds to existing data on methylation status of the corresponding region of the Chromosome.

While the correlation between Wavelet-based feature distributions is weak in fungi it is striking in higher eukaryotes. The high level of correlation amongst mammals is clear and reinforces the hypothesis that wavelet analysis biased towards di-nucleotide frequencies identifies biologically meaningful genomic regions. Surprising is the finding that in higher eukaryotes similarity is seen between honeybee and higher mammals while fly is highly dissimilar. This finding is of particular interest in light of differences between these two insects and the extent to which they have active CpG methylation mechanisms.

In parallel with the publication of the honeybee genome was the publication of the finding that A. mellifera contains and exploits a full complement of DNA methyl-transferases [60,79] while *D. melanogaster* does not methylate CpG di-nucleotides [80]. The similarity of an insect to mammals in the context of DNA methylation would have been highly contentious prior to the completion of the honeybee genome due to the lack of DNA methylation in *D. melanogaster* [80,81]. However, the publication of the honeybee genome carried with it the recognition that *A. mellifera* contains a full complement of DNA methyl-transferase genes as opposed to other insects [60,82]. The finding that Wavelet-based feature distribution in *A. mellifera* closely resembles that of mammals as opposed to *D. melanogaster* provides further evidence for the use of *A. mellifera* as a model for studying DNA methylation in a social context and genomic imprinting as a model for metazoans [61,83].

Efforts are underway to sequence a number of genomes from other Drosophila species (). With a number of genomes available from Drosophila species it will be possible to perform an analysis similar to that presented in this study on multiple

genomes from a single genus. The availability of a genus specific data set would provide a significant advantage in examining the connection between the Chromosomal features identified through Wavelet analysis and their biological significance. Of particular interest would be an examination of whether the distinction seen in this study between *D. melanogaster* and *A. mellifera* holds for the other Drosophila species. Existing data has shown that amongst insects, particularly Drosophila and mosquito, there is conservation seen of the *DNMT2* locus [84,85]. Surprising is that there is existing evidence that Drosophila exhibit low levels of DNA methylation and that the primary targets for DNA methylation are not CpG di-nucleotides.

It is initially surprising that *Arabidopsis thaliana* was not seen to cluster tightly with the mammals. Arabidopsis has the largest complement of DNA methyl-transferases of any sequenced eukaryote [67]. Evidence to date has implied that the degree of methyl-transferase activity in Arabidopsis is low and has limited phenotype. In particular DNA methyl-transferases in plants and filamentous fungi have been most widely associated with targeting transposons and repetitive DNA elements. Mutation studies on the DNA methylation genes within Arabidopsis have shown that the impact of DNA methylation at CpG di-nucleotides is most tightly associated with recombination events and not heredity. In Arabidopsis, as opposed to metazoans, DNA methylation is maintained by at least an additional class of genes, the Chromomethylases (e.g. *CMT3*) [86]. Arabidopsis has homologs to the DNA methyl-transferases of metazoans which regulate CpG di-nucleotide methylation (e.g. *MET1*). Where Arabidopsis differs is in the maintenance of methylation of CpNpG and CpNpN sites through chromomethylases. Recent work using the loss-of-function mutants *met1* and *cmt3* has shown that mutations in the *MET1* and

CMT3 locus abnormally alter embryogenesis in Arabidopsis. Thus Arabidopsis provides an intriguing, highly complex organism in which to study DNA methylation as opposed to metazoans. Therefore given that the numerical encoding employed here was biased to CpG di-nucleotides the finding that Arabidopsis is more similar to fungi and D. melanogaster as opposed to mammals is supported.

Future work on Wavelet analysis for Arabidopsis should focus on identifying additional encoding schemes which exploit CpGpN and CpNpN tri-nucleotides, such work may be able to identify the genomic features associated with Arabidopsis specific mechanisms of DNA methylation. Given that evidence points to non-CpG targeted methylation in Drosophila species, insight derived in Arabidopsis may have implications in insects. As such, future work should involve comparative genomic analyses between Arabidopsis and the Drosophila species being actively sequenced.

While the data examined here suggest strong correlations between Wavelet-based features and DNA methylation it will require direct genetic evidence in order to associate the identified Wavelet-based features as biologically important. As such loss-of-function mutants will need to be acquired for a set of Wavelet-based features. Through the assessment of the loss-of-function mutants it will be possible to determine whether a given Wavelet-based feature is of discrete biological importance. Further if the loss-of-function mutation were derived through the deletion of the genomic region containing the Wavelet-based feature it would be possible to examine the heritability of that mutation. Therefore future work should address the identification of target Wavelet-based features in model organisms, the subsequent generation of loss-of-function mutations within the Wavelet-based features leading to the phenotypic and genetic analysis of such mutants.

Work to investigate the methylation sensitivity of select mutations could be undertaken by exploiting the existing loss-of-function mutations *met1* and *cmt3* in *Arabidopsis thaliana*. If a Wavelet-based allele was methylation sensitive at CpG di-nucleotides then a phenotypic difference should arise between the double mutant *wavelet/met1* versus a *wavelet* mutation in a wild type background. Further, the phenotype should be less severe or absent in a *wavelet/cmt3* double mutant.

With available mutants for Wavelet-based alleles it would be desirable to examine whether any associated phenotypes are methylation dependant. To that end efforts in model organisms to identify mutations in Wavelet-based alleles would provide an initial basis with which to assess whether there exist phenotypes associated with loss-of-function at Wavelet-based locations. In the next chapter (Chapter IV: **Development of a reverse genetics system for assesing the biological significance of genomic information**) I present work to develop a reverse genetics resource for the identification of deletion alleles in the model plant *Arabidopsis thaliana*. Through the exploitation of a deletion mutant resource for Arabidopsis the possibility exists to establish the heritability of Wavelet-based alleles and corresponding discrete phenotypes.

CHAPTER IV

DEVELOPMENT OF A REVERSE GENETICS SYSTEM FOR ASSESING THE BIOLOGICAL SIGNIFICANCE OF GENOMIC INFORMATION

Introduction

Bioinformatics focuses on the use of computer science to decipher meaning from biological data. While bioinformatics is capable of determining patterns within biological data, it is the creation of novel hypotheses which drives science and which is the goal of bioinformatics. The challenge inherent for bioinformatics is to derive hypotheses *in silico* and subsequently evaluate them *in vivo*. In the case of a novel set of DNA features (E.g. Wavelet-based features as described in Chapter III **Wavelet analysis of Eukaryotic Linkage groups**) there is a need to biologically validate the features as units of DNA that functionally contribute towards the survival and thereby propensity for adaptation of a given organism. Upon showing that a novel class of feature is heritable and functional it would be possible to describe a new class of gene. In order to genetically dissect the function of a locus it is necessary to show that there is a heritable phenotype associated with a particular chromosomal region. The study of genetic mutants is a powerful approach to examine the function by comparison between a mutant and a wild type organism. Further, by examining mutants with aberrations in known, or map-able, locations the possibility exists to associate the mutant phenotype with the genetic location underlying the mutation.

Arabidopsis thaliana serves as an attractive model organism for angiosperms. With its short generational time, small stature, ease of growth, self compatibility, a wealth of genetic data and a fully sequenced genome [87], Arabidopsis has been widely accepted as the premier plant model and serves as a benchmark for identification and

functional determination of genes in other plant species. The study of gene function within Arabidopsis is extremely valuable as insight gained with Arabidopsis impacts plant species in general. To date the dissection of gene function in *Arabidopsis thaliana* relies heavily on the exploitation of the extensive mutant resources that are available, including the development of reverse genetic resources for the phenotypic analysis of allelic variants at select loci. The creation of large mutant populations as reverse genetic resources is not unique to Arabidopsis research. Where Arabidopsis stands apart from other model organisms is the strong culture of public availability of mutant resources which typifies the global Arabidopsis research community. At the time of this study, mutant alleles were available for approximately 70% of the ~31,000 known or predicted genes in Arabidopsis (www.arabidopsis.org) compared to about 10% of known or predicted genes in mouse (www.jax.org).

The impact of large mutant populations in Arabidopsis is best exemplified in the widespread availability and exploitation of sequence-indexed mutations derived from a large T-DNA mutagenized population (the 'SiGNaL' repository) developed for Arabidopsis [88]. In the case of the SiGNaL population, T-DNA insertion within a coding region is generally believed to negatively affect the production of the corresponding protein, thereby generating a loss-of-function allele[89]. The SiGNaL population was sequence-indexed based on sequencing of the T-DNA/genome junction at the point of insertion of the trans-gene. Given the availability of the *Arabidopsis thaliana* genome sequence it was thus possible to identify the insertion sites based on the sequence similarity between the T-DNA/genome junction and the sequence of the *Arabidopsis thaliana* linkage groups. Given that the SiGNaL population contains at least 225,000

insertion events it was originally estimated that the probability of an insertion within any given gene was approximately 96%. In practice it is widely accepted that at least 1 insertional mutant allele is available for approximately 70% of all known or predicted genes in Arabidopsis amongst all reverse genetic resources publicly available (~380,000 lines). Thus, no corresponding loss-of-function allele is available for an estimated 30% of Arabidopsis genes.

The widespread exploitation of the SiGNaL population can be largely attributed to the transparency with which the population was made publicly available. All sequenced T-DNA/genome junction sites were deposited in GenBank in conjunction with the seed stocks for the corresponding lines being made publicly available through international stock centres; the Arabidopsis Biological Resource Centre (ABRC; Columbus OH) in North America and the Nottingham Arabidopsis Stock Centre (NASC) in Europe. In addition to the public release of the sequenced T-DNA/genome junction data, the SALK institute made the SiGNaL population and corresponding seed stocks publicly available and searchable via a web portal [90] The SiGNaL population represents a large coverage of the Arabidopsis thaliana genome, nevertheless, there remains an estimated 30% of the genome for which no known T-DNA insertional mutant is available, or for which the T-DNA insertional does not result in high gene penetrance and a true loss-of-function mutation.

The SCF class of E3 ubiquitin ligases is comprised of a family of multi-subunit holoenzymes formed by the quartet of a canonical *SKP*, a *Cullin*, *Rbx1* and an Fbox subunit protein. E3 ubiquitin ligases serve to conjugate one or more ubiquitin adducts to a specific target protein and thereby signal that protein for degradation via the 26S

proteosome. Selective protein ubquitination and degradation is mediated largely through specific protein-protein interactions between the Fbox member of the SCF complex and the specific target protein. In contrast to yeast where a single Skp1-like gene is present, the Arabidopsis contains 21 *ASK* (***A***rabidopsis ***SK***p1-like) genes and a complement of over 700 Fbox genes. Recently published data on rates of gene duplication in Arabidopsis and Rice have argued that the *ASK* gene family represents a gene duplication rate 10 times the average rate within the genome [91]. As core components of the SCF family of E3 ubiquitin ligases *ASK* genes serve a crucial role in targeting proteins for degradation. With a complement of > 700 Fbox genes and 21 ASK genes the combinatorial complexity of the potential SCF complexes that could form across developmental time and space, suggests that complexity of posttranslational protein targeting mechanisms in plants far exceeds that found in yeast and *Homo sapiens*.

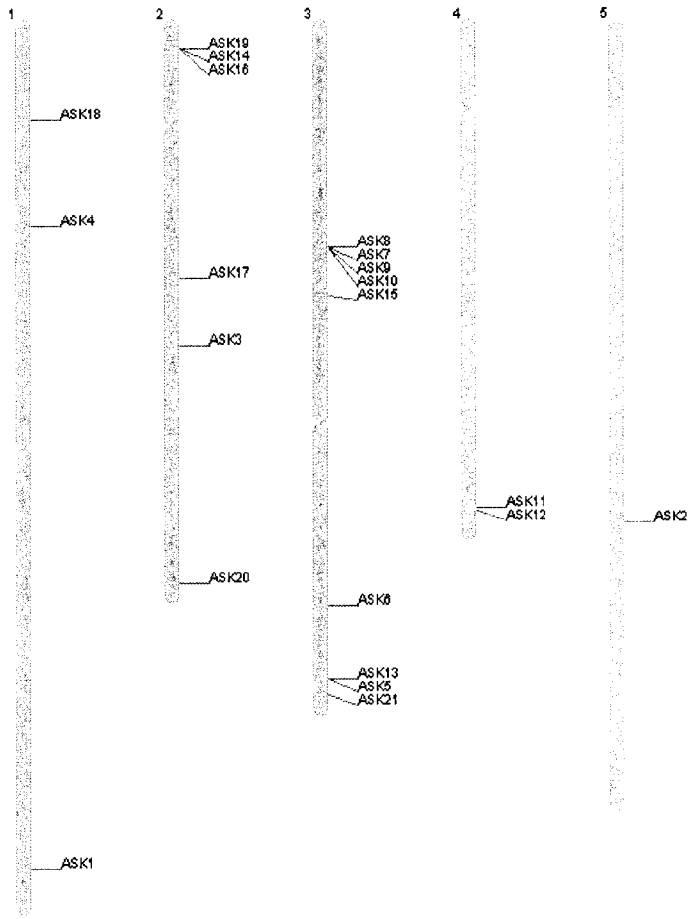

Figure 30 Chromosomal location and distribution of ASK gene family in the genome of *Arabidopsis thaliana*.

Based on yeast 2-hybrid studies it has been shown that within the ASK family *ASK1*, *ASK2* and *ASK11* demonstrate a general and similar protein-protein interaction patterns with Fbox proteins, whereas other members of the ASK family present some level of specificity in their interaction with Fbox genes [92]. Given that there is a significant level of sequence similarity amongst members of the ASK gene family it has been proposed that there is likely some level of redundancy amongst clades. The two best studied ASK genes are *ASK1* and *ASK2* which share75% sequence similarity and where a significant level of redundancy between *ASK1* and *ASK2* has been observed.

Investigation into the Fbox binding partners of *ASK1* and *ASK2* revealed similar patterns [92] while investigation into the phenotypes of loss of function mutants revealed that *ASK1* and *ASK2* have overlapping but distinct roles in Arabidopsis [93].

Initial investigation into the function of *ASK1* demonstrated a role for *ASK1* in male meiosis due to observed unequal chromosome separation in anaphase I in the mutant *ask1* [94]. Further, expression of *ASK1* and *ASK2* were observed to be crucial for wild type embryonic development as demonstrated by the loss-of-function mutants *ask1* and *ask2* [95]. The double mutant *ask1* (*Ds* insertion into coding region) and *ask2* (T-DNA insertion in the coding region) are lethal [96]. Complementation studies using over-expression constructs for *ASK1* and *ASK2* in an *ask1* mutant background revealed that while both genes can restore the wild type flower development *ASK2* is only partially able to compensate for the *ask1* loss-of-function. *ask1* mutants transformed with *ASK1* over-expression constructs have wild type flower development whereas *ASK2* over-expression leads to short siliques and much reduced pollen fertility [97].The partial rescue by *ASK2* and reduced expression of *ASK2* relative to *ASK1* in wild type plants during male meiosis have lead to *ASK1* and *ASK2* being associated with overlapping but distinct functions.

Given the evidence for complementation between *ASK1* and *ASK2* it is reasonable to hypothesize that similar patterns of genetic redundancy exists amongst other clades within the ASK family [92]. In order to dissect the interplay amongst members of a given clade it will be necessary, as was the case for *ASK1* and *ASK2*, to study the phenotype of plants carrying multiple loss-of-function alleles. In the case of *ask1* and *ask2* simply crossing two mutants (one for each allele) will yield a reasonable frequency of the double

mutant amongst segregating progeny since the genes reside on independent linkage groups (chromosomes 1 and 5 respectively see Figure 30). In the case of certain clades (e.g. *ASK7-10*) combining multiple mutant alleles via genetic recombination would be complicated by the tight physical clustering of the 4-gene *ASK7-10* clade within and 8kb region on Chromosome 3. Thus, alternative methods for identifying and/or combining loss-of-function alleles at these 4 loci will be required in order to dissect the individual and collective function of these genes. Of particular interest in this study is the identification and study of deletion mutants which, depending on the size of the deletion involved, may span and oblate the function of more than one member of the *ASK7-10* clade. From a genetic standpoint, the only absolutely penetrant loss of function allele for a given gene is a deletion. Thus, phenotype of deletion mutants are often more definitive than other classes of mutant allels, including insertion based mutations. There exists significant need to expand the existing reverse genetic resources for Arabidopsis to include and exploit deletion mutants, yet a large repertoire of such mutants is not currently publicly available.

Seeds exposed to Fast neutron irradiation accumulate a high frequency of double strand breakage and subsequent deletions in Arabidopsis [98-101]. By co-harvesting seeds with plant material for the preparation of DNA it is possible to generate low-complexity pools of Fast Neutron irradiated mutants for the analysis of plants carrying deletion alleles in select genes of interest. By performing nested PCR on mutant DNA pools it has been shown previously that it is possible to identify large deletions (up to 12 kb) from Fast neutron irradiated Arabidopsis mutants.

To date there has been at least one attempt to develop a deletion mutant resource for Arabidopsis [102,103], but this resource is not publicly available. As shown here we have developed an alternate deletion mutant resource for *Arabidopsis thaliana* which should provide a basis for the study of defined deletion alleles as a community resource. In this study I describe the development of the resource, the validation of the DNA as suitable for long PCR-based genotypic analysis, an investigation into the use of whole genome based amplification of the mutant DNA pools, together with an explicit demonstration as to the limit of mutant allele detection.

M2 mutant DNA and seed pools

Seed for M2 plants were obtained from Lehle Seeds (Round Rock Texas, USA stock # M2F-01A-04). The seeds were from M1 pools of an Arabidopsis Col-3 population (53,856 individuals) irradiated with 60 Gys of fast neutron irradiation (Col-3 differs from the Col-0 genetic background by a single polymorphism in the *GLABROUS* locus, affecting trichome formation). M1 plants were pooled with an average size of 1122 plants / pool. For each M1 parental pool, seed was sown into 2 bedding flats (denoted α and β). Each bedding flat was dividing into quadrants (denoted a,b,c,d). Thus for each M1 parental pool seed was sown into 8 quadrants resulting M2 pools representing approximately 140 distinct parental plants. Arabidopsis seed pools were sown in Premier Biomix BL with sub-irrigation watering, and plants were grown in a greenhouse facility kindly provided by Agriculture and Agri-Food Canada Greenhouse and Processing Crops Research Centre located in Harrow, Ontario.

For each M2 quadrant and pool, plant material was harvested when the initial bolt had achieved an average height of 10cm. Plant material from the base of the bolt through

the apical meristem was collected, the tissues flash frozen in liquid nitrogen in preparation for storage at -80C until DNA extraction could be performed. Plants from which bolt and meristem had been harvested were returned to the greenhouse and allowed to bolt a second time and to flower. Plants were allowed to self pollinate and set seed. Seed stocks for each M2 pool were harvested and weighed prior to storage at the University of Windsor.

Modified DNeasy DNA purification procedure

DNA preparations from plant material harvested from each of the M2 pools were generated using a modified version of Qiagen's DNeasy protocol (mDNEasy). Plant tissue was ground to a fine green powder in a mortar and pestle using liquid nitrogen. Due to the number of samples the manual disruption was performed in batches with the tissue being returned to -80C prior to continuing with the DNA extraction. Plant material was transferred to a 13 ml round bottom poly propylene tube and 4mls of a Grinding buffer (0.8M Sucrose, 50mM Tris pH 7.5, and 10mM EDTA) was added. In the case of larger samples Grinding Buffer was added in 1 ml aliquots until the tissue was submerged. Tissue was macerated using a Tissumizer (IKA, Wilmington North Carolina, TP18) 3 times at 45rpm for 12 seconds each. In order to isolate the aqueous fraction from the tissue debris, samples were passed through a layer of Miracloth^{TM} into a 50ml conical bottom polypropylene tube. The organellar fraction was isolated by transferring the flow-through to a 15ml conical bottom tube and centrifuged at 3,000g for 15 minutes at 4C. The supernatant was discarded by decanting and the organellar pellet was re-suspended in 400µl of buffer AP1 (Qiagen) and samples were transferred to 1.5ml micro-centrifuge tubes. RNA present in the samples was actively degraded through the addition of 13.2U

of Qiagen's RNaseA to each sample and subsequent incubation in a hot water bath for 10 minutes at 65 C with the samples being mixed by inversion every 2 minutes. 130 µls of the lysis Buffer AP2 (Qiagen) was added and the samples were incubated on ice for 5 minutes. The lysed solution was centrifuged for 5 minutes at 20,000g and applied to a QIAshredder mini spin column placed in a 2ml collection tube and centrifuge for 2 minutes at 20,000g. The flow through fraction was added to a fresh 2ml micro centrifuge tube and 1.5 volumes of Buffer AP3/E (Qiagen) was added and mixed by pipetting. Samples were passed through DNeasy spin columns in 650µl aliquots. For each aliquot the column was spun @ 6000g for 1 min and the flow through was discarded. The DNeasy column was placed in a new 2ml collection tube and 500µls of wash Buffer AW (Qiagen) was added. The Wash buffer was passed through the column by centrifugation for 1 min @ > 6000g. A second wash was performed with 500µls of Buffer AW (Qiagen) added to the DNeasy column and centrifugation at 20000g for 2 minutes to dry the membrane. The DNeasy column was added to a 1.5 ml micro-centrifuge tube and eluted twice; each time 100µls of the AE buffer (Qiagen) was added, samples were incubated for 5 minutes at room temperature and then centrifuged for 1 minute at 6000g to elute the purified genomic DNA. The mDNeasy DNA pools along with controls (PCR grade H_2O) along with wild type DNA samples (Col-0, Col-3, C24, Le) were arrayed into four 96-well skirted microtitre plates (ThermoFast, AbGene Rochester New York, USA) and labelled 'Master plates'.

Use of BACs as discrete templates for PCR

To provide a discrete template source for the optimization of PCR Bacterial artificial chromosomes (BACs) were obtained from The Arabidopsis Information

Resource (TAIR). DNA from the Bacterial artificial chromosomes was purified from overnight E. coli cultures using Promega's Wizard Plus SV miniprep kit (Madison, Wisconsin, #A1330). The BACs F19P19 and MSD21 span the genomic regions at the *CRY2* and *ASK7* loci respectively.

Primer design strategy

A nested primer design strategy was employed to determine primers suitable for Phusion™ PCR of long templates in complex mixtures. Perl scripts were employed to design a set of nested primers where the flanking converging external primers are situated approximately 10kb apart centered on a given locus (e.g. *CRY2*). A nested set of primers (internal) was designed to be within 2kb of each of the external primers. A third set of 'control' primers were derived from the reverse compliment of the internal primers. The control primers provided the ability to validate the primer landing sites. By using the each control primer in conjunction with the external primers it was possible to selectively amplify a small (< 2kb) product which provides a diagnostic for both the internal and external primer sites, *CRY2* and *ASK7* primer design are demonstrated in Figure 31 Figure 32 respectively.

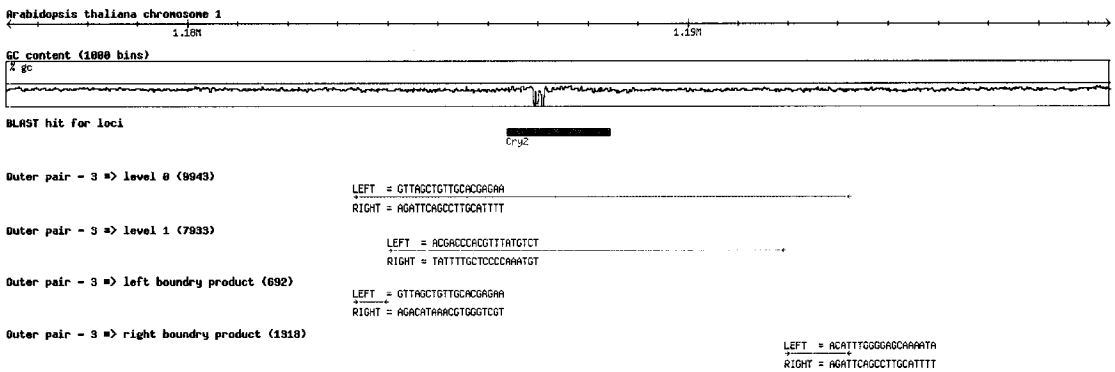

Figure 31 Primer design strategy for *CRY2*. The *CRY2* locus on Arabidopsis chromosome 1 carries a depressed region of GC content as indicated in the upper panel. Primers designed for this locus are

indicated with there putative amplicons in green. The primer design included a nested design strategy. The reverse compliments of the internal pair of primers are used to amplify left and right boundary products as shown.

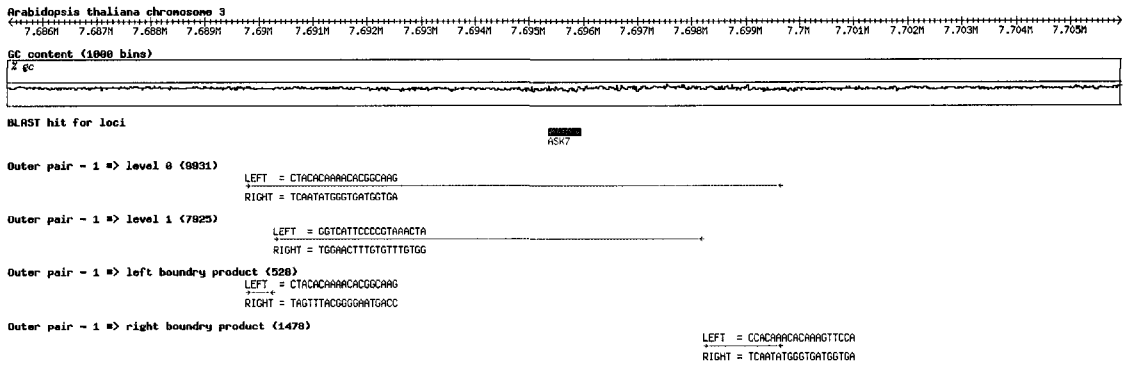

Figure 32 Primer design strategy for *ASK7*. Predicted primers for the *ASK7* locus are indicated with there putative amplicons in green. The primer design included a nested design strategy. The reverse compliments of the internal pair of primers are used to amplify left and right boundary products as shown.

All primers designed for *UFO, ASK7, ZTL, CRY2* and *ACT8* were designed to have a Tm within 1.5C of each other as can be seen in Table 5.

Table 5 Primers used for Taq- and Phusion™-based PCR. Tm was calculated using Finnzymes Tm calculator (http://www.finnzymes.com/tm_determination.html).

Name	Sequence	Tm
UFO-CDS-405-FOR	TTGTGTAACCCTCTTGTCG	60.5
UFO-CDS-405-REV	AATAAGCCTCCCTTTGCTC	61.4
UFO-506-FOR	GGCTTTGTAGCTTGGAATC	60.5
UFO-506-REV	AAAACCCTGCAAGACCTC	61.0
UFO-600-FOR	TCTCCTTACGCTGTGAAAA	60.3
UFO-600-REV	AAAACCCTGCAAGACCTC	61.0
UFO-703-FOR	GAGTCAGTTGCCACCAATA	60.4
UFO-703-REV	AAAACCCTGCAAGACCTC	61.0
UFO-906-FOR	AGAGGAGGAACAAACGATG	61.0
UFO-906-REV	AAAACCCTGCAAGACCTC	61.0
ACT8-1001-FOR	TCTTAACCCAAAAGCCAAC	60.8
ACT8-1001-REV	AAGCATTTTCTGTGGACAA	60.0
ZTL-1164-FOR	ATCGCCGTCTTCATAATCT	60.3
ZTL-1164-REV	AGCAAACGGTCCTCTACAT	60.4
ZTL-1451-FOR	TTACAGGGTATCGTGCTGA	61.0
ZTL-1451-REV	CAAGGGTGGTCAGTTCTCT	61.0

ZTL-1956-FOR	TTTCATGGAGTGGGACAG	61.5
ZTL-1956-REV	AAACACATCGTTCAGCAAA	61.0
ZTL-2455-FOR	GTTGTTACTGATGCCGTTG	60.9
ZTL-2455-REV	TGACCACCTAGCACTATCG	60.6
CRY2-9943-FOR	GTTAGCTGTTGCACGAGAA	60.9
CRY2-9943-REV	AGATTCAGCCTTGCATTTT	60.5
CRY2-7933-FOR	ACGACCCACGTTTATGTCT	61.2
CRY2-7933-REV	TATTTTGCTCCCCAAATGT	61.0
CRY2-9943-FOR-TE	AGACATAAACGTGGGTCGT	61.2
CRY2-9943-REV-TE	ACATTTGGGGAGCAAAATA	61.0
ASK7-9931-FOR	CTACACAAAACACGGCAAG	61.1
ASK7-9931-REV	TCAATATGGGTGATGGTGA	61.5
ASK7-7925-FOR	GGTCATTCCCCGTAAACTA	60.5
ASK7-7925-REV	TGGAACTTTGTGTTTGTGG	61.3
ASK7-9931-FOR-TEST	TAGTTTACGGGGAATGACC	60.5
ASK7-7925-TEST	CCACAAACACAAAGTTCCA	61.3
	min	60.0
	max	61.5
	mean	60.9

Fast Neutron derived deletion allele suitable for reconstruction experiments

Genomic DNA prepared from *Arabidopsis thaliana* plants carrying a known 2/3 deletion of the *CRY2* locus was used in reconstruction experiments designed to validate PCR conditions used for the detection of deletion alleles at this locus[104]. *cry2* mutant plants present aberrant growth rates relative to Col-0 when grown in the absence of blue light. The seed stock for *cry2* mutant was obtained from TAIR (Germplasm identifier: CS3732). Confirmation of the mutant phenotype was performed by growing plants from CS3732 and wild type (Col-0) under constant white light in a growth cabinet lacking a blue light source.

Validation of the *CRY2* primer pairs for the amplification of the wild type allele was carried out by selective amplification of the left and right converging boundary products as well as the 10kb target (Figure 31). Amplifications were performed using a discreet BAC DNA template. In order to reduce complexity in the optimization PCR

reactions a BAC (F19P19) that spanned the region of Arabidopsis Chromosome 1 containing the *CRY2* locus was used as amplification template (10ng).

In order to investigate the use of Phusion™ DNA polymerase for screening the mutant DNA pools, a series of reconstruction PCR experiments were undertaken involving mDNeasy genomic DNA prepared from wild type Col-0 and mutant *cry2* plants. A series of dilutions (1:5, 1:25, 1:125, and 1:625) of *cry2* DNA into Col-0 genomic DNA were prepared and used as template in PCR amplification experiments performed in combination with *CRY2* locus specific primers.

Taq DNA polymerase based PCR

The suitability of DNA templates for use in single target (uniplex) PCR was assessed using purified Taq DNA polymerase (W. Crosby). PCR was conducted in a standard 20µl reaction containing: 1x buffer from Promega (Mg^{2+} free), 1.5mM $MgCl_2$, 0.5mM dNTPs (New England Biolabs), 10-50ng template genomic DNA and 20pmol each primer. Taq based PCRs were cycled in a Mycycler™ (BioRad) thermocycler with an initial cycle of 30 seconds at 96°C, 30 seconds at 53°C followed by 1 minute at 72°C. Samples were subjected to 35 rounds of amplification with 30 seconds at an annealing temperature of 94°C followed by 30 seconds at 53°C and 30 seconds / kbps at the extension temperature of 72°C. A final amplification cycle was performed with an extension time of 30 minutes at 72°C.

In order to assess the suitability of mDNeasy purified genomic DNA for uniplex PCR a series of varied size targets were amplified from single copy genes. The full length gene models for selected loci were obtained from TAIR for *Unusual Floral Organs* (*UFO*), *Actin-8* (*ACT8*) and *Zeitlupe* (*ZTL*). *UFO* is a 1.3 kbps introns-less gene located

on Chromosome 1 [93,105,106] that encodes an Fbox protein and was previously shown to be involved in the establishment of both floral organ identity as well as the maintenance of cadastral boundaries during floral morphogenesis [107]. *ACT8* encodes one of the actin family of proteins and is also situated on Chromosome 1, but has a more complex gene model than *UFO* with 4 introns spanning 2.3kbps [108]. *ZTL* is a 3.1kbps gene consisting of two exons residing on Chromosome 5 and whose gene product is also an Fbox protein and involved in the regulation of circadian clock periodicity [92,109-112]. Primers were designed to amplify various length genomic fragments from *UFO*, *ACT8*, and *ZTL* loci (). Validation of mDNeasy DNA for Taq DNA polymerase-based uniplex PCR was performed by amplifying a series of 405, 506, 600, 703 and 906 bps amplicons spanning the *UFO* locus. As a positive control, amplification of a 405bp amplicon from a cloned cDNA for *UFO* (W.Crosby) was used. Single copy genomic targets up to 2.5 kb were validated using 50ng of purified genomic DNA for targets from *ACT8* (1001bps) and *ZTL* (1164, 1451, 1956, and 2455 bps).

Phusion™ based PCR

Long-product uniplex PCR (2.5-10kb) was performed using Phusion™ DNA polymerase (Finnzymes; NEB #F-553). Phusion™ is a recombinant enzyme based on a DNA polymerase from a *Pyrococcus* species with a 5' to 3' polymerase activity and a 3' to 5' exo-nuclease proofreading activity. Through the addition of double stranded DNA binding domain, Phusion™ is able to achieve very high processivity and is thereby suited for the amplification of very long targets. Likely Phusion™ contains an Sso-7d double stranded binding domain fused to the C terminus of a cloned PFU DNA polymerase as described by Wang *et. al* [113]. In order to optimize PCR conditions for Phusion™ a custom

5x Mg^{2+} free buffer (GC version) was obtained directly from the manufacturer (Finnzymes) with the assistance of New England Biolabs (NEB). Standard reactions were conducted in a 20µl volume containing: 1x Phusion™ Mg^{2+} free buffer (GC version), 3mM $MgCl_2$, 0.2mM dNTPs (NEB), 0.4U Phusion™ DNA polymerase, 10-50 ng template DNA and 20pmol each primer. The amount of $MgCl_2$ found to be optimal was at a concentration of 2.8mM as opposed to the manufacturers recommended 0.5-1.0mM above the dNTP concentration. Thermo-cycling was performed in a BioRad Mycycler™ with an initial cycle of 100°C for 3 minutes followed by 35 cycles of amplification with denaturing occurring for 10 seconds at 100°C followed by a 15 second annealing phase at 61°C (*CRY2* primers) and 67°C (*ASK7* primers) with a 30 second / Kbps extension at 73°C. A final extension cycle was done at 73°C for 10 minutes.

Whole genome amplification and DNA quantification using GenomiPhi™

Examination of the suitability of Phi^{29} amplified genomic DNA for Phusion™ based PCR was performed using Genomiphi™ from GE Health Care (Formerly Amersham Biosciences). While Phi^{29} amplified genomic DNA has been used previously for uniplex PCR [114,115] there was no data available at the time of this study which had evaluated the suitability of Phi^{29} amplified DNA as template for Phusion™-based long PCR. Accordingly five-fold serial dilutions of *cry2* mutant DNA into Col-0 DNA were made and subsequently amplified by Phi^{29} using the commercially available Genomiphi™ reagent kit. The resulting Phi^{29} product was used as template for amplification of the *cry2* deletion allele using primers diagnostic for the mutant deletion (*CRY2*-9943-FOR and *CRY2*-7933-REV).

DNA quantification was performed using Quant-iT PicoGreen™ reagent (Invitrogen) and interpretation of sample fluorescence relative to a standard curve for lambda phage DNA. Quant-iT PicoGreen™ was added to Tris buffers as per the manufactures instructions. Samples were loaded into 384 well microtitre plates (UV Star, Greiner Bio One, Monroe North Carolina) and fluorescence was measured using a Victor microtitre plate reader (Perkin Elmer) equipped with a D480/30x Excitation filter and an F520 emission filter.

Pin tool based assembly of Phusion™-based PCR

Pin tools were initially developed for replicating libraries (yeast/E. coli) but have also become useful for the reproducible transfer of very small liquid volumes [116]. Disposable pin tools made of polypropylene are available in formats such that pins are arrayed complementary to 96- and 384-well microtitre plates. Polypropylene tools in 96-well format were used in this study (V&P Scientific, San Diego, California). Efficacy of the pin tools to transfer DNA was assessed in two ways: transfer of a Lambda phage DNA standard and in the assembly of Phusion™-based PCRs. Lambda phage DNA standard was assembled as per manufacturer's instructions (Invitrogen). Pin tools were used to transfer ~ 130nl of the Lambda DNA standard into a Tris buffer and the DNA concentration was read as described for the use of Quant-iT PicoGreen™ (Invitrogen). Phusion™-based PCRs were assembled using 96-pin pin tools to transfer template DNA from M2 mutant DNA pools and a 7925 bps amplicons was targeted using the primers *ASK7*-7925-FOR and *ASK7*-7925-REV.

Results

Amplification of discreet targets for *UFO*, *ACT8*, and *ZTL* are shown in Figure 33 and demonstrate the suitability of mDNeasy purified genomic DNA for amplification of amplicons less than 1 kb in size.

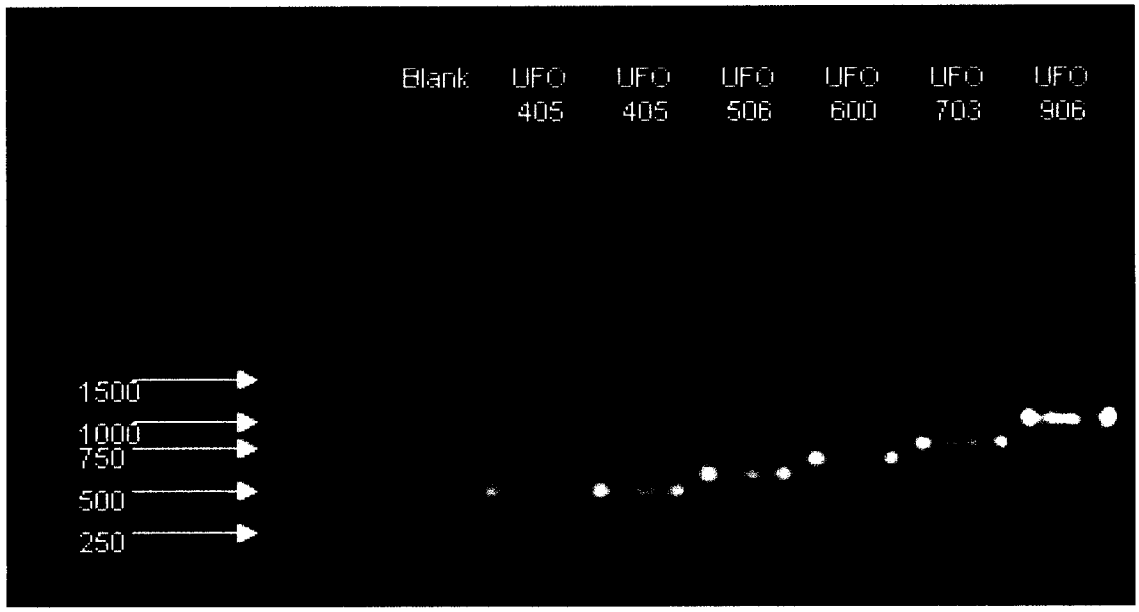

Figure 33 Suitability of mDNeasy purified DNA as template for Taq based PCR of the *UFO* locus. Lanes from left to right: Molecular size standard, negative control, positive control amplicon from cloned cDNA; all other amplicons are produced from genomic DNA templates. Amplicon sizes are indicated at the top of the lanes.

Figure 34 demonstrates the amplification of discreet amplicons of 1kbps – 2.5 kbps from mDNeasy purified genomic DNA.

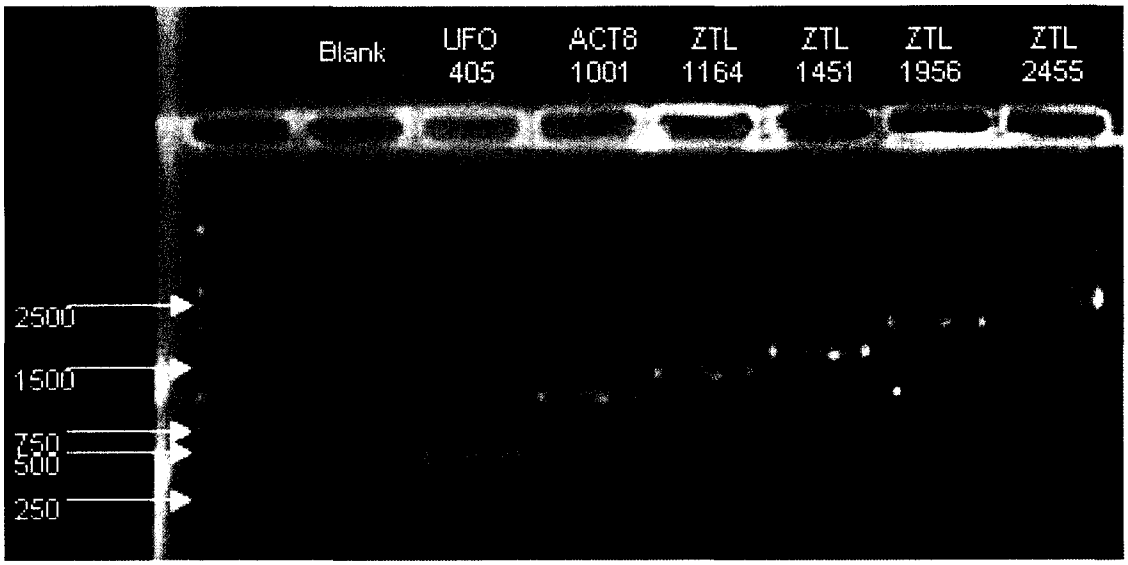

Figure 34 Validation of genomic DNA for targets up to 2.5kb from the loci *UFO, ACT8,* and *ZTL*. Lanes from left to right: Molecular size standard, negative control (lacking template DNA), positive control amplicon from cloned cDNA; all other amplicons are produced from genomic DNA templates. Amplicon sizes are indicated atop each lane.

In order to investigate the feasibility of PCR based screening for deletion mutants it was necessary to derive PCR primers suitable for the amplification of the *CRY2* locus by Phusion™ DNA polymerase. The primer design strategy employed in this study involved the design of 2 pairs of nested primers (Figure 31). By using primers which were the reverse compliment of the inner pair of primers it was possible to validate the primer landing sites and thereby optimize the choice of primers for amplification of target loci. The use of control BAC templates as shown in Figure 35 allowed for the optimization of the amplification reaction including functional validation of the primers used.

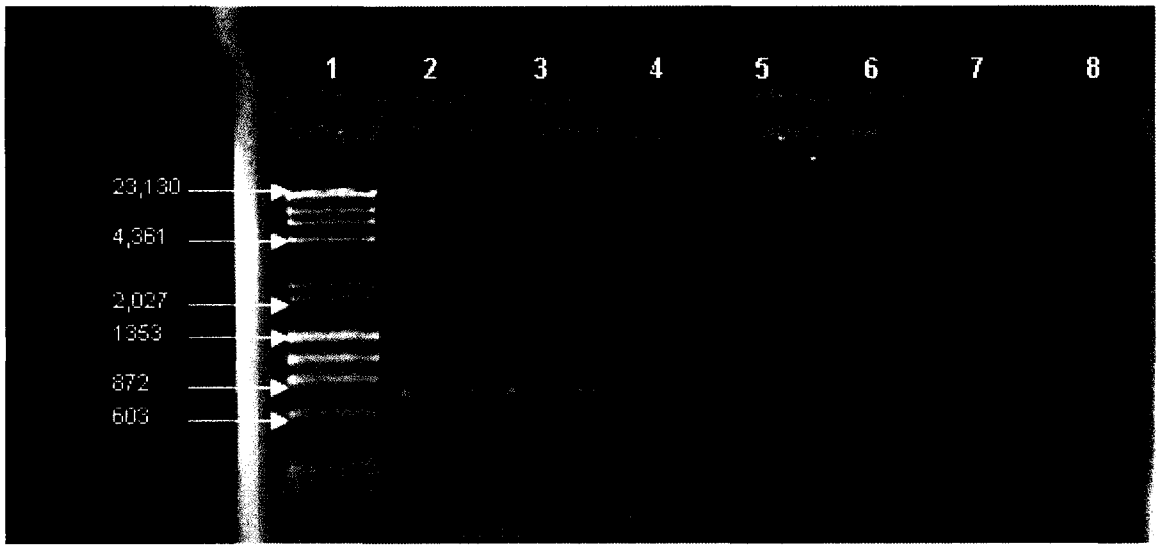

Figure 35 Validation of *CRY2* primers. Lanes 2 and 3 are replicates of the amplification of the left boundary product (692 bps using primers *CRY2*-9943-FOR and *CRY2*-9943-FOR-TE). Lanes 4 and 5 represent failed amplification of the 9943bps target (primers *CRY2*-9943-FOR and *CRY2*-9943-REV). Lanes 6 and 7 are amplifications of the right boundary product of 1318 bps (primers *CRY2*-9943-REV-TE and *CRY2*-9943-REV). Lane 8 is a negative control. All reactions are of 10ng of DNA from the BAC F19P19 which on which the *CRY2* locus is situated.

In diagnosing the amplification pattern seen in Figure 35 problems with at least one of the external primer sites were evident. Following reaction optimization, the primer pair *CRY2*-9943-FOR and *CRY2*-7933-REV was used to amplify a product from both the wild type (Col-0) and the mutant *cry2* it was seen in Figure 36 that these primers were diagnostic for the large deletion within the *cry2* mutant.

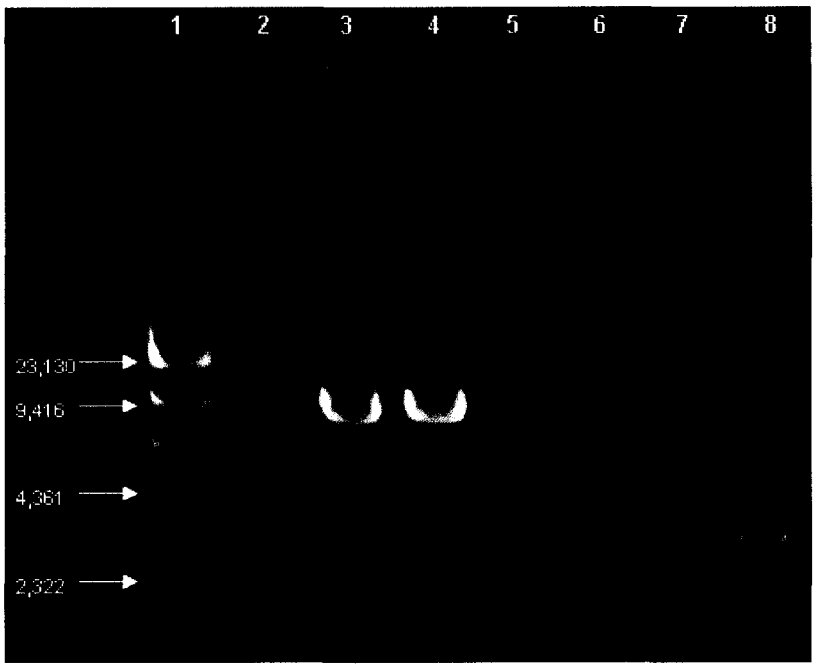

Figure 36 Validation of primers for amplification of a 8625 bp amplicon spanning the *CRY2* locus (*CRY2*-9943-FOR and *CRY2*-7933-REV). Lane 1; molecular size standard, Lane 2; negative control (lacking template DNA), Lanes 3 and 4; biological replicates from the wild type (Col-0), Lanes 5 and 6; technical replicates of *cry2* mutant template DNA, Lanes 7 and 8; technical replicates of a second biological replicate of the *cry2* mutant template DNA. All samples are based on 50ng of mDNeasy purified genomic DNA as template.

With a primer pair validated and diagnostic for the *cry2* deletion allele it was possible to investigate whether Phusion™ DNA polymerase based PCR was able to detect the mutant allele in a genomic DNA sample of similar complexity to the M2 mutant DNA pools developed in this study. As described in the Materials and Methods the DNA pools prepared from the Fast Neutron irradiated population represented a complexity of about 140 distinct M1 individuals per pool. As shown in Figure 37, PCR reactions in combination with Phusion™ DNA polymerase was suitable for the detection of the predicted ~3Kbps mutant amplicon where *cry2* mutant DNA was diluted to a 1 in 625 fraction of wild type DNA. The results shown here demonstrate the suitability of

Phusion™ DNA polymerase for the detection of deletion mutant alleles in a complexity exceeding that of the M2 DNA pools.

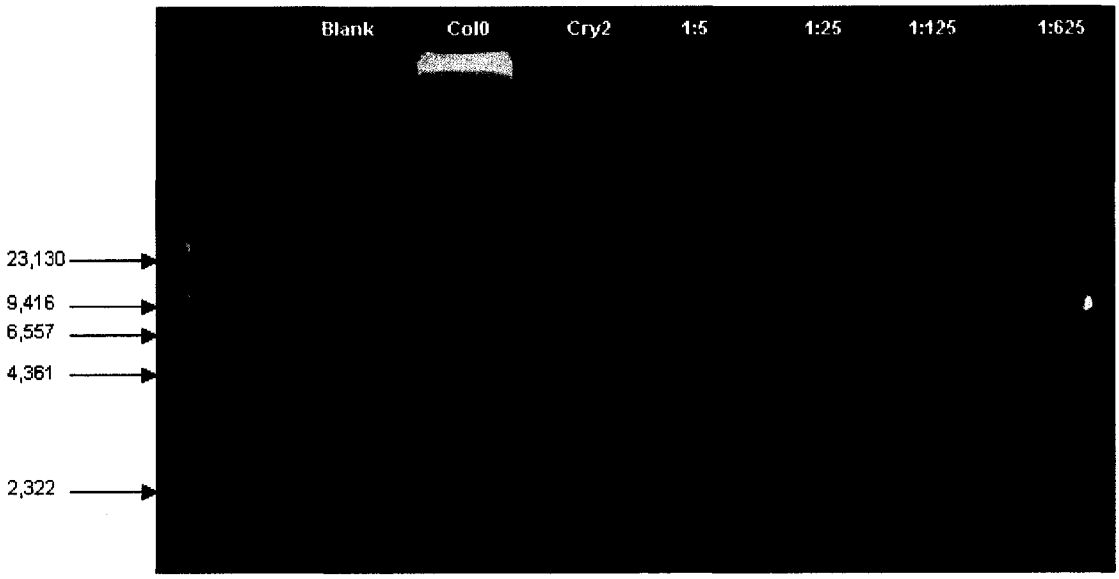

Figure 37 Reconstruction of M2 DNA pool complexity using the *cry2* mutant. 5-fold serial dilutions of *cry2* into Col-0 were made and 50ng of the complex sample was used as template for Phusion™ based PCR. Amplification in the wild type genome is of the 8625bps target (primers *CRY2*-9943-FOR and *CRY2*-7933-REV).

DNA preparation of the M2 plant pools created of the Fast neutron irradiated population using the mDNeasy protocol resulted in a limited quantity of DNA template for application in PCR screening reactions. The average yield from the mDNeasy purification protocol yielded sufficient template for approximately 200 screens from each mutant pool (10μg / pool). In an attempt to immortalize the DNA from the mutant pools (without the need of growing out a subsequent generation of plants) a Phi^{29} -based commercial genomic DNA amplification system was used to non-selectively amplify the mutant DNA pools, and the product was subsequently assessed for use in PCR screening reactions.

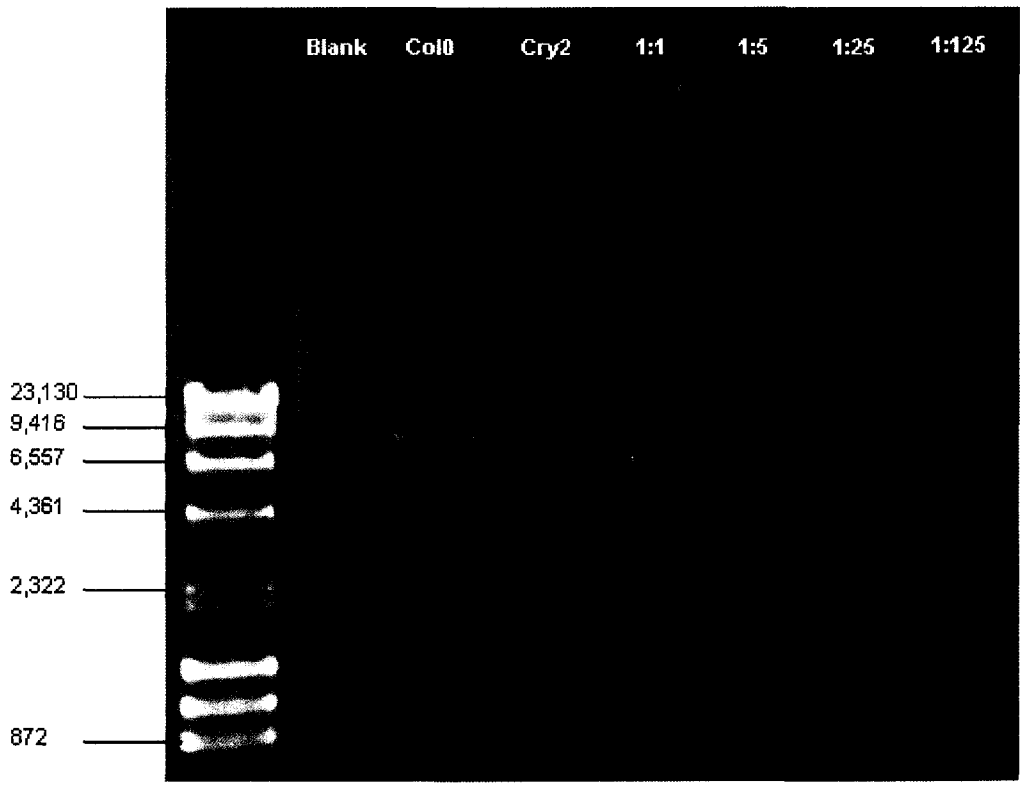

Figure 38 Use of Phi^{29} amplified mDNeasy DNA. Lane 1 is a molecular size standard. Lane 2 is a negative control for both the Phi^{29} and Phusion^{TM}-based amplifications. Lane 3 is the wild type allele, Lane 4 shows the mutant allele from *cry2*. Lanes 5 through 8 show the Phi^{29} amplification and subsequent Phusion^{TM}-based PCR of 5-fold dilutions of cry2 into Col-0. Note the low molecular

weight products (~800bps) and the reduced efficacy relative to non Phi^{29} amplified template. Amplification in the wild type genome is of the 8625bps target (primers *CRY2*-9943-FOR and *CRY2*-7933-REV).

As shown in Figure 38 the Phi^{29} amplified genomic DNA product was suitable for the detection of the mutant allele, although there were significant problems which arose for the detection of the amplicon from Phi^{29}-amplified templates. More specifically, Phusion^{TM} based PCRs in combination with Phi^{29} amplified templates were subject to a higher frequency of low molecular weight non-specific amplification products (Figure 38 lanes 3-8) and the efficacy the reaction was generally reduced as evidenced by a reduced yield of the desired amplicon product. PCR failures were common place with Phi^{29}-amplified templates such that reactions involving Phi^{29}-amplified templates would not be recommended for future development of this, or similar, mutant resources without further validation.

Given the lack of suitability of the Phi^{29} amplified template in long PCR an initial mutant screen was performed using Phusion^{TM} DNA polymerase to directly amplify targets from the mutant pools using primers specific for the *ASK7* locus. As seen in Primer design strategy and Figure 32, primers were designed centered on the *ASK7* locus in a manner analogous to those designed for *CRY2* with two sets of nested primers along with the left and right boundary product primers.

Validation of the *ASK7* primers was performed in order to select the optimal pair with which to screen the mutant DNA pools. In Figure 39 the results of amplification with the battery of the *ASK7* primers is presented. Figure 39 demonstrates the benefit of the primer design used in this study as it is clear that the primers of choice are those from

lane 6 (*ASK7*-7925-For and *ASK7*-7925-REV) since they produce a prominent 8Kbps amplicon indicative of the wild type allele.

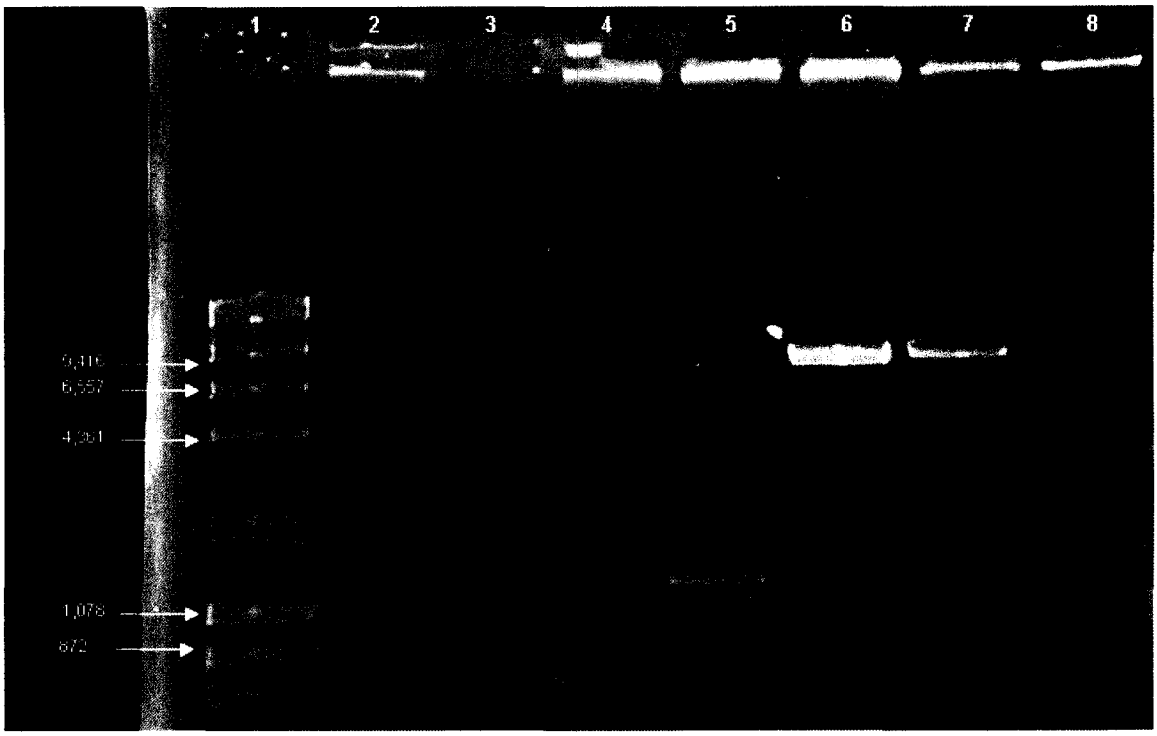

Figure 39 Validation of Primer sets for amplification of the *ASK7* locus from Col-0 template DNA. Lane 1; molecular size standard, Lane 2; negative control (no template DNA), Lane 3; boundary product produced with primers *ASK7*-9931-FOR and *ASK7*-9931-FOR-TEST (528 bps), Lane 4; failure of amplification using *ASK7*-9931-FOR and *ASK7*-9931-REV primers (9931 bps), Lane 5; is successful amplification of the right boundary product using primers *ASK7*-7925-TEST and *ASK7*-9391-REV (1478 bps), Lane 6; amplification of *ASK7* locus using the primers *ASK7*-7925-For and *ASK7*-7925-REV (7925 bps). Primers used in Lane 7 (*ASK7*-9391-FOR and *ASK7*-7925-REV) were also capable of producing a large amplicon (8453 bps). Lane 8 employed primers *ASK7*-7925-For and *ASK7*-9391-REV (9403 bps).

Using a set of primers diagnostic for the *ASK7* locus an initial screen was performed on the mutant DNA pools. For the initial *ask7* screen ¼ of the population was screened (Master plate 1). Figure 40 shows the results of the initial screen for *ask7*. Samples were loaded in an interleaved fashion such that successive rows of the microtitre

plate were interleaved with one another. Well F2 from Master plate 1 showed a possible positive result for a deletion at the *ASK7* locus. As can be seen in Figure 40 there appeared to be 2 lower molecular weight bands which may have been indicative of a mutant in the corresponding pool carrying a deletion within the primers (*ASK7*-7925-FOR and *ASK7*-7925-REV).

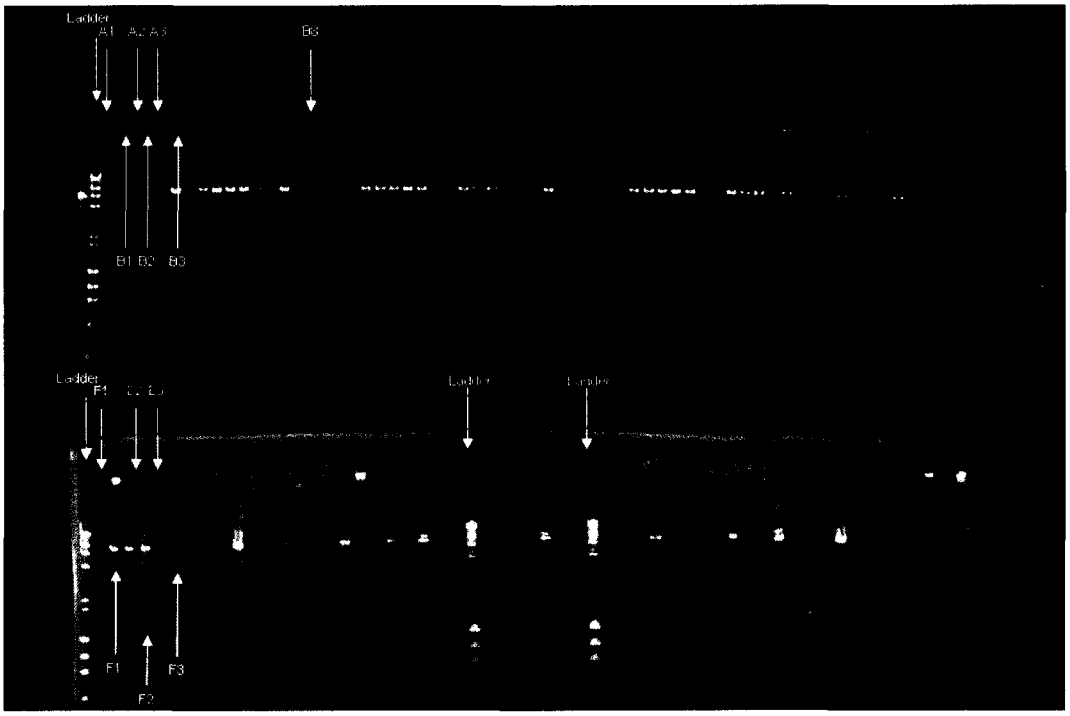

Figure 40 Initial screen for deletion alleles at the *ASK7* locus. A1 represents a negative control. A2 through F2 correspond to DNA templates from Master plate 1 of DNA pools. F2 shows a possible deletion allele for *ask7* due to the lower molecular weight bands seen in the lane. The wild type amplicon for primers *ASK7*-7925-For and *ASK7*-7925-REV (7925 bps) can be seen in lanes B3, E2, and F1.

The suspected deletion allele detected in Figure 40 corresponds to Well F2 from Master Plate 1 which was derived from the M2 mutant DNA pool β11C; due to the volume of tissue harvested from pool β11C the subsequent DNA is actually represented twice in Master Plate 1 in wells B8 and F2. Thus the initial screen for *ask7* mutant alleles

included a technical replicate for β11C in B8 as well as F2. Seen in Figure 40, Master Plate 1 well B8 showed no amplification. Since two primer pairs were observed to produce large amplicons for the wild type *ASK7* allele (Figure 39) these were used to validate whether the amplification pattern seen for β11C was a true positive. Both amplicons relied on the use of *ASK7*-7925-REV and so the forward primer was varied between *ASK7*-7925-FOR and *ASK7*-9931-FOR in samples containing the wild type (Col-0) or Master plate 1 wells B8 and F2 as template (Figure 41).

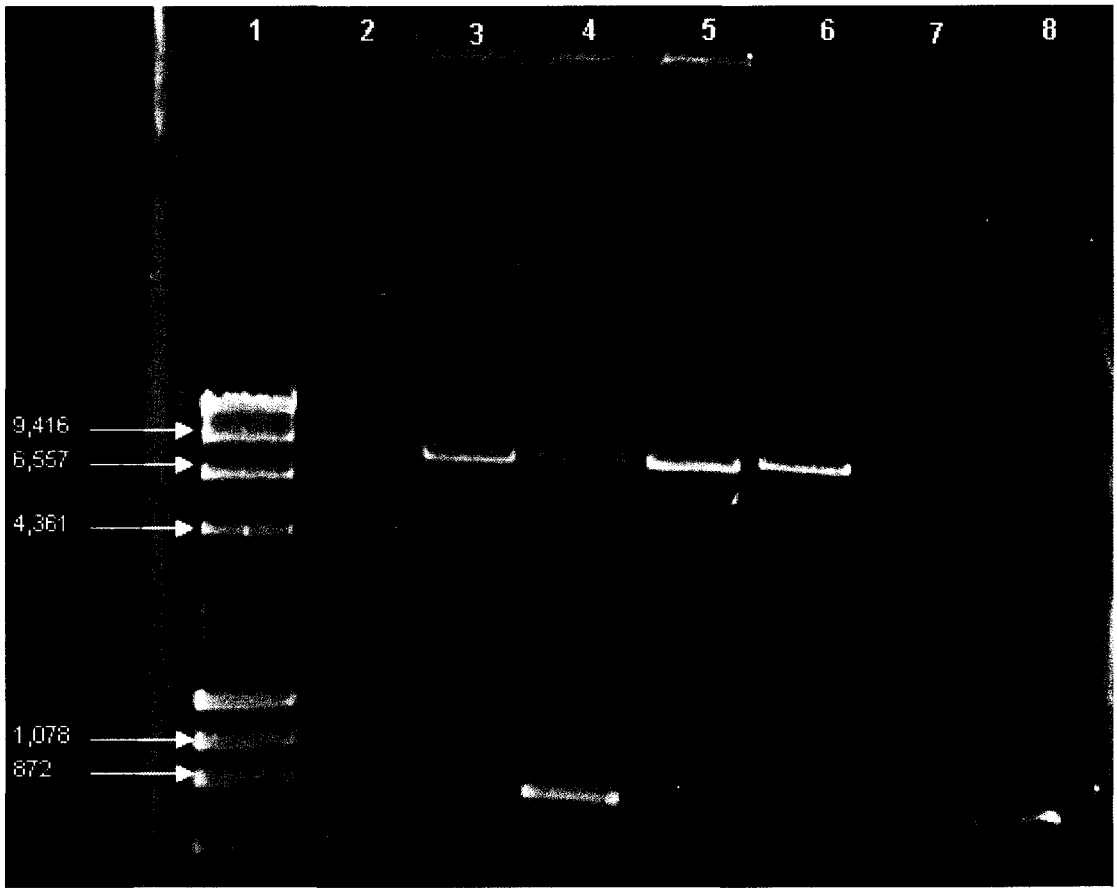

Figure 41 Validation of potential *ask7* deletion alleles. All reactions used ASK7-7925-REV as the reverse primer in combination with template DNA from well F2 of Master plate 1. Lanes 2, 3, 5 and 7 used *ASK7*-7925-FOR as the forward primer (7925 bps). Lanes 4, 6 and 8 used *ASK7*-9931-FOR as the forward primer (8453 bps). Lane 2; negative control (no template DNA), Lanes 3 and 4; wild-

type (Col-0) template DNA, Lanes 5 and 6; template DNA from Master plate 1 well F2, Lanes 7 and 8; template DNA from Master plate 1, well B8.

As shown in Figure 41 the amplification of pool β11C samples did not result in the amplification of a deletion allele for the ASK7 locus. While the initial partial screen was unsuccessful in identifying a pool with an *ask7* mutant, the screening methodology presented here does provide a technical validation of the reverse genetics system for subsequent application in screening of mutant DNA pools.

The efficacy of pin tool based transfer as examined by using a pin tool to transfer a Lambda DNA standard. As is presented in Figure 42 and Figure 43 the lambda DNA standard had an R^2 of 0.99. Pin tool based transfer did introduce variation, leading to an R^2 of 0.91, but given that the coefficient of correlation was high pin tools were investigated for their efficacy in the assembly of Phusion™-based PCRs.

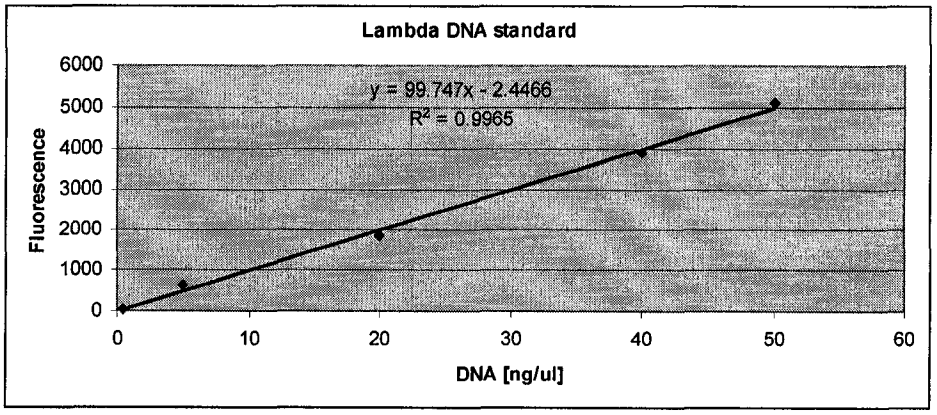

Figure 42 Lambda DNA standard curve quantified with PicoGreen.

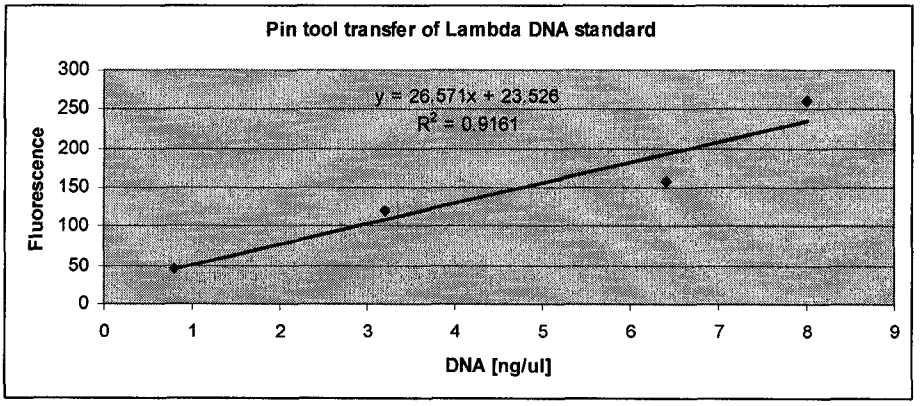

Figure 43 Lambda phage DNA standard curve using pin tool transfer. DNA was transferred using a 96-well pin tool and quantified as described for Figure 42.

Evaluation of pin tool based transfers of template DNA was performed through the use of pin tools to assemble Phusion™-based PCRs. Transfers were performed on M2 DNA pools to transfer DNA templates from 96 well microtitre plates. As is presented in Figure 44 the pin tool based transfers provide a reasonable approach to assembling PCRs but further work will be required to evaluate the efficacy of pin tool based M2 DNA transfers at larger scales (entire 96 well plates).

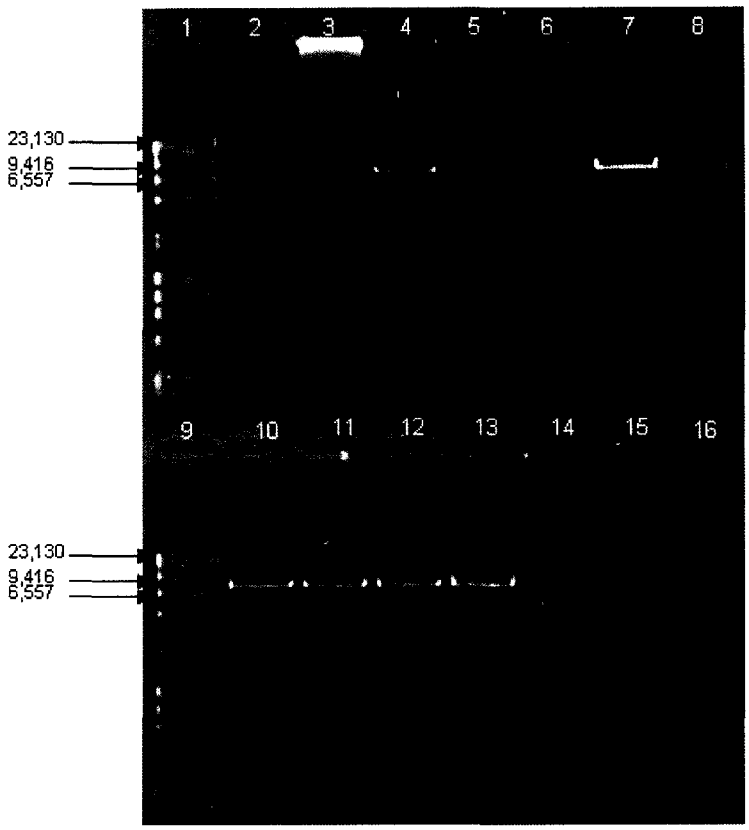

Figure 44 Pin tool transfer for the assembly of Phusion™-based PCR. All reactions used the primers *ASK7*-7925-FOR and *ASK7*-7925-REV and all reactions had the template DNA transferred using a pin tool. Lanes 1 and 9; molecular size standards, Lane 2; negative control (no template DNA), Lane 3; positive control amplification using Col-0 template DNA, Lanes 4-8 and 10-16; amplifications of DNA prepared from different M2 DNA pools.

Discussion

In this study I present details outlining the development of a mutant reverse genetic population and resource suitable for the identification of deletion alleles of select loci in Arabidopsis. Seed and DNA stocks were collected in pools comprising about 140 distinct M1 plants. Using a series of primers for distinct loci (*UFO*, *ACT8* and *ZTL*) it has been shown (Figure 33 and Figure 34) that the genomic DNA produced by the modified Qiagen DNeasy protocol used in this study is suitable for uniplex PCR of targets up to

2.5kbps in size using Taq DNA polymerase. This study also demonstrates the ability to amplify large (~ 9kbps) targets (*CRY2* and *ASK7*) using Phusion™ DNA polymerase.

While this study included a complete protocol for exploitation and screening of deletion mutants of Arabidopsis, further development and optimization of the actual screening procedure will be required. As described here, the mutant DNA pools have been organized into four 96-well microtitre plates, given the ready availability of 384-well thermocyclers and thereby the ability to screen the entire population in 1 microtitre plate the mutant DNA pools should be re-assembled and deployed in this high density format. Upon organization in 384-well format there are a series of quality control assessments and reconstruction experiments should then be investigated, as described here for the 96 well format.

Give the cost differential between Phusion™ and Taq based PCRs, it may be advantageous to focus future applications on the amplification of targets like the 2.5kbps target from *ZTL*. By focusing on the amplification of smaller products in combination with Taq DNA polymerase it should be possible to optimize the logistics surrounding the large scale screening (issues of scale) while reducing the costs associated with the development cycle. Similar investigation of alternative sample handling (reduction of mechanical agitation and reduction in DNA template) could also be performed using Taq DNA polymerase prior to validating the same findings with Phusion™ DNA polymerase.

With the improvement of the screening techniques a complete screen for *ask7* mutant alleles should be performed so as to complete the validation of the methodology presented here for the identification of deletion mutants. Subsequent genes of interest would likely be expanded to other members of the ASK gene family in the hope that the

issue of genetic redundancy between select gene family members could be more generally analyzed.

Another suggested target for early screening would be the C-Repeat Binding Factor2 (*CBF2*) locus. The *CBF1* locus encodes a single exon gene which regulates the expression of cold responsive genes in Arabidopsis. CBF genes have been shown to be involved in the regulation of cold tolerance in *Arabidopsis thaliana* [117-121]. There are four genes which seem to comprise the CBF gene family in *Arabidopsis thaliana*, three of which are encoded together on Chromosome 4 (*CBF1*, *CBF2*, and *CBF3*). By focussing on the *CBF2* locus (for which no T-DNA based mutant allele is available) and screening for a 10kb deletion it may be possible (as proposed for the *ASK* gene family) to identify a deletion spanning multiple genes of the *CBF* family.

The major limitation to the deletion population described in this study is the number of deletion screens possible. The mDNeasy procedure typically yielded 10μg of genomic DNA / mutant pool. For Phusion™ based PCR as performed in this study the amount of template required in ~ 50ng. Therefore this mutant population represents enough template DNA to perform ~ 200 reactions. Depending on the number of optimization control reactions required (following up on suspected positives with secondary primer pairs) the mutant DNA pools described in this study probably represent enough template to screen for 100-150 distinct loci.

In order to generate additional mutant DNA for additional screening it would be possible to grow out another generation from the M2 seed and subsequently perform mDNeasy DNA extractions and organize seed for pools of M3 plants. Whole genome amplification of the existing M2 DNA pools could avoid the need for a subsequent large-

scale grow-out, however, the use of Phi^{29}-based whole genome amplification was found not to be appropriate for use in Phusion^{TM} based PCR screening reactions. The use of narrow-bore micropipette tips would be anticipated to contribute to DNA shearing during the isolation of genomic template DNA. It is likely that the use of normal pipette tips in the mDNeasy purification of DNA from the mutant pools results in DNA ≤100kbps in length. Phi^{29} has been shown to produce at best 70kbps products [114]. It therefore seems likely that a major factor affecting the ability to use Phi^{29} amplified mDNeasy purified genomic DNA is that the template product length is too short. If the template DNA length is ≤ 70kbps then there will be a reduction in the number of full length targets available in the subsequent Phusion^{TM} based PCR involving large-span locus specific primers. At the time of this study only one formulation for Genomiphi^{TM} was available, although the vendor (GE Healthcare) announced the development of two additional formulations of Genomiphi^{TM} for the generation of longer products, but which were not available for evaluation in this study.

In addition to investigating the efficacy of new Phi^{29} formulations initial work presented in this study has been done to evaluate the use of pin-tool transfers to add template DNA during the assembly of Phusion^{TM}-based PCRs. By examining the efficacy of Pin tools to transfer a DNA standard it was demonstrated in this study that the consistency exhibited by pin tools is suitable for PCR. Further, Pin tools were tested for their ability to transfer sufficient template to detect the wild type *ASK7* amplification pattern with success. Future use of pin tools for transfers of mutant DNA templates would serve two functions: adherence-based transfer would limit shearing (since no mechanical agitation or pipetting of the DNA template would be involved) and, secondly,

would likely extend the number of deletion screens possible with this resource. If 130 nl transfers prove feasible for the deletion screening then the existing mutant DNA pools (200µl / pool at 50ng /µl) would potentially enable the screening of 1500 different loci without dilution of the existing DNA stocks.

The limitation on the number of mutant screens possible is the principle factor limiting the wide distribution of a mutant resource such as the one described in this study. Future work to investigate the efficacy of new methodologies for whole genome amplification and logistical approaches to maximize the utilization of the mutant DNA pools will likely serve to extend the mutant DNA pools developed in this study to the level of a resource available for the wider Arabidopsis research community.

At present the Arabidopsis research community relies heavily on T-DNA insertion mutant collections as a mechanism of identifying null phenotypes. T-DNA mutants in particular have received wide community utilization. While T-DNA mutants provide an approach to identifying null-phenotypes for target alleles there are issues with the availability as well as genetic penetrance of the available mutant alleles using such an approach. It is commonplace that publications relying on T-DNA mutants will require multiple insertions in a target allele, supplemented by gene expression data for genes immediately proximal to the insertion, in order to be informative for the description of the true null phenotypes. Thus the effort is significant in order to demonstrate, with certainty, the phenotype associated with an arbitrary T-DNA insertion as a null-allele. The advantage of a deletion mutant is that, generally, such alleles are more definitive with respect to the genetic basis of any observed phenotype. There is a logistic factor with respect to the identification of a deletion mutant as described in this work. Once an

individual plant has been identified as carrying a deletion allele of interest, it is necessary to cross the mutant with the wild type and show that the phenotype faithfully segregates with the deletion through multiple successive out-crosses to the wild type. Existing evidence in Arabidopsis suggests that a Fast neutron dose of 60Gys results in deletions of an average size Kb [102,103]. Thus at least 5 outcross generations would be required to reasonably isolate the mutant allele of interest from other background deletions present in the M1 parent.

At present there is limited existing data regarding the frequency of other mutations such as recombination in Arabidopsis plants subjected to fast neutron irradiation. Through collaboration which intersected with this study attempts were made to survey individual deletion mutants by genomic hybridization to whole genome microarrays for Arabidopsis. The ability presently exists to employ whole genome tiling arrays to analyze the genome of Arabidopsis plants at high resolution [122-124]. Through the use of whole genome tiling arrays the possibility exists to map multiple, and potentially all, deletions within a Fast neutron irradiated mutant. Also given that array based technologies can typically interrogate multiple samples at once through the use of multiple fluorescent dyes and that most array platforms can be re-used multiple times, future work to employ array based mapping of deletion alleles will likely provide higher throughput than PCR based screening approaches as described in this study. To date work to probe whole genome arrays with fast neutron mutated genomic DNA has not proved fruitful. Future developments involving the use of array based procedures will likely result in the ability for high-throughput identification of deletion mutants and provide initial data on the affect of fast neutron irradiation at the whole genome level.

The impact that reverse-genetic resources such as the SiGNaL T-DNA population have made on the Arabidopsis research community, not to mention plant science in general, has been immense. For example, as of May 2007, the foundation publication describing the Arabidopsis SiGNaL T-DNA mutant resource [88] has been cited over 780 times and the citation frequency is growing making this among the most influential plant publications of the decade. Given that there are issues relating to the penetrance of the insertional alleles generated by T-DNA based mutagenesis, together with lack of complete mutational coverage of the Arabidopsis genome it is likely that a publicly available deletion mutant resource such as that described here would be of similar value to the extend the existing genetic resources for Arabidopsis.

APPENDICES

APPENDIX A

Wheat EST resources for functional genomics of abiotic stress

Houde, M., Belcaid, M., Ouellet, F., Danyluk, J., Monroy, A. F., Dryanova, A., Gulick, P., Bergeron, A., Laroche, A., Links, M. G., MacCarthy, L., Crosby, W. L., and Sarhan, F. (2006) *BMC. Genomics* 7, 149

Sequencing tracefiles were obtained for 110,544 cloned cDNAs. The wheat EST libraries involved represented a series of tissues and developmental stages as described in [125]. Computationally based annotation of the 110,544 EST sequences significantly increased the global data for *Triticum aestivum* through the deposition of 80,821 EST sequences into dbEST [126]. As of July 31, 2006 the international repository, Unigene build # 46 for *T. aestivum,* represented 746,185 ESTs. Therefore the collection analyzed in this work represented 14% of the total data publicly available. Exploiting the EST data it was possible to contribute in an international consortium through the design of a 17,000 feature microarray for *T. aestivum*. Drawing on the non-redundant set of contigs and singletons derived from the EST sequences in addition to collections from the United States Department of Agriculture and the Waite Campus in Australia; 17,000 distinct cDNA sequences were identified as targets for further gene expression studies. For each cDNA target a 70 nucleotide probe sequence was designed. Using a parallel distributed program (unpublished) probes were designed by testing 70mers in the cDNA from a 3' to 5' direction. Probes were identified on the basis of Tm, GC, quality, and specificity heuristics. The resulting probe set has been designed and is presently being used to investigate gene expression in *T. aestivum* by a number of labs throughout the world.

BMC Genomics

BioMed Central

Research article

Open Access

Wheat EST resources for functional genomics of abiotic stress

Mario Houde[1], Mahdi Belcaid[2], François Ouellet[1], Jean Danyluk[1], Antonio F Monroy[3], Ani Dryanova[3], Patrick Gulick[3], Anne Bergeron[2], André Laroche[4], Matthew G Links[5], Luke MacCarthy[6], William L Crosby[5] and Fathey Sarhan*[1]

Address: [1]Département des Sciences biologiques, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montréal QC, H3C 3P8, Canada, [2]Département d'Informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montréal QC, H3C 3P8, Canada, [3]Biology Department, Concordia University, 7141 Sherbrooke Street West, Montreal QC, H4B 1R6, Canada, [4]Agriculture et Agroalimentaire Canada, Centre de recherches de Lethbridge, 5403, 1st Avenue South, C.P. 3000, Lethbridge AB, T1J 4B1, Canada, [5]Department of Biological Sciences, University of Windsor, 401 Sunset ave, Windsor ON, N9B 3P4, Canada and [6]Department of Computer Science, University of Saskatchewan, 176 Thorvaldson Building, 110 Science Place, Saskatoon SK, S7N 5C9, Canada

Email: Mario Houde - houde.mario@uqam.ca; Mahdi Belcaid - belcaid.mahdi@courrier.uqam.ca; François Ouellet - ouellet.francois@uqam.ca; Jean Danyluk - danyluk.jean@uqam.ca; Antonio F Monroy - amonroy@power2will.com; Ani Dryanova - adryanov@alcor.concordia.ca; Patrick Gulick - pgulick@alcor.concordia.ca; Anne Bergeron - bergeron.anne@uqam.ca; André Laroche - laroche@agr.gc.ca; Matthew G Links - links@uwindsor.ca; Luke MacCarthy - mccarthy@cs.usask.ca; William L Crosby - bcrosby@uwindsor.ca; Fathey Sarhan* - sarhan.fathey@uqam.ca

* Corresponding author

Published: 13 June 2006

BMC Genomics 2006, **7**:149 doi:10.1186/1471-2164-7-149

Received: 08 March 2006

Accepted: 13 June 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/149

Abstract

Background: Wheat is an excellent species to study freezing tolerance and other abiotic stresses. However, the sequence of the wheat genome has not been completely characterized due to its complexity and large size. To circumvent this obstacle and identify genes involved in cold acclimation and associated stresses, a large scale EST sequencing approach was undertaken by the Functional Genomics of Abiotic Stress (FGAS) project.

Results: We generated 73,521 quality-filtered ESTs from eleven cDNA libraries constructed from wheat plants exposed to various abiotic stresses and at different developmental stages. In addition, 196,041 ESTs for which tracefiles were available from the National Science Foundation wheat EST sequencing program and DuPont were also quality-filtered and used in the analysis. Clustering of the combined ESTs with d2_cluster and TGICL yielded a few large clusters containing several thousand ESTs that were refractory to routine clustering techniques. To resolve this problem, the sequence proximity and "bridges" were identified by an e-value distance graph to manually break clusters into smaller groups. Assembly of the resolved ESTs generated a 75,488 unique sequence set (31,580 contigs and 43,908 singletons/singlets). Digital expression analyses indicated that the FGAS dataset is enriched in stress-regulated genes compared to the other public datasets. Over 43% of the unique sequence set was annotated and classified into functional categories according to Gene Ontology.

Conclusion: We have annotated 29,556 different sequences, an almost 5-fold increase in annotated sequences compared to the available wheat public databases. Digital expression analysis combined with gene annotation helped in the identification of several pathways associated with abiotic stress. The genomic resources and knowledge developed by this project will contribute to a better understanding of the different mechanisms that govern stress tolerance in wheat and other cereals.

Background

Cold acclimation (CA) allows hardy plants to develop the efficient freezing tolerance (FT) mechanisms needed for winter survival. During the period of exposure to low temperature (LT), numerous biochemical, physiological and metabolic functions are altered in plants, and these changes are regulated by LT mostly at the gene expression level. The identification of LT-responsive genes is therefore required to understand the molecular basis of CA. Cold-induced genes and their products have been isolated and characterized in many species. In wheat and other cereals, the expression of several genes during cold acclimation was found to be positively correlated with the capacity of each genotype and tissue to develop FT [1]. Furthermore, abiotic stresses that have a dehydrative component (such as cold, drought and salinity) share some responses. It is therefore expected that, in addition to the genes regulated specifically by each stress, some genes will be regulated by multiple stresses. The availability of wheat genotypes with varying degree of FT makes this species an excellent model to study freezing tolerance and other abiotic stresses. The identification of new genes involved in the cold response will provide invaluable tools to further our understanding of the metabolic pathways of cold acclimation and the acquisition of superior freezing tolerance of hardy genotypes.

Major genomics initiatives have generated valuable data for the elucidation of the expressed portion of the genomes of higher plants. The genome sequencing of *Arabidopsis thaliana* was completed in 2000 [2] while the finished sequence for rice was recently published [3]. The relatively small genome size of these model organisms was a key element in their selection as the first plant genomes to be sequenced with extensive coverage. On the other hand, the allohexaploid wheat genome is one of the largest among crop species with a haploid size of 16.7 billion bp [4], which is 110 and 40 times larger than *Arabidopsis* and rice respectively [5]. The large size, combined with the high percentage (over 80%) of repetitive non-coding DNA, presents a major challenge for comprehensive sequencing of the wheat genome. However, a significant insight into the expressed portion of the wheat genome can be gained through large-scale generation and analysis of ESTs. cDNA libraries prepared from different tissues exposed to various stress conditions and developmental stages are valuable tools to obtain the expressed and stress-regulated portion of the genome. This approach was used in several species such as oat [6], barley [7], tomato [8] and poplar [9]. The sequencing of cDNAs gives direct information on the mature transcripts for the coding portion of the genome that can subsequently be used for gene identification and functional studies. The availability of wheat genomics data in the public datasets has grown rapidly through major initiatives [10,11]. However, additional ESTs are needed to complete the identification of the expressed genes under different growth conditions and from different genotypes. This will contribute to a more complete representation of the genome through identification of new genes and extension of contigs for the majority of genes that have incomplete sequence coverage. Towards this goal, the Functional Genomics of Abiotic Stress (FGAS) program initiated an EST sequencing effort directed toward the study of abiotic stress, with an emphasis on cold acclimation [12]. To increase gene diversity in the EST population and increase the probability of identifying those associated with freezing tolerance, different cDNA libraries were prepared from winter wheat tissues exposed for various times to low temperature, together with select libraries derived from tissues exposed to other stresses or at different developmental stages. In this report, we describe the generation of 73,521 high quality ESTs from wheat stress-associated cDNA libraries. In order to perform the assembly and digital expression analyses, these ESTs were supplemented with wheat ESTs for which sequence quality data was available. These include the NSF [13] and DuPont datasets, which will be referred to as the 'NSF-DuPont' dataset in this report. Digital expression analyses identified a large number of genes that were associated with cold acclimation and other stresses. Expression analyses and functional classification provided important information about the different metabolic and regulatory pathways that are possibly associated with cellular adjustment to environmental stresses. These new EST resources are an important addition to publicly available resources especially in relation to the study of abiotic stresses in cereals.

Results and discussion

The large-scale FGAS wheat EST sequencing project was undertaken to identify new genes associated with abiotic stress and to provide physical resources for functional studies. We have developed a unique wheat EST resource from eleven cDNA libraries prepared from tissues at different developmental stages and exposed to different stress conditions (Table 1). The EST collections from FGAS, NSF and DuPont were analyzed and classified into functional categories.

Assembly and identification of new wheat genes

We have used EST sequences and quality values from the corresponding tracefiles of large datasets (FGAS, NSF and DuPont) to assemble 75,488 different wheat sequences (31,580 contigs, 36,388 singletons and 7,520 singlets). Among these datasets, the FGAS project produced 11,225 unique sequences (2,824 contigs, 6,663 singletons and 1,738 singlets) indicating that the FGAS ESTs encompass a large subset of unique transcripts. These sequences were analyzed using BLASTN on the db_est database and filtered for wheat sequences with two different cut-off e-val-

ues to identify new wheat genes. With an e-25 cut-off value, we found that 2,304 genes had no homologous wheat ESTs (Table 2). After filtering these genes against the wheat protein database with TBLASTX, there were still 2,243 proteins showing no homology to known proteins. With an e-05 cut-off, 1,581 genes had no homologs in wheat. After filtering these against the protein database, 1,470 non-homologous sequences remained. These unique wheat sequences were then BLASTed against *Arabidopsis*, rice, and finally nr db EST (Table 2). In *Arabidopsis*, we found that only 5 of the remaining FGAS wheat sequences had a strong (e-25) similarity using BLASTN while 253 of the remaining sequences had homologs when filtered with the *Arabidopsis* protein database (count down to 1,985). A similar trend was found in *Arabidopsis* using a lower sequence similarity cut-off (e-05). The remaining unique gene count was reduced by several hundred after comparing protein homologs in rice (counts down to 1674 at e-25 and down to 855 at e-05) demonstrating that several genes common between rice and wheat are absent in *Arabidopsis* (Table 2). The remaining unique ESTs were BLASTed against the non redundant database to determine whether homologs were present in other organisms. At an e-05, there were 795 ESTs showing no significant similarity to known domains in genes from other species. It is possible that some of these genes derive from unknown micro-organisms contaminating the plant tissues, and/or from residual genomic DNA in the RNA samples used for cDNA synthesis. However, the majority of these sequences have ORFs encoding proteins larger than 30 amino acids, with an average predicted protein size of over 100 amino acids. This suggests that the unidentified genes do represent novel wheat genes.

The Institute for Genomic research (TIGR) wheat gene index (Release 10.0) shows that only 6,431 of the 44,954 wheat contigs (14%) were successfully allocated a known Molecular Function using Gene Ontology, compared to the classification done for *Arabidopsis* in which 12,558 of the 28,900 contigs (42%) have a known Molecular Function. Therefore, prior to this report, *Arabidopsis* had almost twice as many genes annotated with at least one defined function compared to wheat (12,558 vs 6,431). The classification of the complete dataset (FGAS and NSF-DuPont datasets) allowed the tentative annotation of 43.3% of the genes. As expected, most of the annotated sequences were in contigs (57.6%) while the percentage of annotated singletons/singlets was much lower (30.8%). We have thus been able to functionally annotate 29,556 different sequences, an almost 5-fold increase in annotated sequences compared to TIGR. This is a significant contribution that broadens the available wheat public annotation dataset for downstream functional studies. These results demonstrate that a large number of wheat genes are poorly characterized and stress the fact that major efforts in functional analyses are needed.

Enrichment for stress-regulated genes in the FGAS dataset

Comparative analysis of the FGAS ESTs and NSF-DuPont ESTs based on Gene Ontology (GOslim) showed that several GO classes are more represented in FGAS than in the NSF-DuPont dataset (Figure 1). When general GO classes are compared (GOs 1 to 3; Biological Process, Transcription and Protein Metabolism), no major differences in the number of ESTs were found. Similarly, most GOslim classes showed less than 25% difference between the two datasets. However, GOs 4 and 5 (Enzyme Regulator Activity and Nutrient Reservoir Activity) had a lower representation while GOs 6 to 15 (Transcription Factor Activity, Nuclease Activity, Plasma Membrane, Secondary Metabolism, Response to External Stimulus, Carbohydrate Binding, Response to Abiotic Stimulus, Cell-Cell Signalling, Development and Behavior) were more abundant in the FGAS dataset (Figure 1).

To identify genes that are differentially represented between the two datasets, the relative abundance of ESTs was analyzed and referred to as digital expression analysis. For each contig, the number of ESTs from FGAS (excluding ESTs derived from Suppressive Subtractive Hybridization; SSH) was divided by the number of ESTs from NSF-DuPont and the ratio was normalized to correct for the difference in size between the two datasets (54,032 non SSH EST sequences for the FGAS dataset and 196,041 sequences for the NSF-DuPont dataset). Thus, after normalization, the relative expression level for a contig having 1 EST from each dataset would result in a relative expression of 3.62X in FGAS compared to NSF-DuPont (a ratio of 1 multiplied by 196,041/54,032). Since the SSH technique aims to enrich differentially expressed cDNAs, the ESTs derived from the SSH libraries were analysed separately to avoid a bias in the number of ESTs in a contig, which could invalidate the digital expression analysis approach.

The data indicated that over 75% of the contigs have ratios that vary by less than two-fold, suggesting a similar representation of ESTs between the FGAS (less SSH) and the NSF-DuPont datasets. The remaining 25% of contigs showed more than two-fold difference in abundance (Table 3; see additional file 1: Table1.xls) in the FGAS dataset. When 5- and 10-fold ratios are used as cut-off, 6.6% and 1.7% of the contigs are retained respectively. Most of the differences are due to genes that are over-represented in the FGAS dataset (for the 5-fold cut-off, 1959 genes are over- and 136 genes are under-represented, see Table 3). With a higher cut-off (20-fold differential abundance), only 61 contigs are over expressed and 5 are under-expressed. An analysis of these highly over-repre-

A

GO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
FGAS	26,016	1,378	8,316	262	276	444	230	76	399	971	284	488	31	383	52	39,606
% of Total GOs	65.7	3.5	21	0.66	0.7	1.12	0.58	0.19	1.01	2.45	0.72	1.23	0.08	0.97	0.13	100
NSF-DuPont	86,093	3,704	31,763	2,791	1,878	1,033	522	156	769	1,869	446	615	37	411	53	132,140
% of Total GOs	65.2	2.8	24	2.11	1.42	0.78	0.4	0.12	0.58	1.41	0.34	0.47	0.03	0.31	0.04	100

B

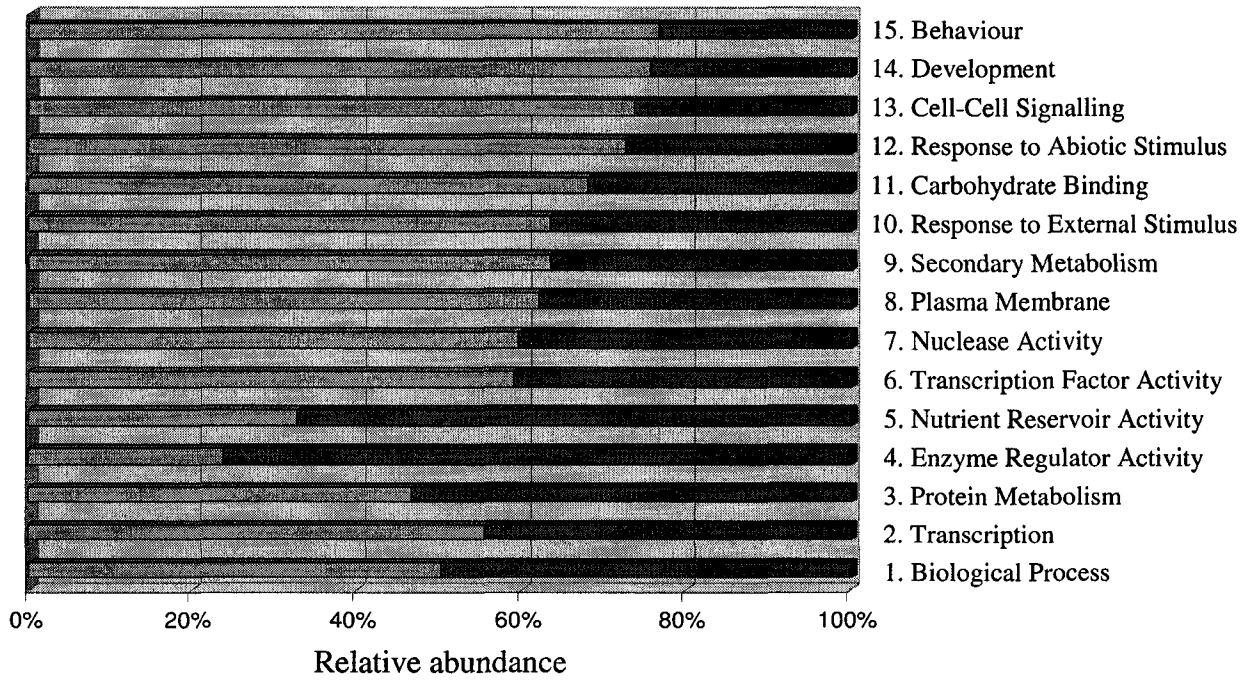

Figure 1

Abundance of annotated ESTs in FGAS contigs relative to NSF-DuPont contigs within select GO classes. A) Number of annotated ESTs. The GO counts were added for each dataset and the percentage of ESTs for each GO was calculated based on this total count. **B)** The relative abundance for each GO is compared between the FGAS (blue) and the NSF-DuPont (red) datasets by comparing the percentage of each GO as determined in A. GO categories: 1. Biological Process GO:0008150; 2. Transcription GO:0006350; 3. Protein Metabolism GO:0019538; 4. Enzyme Regulator Activity GO:0030234; 5. Nutrient Reservoir Activity GO:0045735; 6. Transcription Factor Activity GO:0003700; 7. Nuclease Activity GO:0004518; 8. Plasma Membrane GO:0005886; 9. Secondary Metabolism GO:0019748; 10. Response to External Stimulus GO:0009605; 11. Carbohydrate Binding GO:0030246; 12. Response to Abiotic Stimulus GO:0009628; 13. Cell-Cell Signalling GO:0007267; 14. Development GO:0007275; 15. Behaviour GO:0007610.

sented contigs showed that a good proportion (52%) of these show homology to genes that were previously reported to be over-expressed under stress (see references in Table 4). This high percentage of positive identification suggests that the NSF-DuPont collection was a good reference dataset for digital expression analysis of the FGAS dataset.

Our digital expression analysis relies on the presence of ESTs from both datasets in a same contig (since we cannot

Table 1: Summary of tissues used for the different cDNA libraries generated for the FGAS EST sequencing project.

Library	Growth conditions*	Tissues	High quality EST sequences
Library 2	Control plants; Plants cold acclimated for 1, 23 and 53 days	leaves and crowns	25,240
Library 3	Control plants; Plants cold acclimated for 1, 23 and 53 days; Plants salt stressed for 0.5, 3 and 6 hours	roots	11,382
Library 4	Plants dehydrated on the bench (4 time points) and in a growth chamber (4 time points)	leaves and crowns	2,838
Library 5	Various vernalization and developmental stages through spike formation.	crowns and flowers	6,668
Library 6	Control plants; Plants cold acclimated for short time points (1, 3 and 6 hours) under light or dark conditions	leaves and crowns	7,904
TaLT2	SSH library: Tester: cv. CI14106 cold acclimated for 1 day; Driver: cv Norstar cold acclimated for 21 and 49 days	crowns	2,271
TaLT3	SSH library: Tester: cv. CI14106 cold acclimated for 21 and 49 days; Driver: cv Norstar cold acclimated for 1 day	crowns	1,832
TaLT4	SSH library: Tester: cv. PI178383 cold acclimated for 1 day; Driver: cv Norstar cold acclimated for 21 and 49 days	crowns	2,716
TaLT5	SSH library: Tester: cv. PI178383 cold acclimated for 21 and 49 days; Driver: cv Norstar cold acclimated for 1 day	crowns	2,784
TaLT6	SSH library: Tester: cv. CI14106 cold acclimated for 1 day; Driver: non-acclimated cv. CI14106	crowns	4,961
TaLT7	SSH library: Tester: cv. CI14106 cold acclimated for 21 and 49 days; Driver: non-acclimated cv. CI14106	crowns	4,925

* Libraries 2 to 6 were constructed from wheat cv Norstar.

divide by 0). We have also identified 542 contigs that contained at least 3 ESTs from FGAS but none from the NSF-DuPont dataset (See additional file 2: Table 2.xls). Table 5 lists the 90 genes that contain at least 5 ESTs unique to FGAS, and many of these are similar to genes that have previously been reported to be over-expressed under stress. Although the unique contigs in the FGAS dataset may represent transcripts that are specific to the cultivar used in our study, there is a possibility that they may represent novel genes that are induced by environmental stress.

Table 2: Homology search of FGAS contigs. As a first step, the 11,225 FGAS unique sequences were analyzed using the wheat-filtered db_est (NCBI release 2.2.12, Aug-07-2005). The non-homologous transcripts were then analyzed against the wheat protein database to subtract protein homologs. The remaining transcripts were then analyzed in the same manner against the *Arabidopsis* and rice databases and finally against the nr database. The complete homology search was performed at e-25 and e-05 cut-offs. The numbers indicate the number of genes that do not show any homology at the indicated e-value cut-off.

		e-25	e-05
Wheat	BLASTN db_est	2304	1581
	TBLASTX	2243	1470
Arabidopsis	BLASTN db_est	2238	1470
	TBLASTX	1985	1102
Rice	BLASTN db_est	1845	987
	TBLASTX	1674	855
nr db_est	BLASTN	1623	795

In *Arabidopsis*, microarray experiments have shown that about 10% of the genes are over-or under-expressed by at least two-fold upon exposure to cold acclimation conditions [14]. Based on our previous northern and microarray analyses, we have estimated that the same proportion of wheat genes is cold-regulated (Sarhan *et al.*, unpublished results). If we consider a conservative estimate of 30,000 wheat genes (90,000 if we consider the A, B and D genomes), this means that around 3,000 genes would be cold-regulated. A similar number of genes was identified when we used a 5-fold cut-off differential expression (2,095 differentially expressed contigs, Table 3) and added the 542 contigs having at least 3 ESTs that are unique to the FGAS dataset. Using these criteria, our analyses resulted in a total of 2,637 contigs or 8.4% of the contigs generated in our assembly (31,580 contigs). Considering that 95% of the EST sequences were derived from libraries constructed from cold-acclimated plants, these genes represent candidate genes likely regulated by low temperature and other stresses. However, many of these may be differentially expressed as a consequence of the temperature shift and metabolic adjustment and might not be involved in conferring or regulating increased tol-

Table 3: Contigs containing ESTs that are over or under-represented in the FGAS dataset relative to the NSF-DuPont dataset.

Fold increase/decrease	Over-represented ESTs	Under-represented ESTs	Total	Percent of total contigs (31,772)
20	61	5	66	0.2
10	533	22	555	1.7
5	1959	136	2095	6.6
3	5569	489	6052	19
2	6794	1047	7841	24.7

erance to stress. It would be of interest to analyse these 2,637 genes to identify those relevant to LT tolerance and other stresses in cereals. To verify the conservation of the stress response between wheat and *Arabidopsis*, we first identified the *Arabidopsis* proteins having homology (e-25) to the 2,637 wheat proteins identified in our study, using the TAIR protein database. The homology search resulted in the identification of 1,551 *Arabidopsis* proteins. Most of the genes encoding these proteins are represented on the Affymetrix and MWG microarrays. This allowed us to obtain their expression profiles from the available public data [14,15]. Our analysis indicated that 941 genes are cold-regulated and 890 are drought-regulated (See additional file 1: Table 1.xls and additional file 2: Table 2.xls). There are 678 genes regulated by both stresses, with a total of 1153 different *Arabidopsis* genes that are stress-regulated. Therefore, there are over 44% of the 2,637 putative wheat stress-regulated genes that have a homolog regulated by stress in *Arabidopsis*, suggesting overlapping responses between the two species.

As a complementary approach to identifying new wheat genes that may be differentially expressed, different SSH libraries were produced to identify genes over-expressed after brief (1 day) or long (21–49 days) periods of cold acclimation. Different cultivars that may help to identify other components of freezing tolerance such as pathogen resistance to snow molds were used for these analyses. A total of 3,873 contigs containing 18,610 SSH ESTs were obtained with 2,969 contigs (76.7%) tentatively annotated. Unique contigs from SSH libraries are potentially a good source to mine for new genes associated with cold acclimation. Overall, 225 contigs unique to the SSH libraries (See additional file 3: Table 3.xls) were identified, among which 74 were annotated (Table 6). We found that 11 of the 74 annotated SSH contigs (or 15% of the unique SSH contigs) have corresponding genes (high similarity based on BLASTX e-values) that are over-expressed more than 5-fold in the differentially-expressed FGAS contigs. These results suggest that unique SSH contigs contain candidate genes that could be involved in abiotic stress tolerance.

Metabolic pathways associated with differentially expressed genes

GO slim annotation was used to subdivide the 2,637 stress-regulated genes into function categories to gain insight into their putative role during cold acclimation and abiotic stresses. The results show that a large proportion of these contigs were annotated under a limited number of GO classes (Figure 2). Over 53.7% of the contigs were grouped into 14 GO categories while 27.5% of the contigs were designated "No Gene Ontology" and 4.2% were classified as "Hypothetical Protein", a term used to designate open reading frames predicted from the *Arabidopsis* or rice genomic DNA. The remaining contigs with other GO categories were grouped together in one category (14.6%).

A plethora of physiological and metabolic adjustments occur during cold acclimation and in response to other stresses. The regulation of genes involved in temperature, drought and salt stresses is known to reflect the cross-talk between different signalling pathways [16]. However, few studies have identified multiple genes that are stress-regulated and that belong to a same metabolic pathway. Our analyses enabled us to position several genes in their respective metabolic pathway, suggesting that these pathways are involved in stress responses. Since it is beyond the scope of this report to cover all possible pathways involved, we highlight some of the key elements that likely contribute to the stress response and tolerance. Unless specifically indicated, all enzymes discussed are encoded by transcripts that are over-represented by at least 5-fold in the FGAS dataset.

Amino acid metabolism

Genes encoding proteins involved in primary metabolism pathways have been identified in the contigs with an over-representation of FGAS ESTs and cover several aspects of plant metabolic adjustments. Amino acid metabolism and the TCA cycle are the major pathways that generate precursors for various biological molecules. ESTs encoding several enzymes that are involved in the synthesis of arginine, cysteine, lysine, methionine, serine, phenyla-

lanine, proline and tryptophan are over-represented by more than 5-fold. These amino acids are precursors for the synthesis of several specialized metabolites. Two contigs encode the enzyme delta-1-pyrroline-5-carboxylate synthetase that is involved in proline biosynthesis, a metabolite that was found to increase during cold acclimation and drought stress [17]. Similarly, two contigs encode glutamate decarboxylase (GAD1), which is involved in the synthesis of gamma-aminobutyric acid (GABA), a non protein amino acid known to accumulate during cold acclimation and proposed to function in oxidative stress tolerance [18]. Several contigs encode enzymes involved in the metabolism of cysteine, an important precursor of glutathione involved in the modulation of oxidative stress. These include two different cysteine synthases and a putative O-acetylserine (thiol) synthase (OASTL). Overexpression of different isoforms of OASTL can increase thiol content in different transgenic plants and increase tolerance to abiotic stress such as exposure to elevated levels of cadmium [19].

Lipid metabolism

ESTs encoding different putative lipases and other proteins involved in lipid oxidation (acyl-CoA oxidase, MutT/nudix protein like, dihydrolipoamide acetyltransferase, b-keto acyl reductase, enoyl-ACP reductase, enoyl-CoA-hydratase, 3-hydroxyisobutyryl-coenzyme A hydrolase) are over-represented in the FGAS dataset while the acyl-carrier protein III involved in lipid synthesis is under-represented. These results suggest that lipid degradation occurs concomitantly with a reduction in the synthesis of short chain lipids. On the other hand, ESTs encoding enzymes involved in the synthesis of specialized lipids such as ATP citrate lyase α-subunit and the long chain fatty acid enzyme acetyl-CoA carboxylase are more abundant among FGAS ESTs. ESTs corresponding to several enzymes involved in sterol metabolism are also over-represented, suggesting major lipid modifications in membranes during cold acclimation. ESTs encoding three enzymes involved in the alternate pathway of isopentenyl pyrophosphate and squalene synthesis (1-deoxy-D-xylulose 5-phosphate reductoisomerase, 1-deoxy-D-xylulose-5-phosphate synthase, squalene synthase), three key enzymes of the sterol pathways (cycloartenol synthase, C14-sterol reductase (FACKEL), and 24-methylenelophenol methyltransferase) (Figure 3), and other enzymes such as sterol 4-alpha-methyl-oxidase, which can add to the variety of sterols produced, are also over-represented. The putative over-expression of several enzymes in the sterol pathway supports the previous observation of an increased production of membrane sterols [20]. These authors showed that the concentration of membrane sterols increases during cold acclimation and that this effect is more prominent in tolerant rye cultivars. Interestingly, sitosterol increases while campesterol decreases during acclimation, suggesting that the C24 methyltransferase that is putatively over-expressed in the FGAS dataset may be the SMT-2 transferase that diverts the methylenelophenol into the sitosterol pathway (see Figure 3; [21]). A search through the protein database has shown that the C24 methyltransferase has a much greater homology with SMT2 (7e-143) than with SMT1 (4e-63) supporting that the C24 methyltransferase is SMT2. The over-representation of FGAS ESTs in two contigs encoding stearoyl-acyl-carrier protein desaturase and two contigs encoding CDP-diacylglycerol synthase suggests that other important lipid modifying activities also occur in response to cold acclimation. Stearoyl-acyl-carrier protein desaturase is involved in the desaturation of existing lipids to form double bonds rendering the lipids more fluid at low temperature. This is an important adjustment associated with membrane stability at low temperature [20]. The over-expression of CDP-diacylglycerol synthase was previously shown to favour the synthesis of phosphatidylinositol [22]. In addition, one contig encodes a phosphoethanolamine N-methyltransferase. This enzyme is induced by low temperature and catalyzes the three sequential methylation steps to form phosphocholine, a key precursor of phosphatidylcholine and glycinebetaine in plants – metabolites known to be important in conferring tolerance to osmotic stresses such as low temperature, drought and salinity [23].

Secondary metabolism

Several contigs encode key enzymes involved in the biosynthesis of secondary metabolites such as phenylalanine ammonia lyase, cinnamyl alcohol dehydrogenase, and caffeoyl-CoA O-methyltransferase. Several enzymes are involved in the synthesis of methionine and its derivatives. The digital expression data suggest that the S-adenosylmethionine (SAM) cycle becomes more active during stress since contigs encoding three major enzymes of the cycle (S-adenosylmethionine synthetase, methionine S-methyltransferase, and S-adenosylhomocysteine hydrolase) are over-represented in FGAS. This pathway can provide SAM, the precursor molecule needed for nicotianamine biosynthesis. Four different contigs encoding nicotianamine synthase or nicotianamine aminotransferase are over-represented in FGAS. These enzymes are involved in nicotianamine and phytosiderophores synthesis and were found to be induced under iron deficiency [24,25]. The SAM cycle also provides the one carbon precursor for the methylation steps required for methyltransferase activities. At least 20 different contigs encoding methyltransferases contain ESTs that are over-represented in FGAS.

Transport activity

During cold acclimation, the cell mobilizes several transport systems to adapt to cold conditions. One of the major

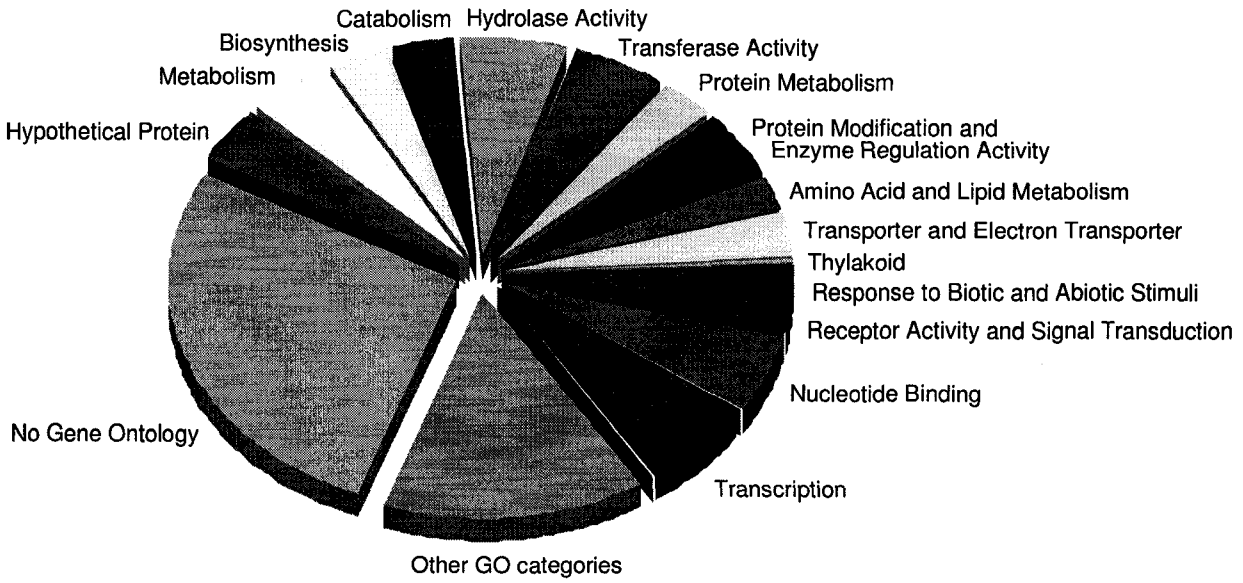

Figure 2

Functional classification of FGAS contigs containing ESTs that are over or under-represented more than 5-fold, or that contain more than 3 unique ESTs. The contigs belonging to the following GO terms were used: GO0008152 Metabolism; GO0009058 Biosynthesis; GO0009056 Catabolism; GO0016787 Hydrolase Activity; GO0016740 Transferase Activity; GO0019538 Protein Metabolism; GO0006464 and GO0030234 Protein Modification and Enzyme Regulator Activity; GO0006519 and GO0006629 Amino Acid and Lipid Metabolism; GO0005215 and GO0005489 Transporter and Electron Transporter Activity; GO0009579 Thylakoid; GO0009607 and GO 0009628 Response to Biotic and Abiotic Stimulus; GO0004872 and GO0007165 Receptor Activity and Signal Transduction; GO000166 Nucleotide Binding; Transcription Factors only from GO0006350 and GO0003677 (other DNA Binding Proteins were transferred to "Other GO categories"); a class was made for the mention "Hypothetical Protein" and for the mention "No Gene Ontology" while the "Other GO Categories" regroups several GO terms with small number of contigs.

effects of extracellular freezing is the reduced apoplastic water pressure and the rapid flow of water from the intracellular compartment to the apoplasm. Some of the consequences include the need for water and ion regulation as well as protection against dehydration. Two different contigs encoding aquaporins are highly abundant in FGAS (a contig with 12 ESTs found only in the FGAS dataset and a contig with ESTs over-represented 18-fold). These proteins likely play an important role in the regulation of the outward water flow. Similarly, several contigs associated with transport of ions or other small solutes are more highly represented, such as anion/sugar transporters, major facilitator superfamily antiporters, MATE efflux family transporters, nitrate transporters, cation exchangers, calcium and zinc transporters, betaine/proline transporters, and amino acid transporters. These different transporters are potential regulators controlling the flow of ions and other solutes that become more concentrated as water is drawn out of the cell during freezing. An interesting transporter activity is the phosphatidylinositol-phosphatidylcholine transfer protein which can contribute to the turnover of these lipids in the membrane. This pathway is involved in the accumulation of the compatible solute betaine that was reported to increase tolerance to drought and freezing [26]. Another mechanism involved in cell protection against higher ionic content include the replacement of water with compatible solutes such glycerol, glucose, sorbitol, proline and betaine. ESTs encoding hydroquinone glucosyltransferase, an interesting enzyme responsible for the synthesis of arbutin, are over-represented over 7-fold in the FGAS dataset. Glycosylated hydroquinone is very abundant in freezing and desiccation tolerant plants. It was suggested to accumulate up to 100 mM in the resurrection plant *Myrothamnus flabellifolia* and to increase membrane stability of artificial liposomes and thylakoids, possibly through the insertion of the phenol moiety in the phospholipid bilayer [27]. These authors showed that the lipid membrane composition is an important element for the cryoprotective effect of arbutin. In support of this observation, several contigs

with an over-representation of FGAS ESTs encoding transporters of compatible solutes and lipid modifying enzymes were identified.

Proteins involved in cryoprotection

One strategy that hardy plants such as wheat use to tolerate subzero temperatures is the accumulation of freezing tolerance associated proteins such as antifreeze proteins (AFPs) and dehydrins [28]. AFPs exhibit two related activities *in vitro*. The first is to increase the difference between the freezing and melting temperatures of aqueous solutions, a property known as thermal hysteresis. The second is ice recrystallization inhibition (IRI), where the growth of large ice crystals is inhibited, thus reducing the possibility of physical damage within frozen tissues [29]. In winter wheat and rye, several AFPs similar to pathogenesis-related proteins such as chitinases, glucanases, thaumatins and ice recrystallization inhibition proteins were identified [30-32]. Many contigs encoding chitinases, β-1,3-glucanases and thaumatin-like proteins contain ESTs that are over-represented in FGAS. Hincha *et al.* [33] reported that different cryoprotective proteins were able to protect thylakoids from freezing injury *in vitro*. Wheat ice recrystallization inhibition proteins are partly homologous to, and were annotated as, phytosulfokine receptors and were present in several contigs containing ESTs over-expressed in FGAS.

The dehydrins are hydrophilic proteins resistant to heat denaturation composed largely of repeated amino acid sequence motifs. They possess regions capable of forming an amphipathic α-helix. These properties may enable them to protect cells against freezing damage by stabilizing proteins and membranes during conditions of dehydration [28]. The most studied dehydrins are the WCS120 family, the WCOR410 and the chloroplastic WCS19 dehydrins. Genes encoding these proteins are highly over-represented in the FGAS dataset (Table 4, Table 5, and see additional file 1: Table 1.xls).

Photosynthesis

During cold acclimation, the chloroplast continues to receive as much light as at normal temperature but its thermal biochemical reactions are reduced. This results in an excess of light energy whereby electrons accumulate mostly in Q_A [34]. The reduced capacity to transfer electrons through PSII requires metabolic adjustments on a short term basis through redox balance, and communication between the chloroplast and the nucleus to modify gene expression for adaptation on a longer term basis. Freezing tolerant plants were previously shown to better cope with photoinhibition than less tolerant cultivars [34]. Although the number of genes classified under the GO "Thylakoids" is only 13, the genes identified indicate that putative changes in expression occur for genes encoding components of both the photosystem I (PSI) and the photosystem II (PSII). Several studies have reported changes in PSII during cold acclimation [34], The D1 and D2 proteins were shown to be sensitive to excess energy and to turn over more rapidly at low temperature and high light [35]. ESTs encoding the D2 protein are over-expressed by 7.2-fold in FGAS suggesting that the PSII adapts to low temperature conditions. On the other hand, the transcript encoding PSII Z is less represented in FGAS. A reduced amount of this protein may lead to a reduction in active antennas and allow a reduction in electron flow towards the PSII. ESTs encoding two other proteins of the PSII complex are over-represented (29.8 kDa and 20 kDa protein). These proteins belong to the same PsbP protein family which has 4 members in *Arabidopsis*. Recent results using RNAi have shown that this lumen protein is both essential and quantitatively related to PSII efficiency and stability. This suggests that their over-expression could improve electron flow through PSII [36,37]. Another limiting factor in the electron flow is the availability of CO_2. Several contigs with over-represented ESTs in the FGAS dataset encode carbonic anhydrase (carbonic anhydrase chloroplast precursor, dioscorin class A and nectarin III). This enzyme is known in C4 plants to concentrate CO_2 at its site of fixation. In the C3 plant wheat, this enzyme was previously shown to be modulated by nitrogen deficiency to maintain optimal CO_2 concentrations [38]. The over-expression of this enzyme could thus help to efficiently use the CO_2 and available light energy at low temperature. Failure to dissipate excess light energy could lead to oxidative stress, which needs to be controlled. A contig encoding a putative serine hydroxymethyltransferase is over-represented in the FGAS dataset. Hydroxymethyltransferases play a critical role in controlling the cell damage caused by abiotic stresses such as high light and salt, supporting the notion that photorespiration forms part of the dissipatory mechanisms of plants to minimize production of reactive oxygen species (ROS) in the chloroplast and to mitigate oxidative damage [39].

Very few studies have documented the modulation of PSI under stress conditions. The excess light or low temperature can decrease stromal NADP/NADPH ratio and it has been proposed that the cytochrome b6f complex can be regulated by the stromal redox potential possibly via a thioredoxin mediated mechanism (see [40]). The PSI components are largely integrated and composed of many subunits making it energetically expensive for the cell to produce. It has been suggested that cells might modulate PSI activity by varying the amount of the small and mobile plastocyanin protein carrying the reducing power [41]. The over-representation of ESTs encoding this protein in FGAS (represented by 27 ESTs within contig CL187Contig5) suggests that this PSI electron relay component becomes more active during cold acclimation and

Table 4: Contigs containing ESTs that are over-represented over 20-fold in the FGAS dataset.

Contig name	Annotation	Fold representation (FGAS/NSF-DuPont)	Reference
CL91Contig4	No Gene Ontology Hit (Wcor413, manual annotation)	163.30	[59]
CL206Contig4	Low molecular mass early light-inducible protein HV90, chloroplast precursor (ELIP)	94.35	[60]
CL386Contig5	Chitinase (EC 3.2.1.14)	68.94	[31]
CL1959Contig1	Legumin-like protein	68.94	[61]
CL117Contig7	No Gene Ontology Hit (Lea/Rab, manual annotation)	61.69	[62,63]
CL10Contig25	Defensin precursor	54.43	[64]
CL347Contig1	COR39 (WCS120 homolog, manual annotation)	52.61	[65]
CL158Contig8	Putative 1-aminocyclopropane-1-carboxylate oxidase	47.17	
CL386Contig1	Chitinase 1	43.54	[31]
CL347Contig2	Cold shock protein CS66 (Wcs120 homolog, manual annotation)	43.54	[65]
CL756Contig2	Hypothetical protein 259116.2b (LEA homolog, manual annotation)	43.54	[66]
CL1620Contig2	No Gene Ontology Hit	32.66	
CL411Contig1	Putative phytosulfokine receptor (Wheat Ice recristallization inhibitor, manual annotation)	32.66	[32]
CL349Contig4	Ferredoxin-NADP(H) oxidoreductase	32.66	[45]
CL1918Contig1	Glycosyltransferase	32.66	[16]
CL2Contig21	Hypothetical protein (Fragment) (Cab binding protein, manual annotation)	32.66	Genbank U73218
CL756Contig3	No Gene Ontology Hit	29.03	
CL2Contig9	Hypothetical protein (Fragment) (Cab binding protein, manual annotation)	29.03	Genbank U73218
CL3270Contig2	No Gene Ontology Hit	29.03	
CL28Contig11	Extracellular invertase (EC 3.2.1.26)	29.03	[67]
CL650Contig2	Cold acclimation protein WCS19	26.30	[68]
CL1442Contig1	Putative major facilitator superfamily antiporter	25.40	
CL1698Contig3	No Gene Ontology Hit	25.40	
CL704Contig4	Legumin-like protein	25.40	[61]
CL4965Contig1	Hypothetical protein P0508B05.10	25.40	
CL4930Contig1	ATP-dependent RNA helicase	25.40	[69]
CL411Contig4	No Gene Ontology Hit (Wheat Ice recristallization inhibitor, manual annotation)	25.40	[32]
CL2910Contig2	CONSTANS-like protein CO6	25.40	
CL117Contig3	No Gene Ontology Hit (Lea/Rab)	25.40	[63]
CL4699Contig1	Cytochrome P450	25.40	
CL4567Contig1	No Gene Ontology Hit	25.40	
CL1631Contig3	Beta-1,3-glucanase	25.40	[33]
CL411Contig3	Putative phytosulfokine receptor (Wheat Ice recristallization inhibitor, manual annotation)	25.40	[32]
CL91Contig8	No Gene Ontology Hit (COR413, manual annotation)	25.40	[59]
CL2020Contig1	No Gene Ontology Hit	23.58	
CL1106Contig2	Putative cytochrome c oxidoreductase	23.58	[70]
CL280Contig5	No Gene Ontology Hit (blt14, manual annotation)	23.58	[71]
CL1911Contig2	Putative cysteine proteinase inhibitor	21.77	[72]
CL3036Contig1	No Gene Ontology Hit hypothetical protein (OSJNBa0062C05.24, manual annotation),	21.77	
CL171Contig6	No Gene Ontology Hit	21.77	
CL202Contig14	No Gene Ontology Hit	21.77	
CL2484Contig2	No Gene Ontology Hit (putative F-Box family, manual annotation)	21.77	
CL3205Contig2	Hypothetical protein At2g43940	21.77	
CL117Contig2	No Gene Ontology Hit (Lea/Rab)	21.77	[63]
CL2663Contig3	Serine carboxypeptidase I precursor (EC 3.4.16.5) (Carboxypeptidase C) (CP-MI)	21.77	
CL4989Contig1	No Gene Ontology Hit	21.77	
CL437Contig6	Putative family II lipase EXL4	21.77	
CL2012Contig3	CIPK-like protein 1 (EC 2.7.1.37) (OsCK1)	21.77	[73]
CL1442Contig3	Putative major facilitator superfamily antiporter (sugar transporter family, manual annotation)	21.77	
CL3511Contig1	Similarity to receptor protein kinase (leucine rich protein similar to TIR1, manual annotation)	21.77	[74]
CL861Contig1	No Gene Ontology Hit	21.77	
CL4814Contig1	Putative cinnamyl alcohol dehydrogenase	21.77	
CL2Contig49	Chlorophyll a/b-binding protein WCAB precursor	21.77	Genbank U73218

Table 4: Contigs containing ESTs that are over-represented over 20-fold in the FGAS dataset. *(Continued)*

CL4798Contig1	No Gene Ontology Hit	21.77	
CL1740Contig2	Hypothetical protein OSJNBa0086E02.13 (Hypothetical protein P0419C04.2) (putative haloacid dehalogenase-like hydrolase, manual annotation)	21.77	
CL4476Contig1	No Gene Ontology Hit (phosphate induced protein, manual annotation)	21.77	
CL4337Contig1	Putative o-methyltransferase	21.77	[75]
CL2623Contig1	No Gene Ontology Hit (lumenal protein subunit of photosystem II, manual annotation)	21.77	
CL3656Contig2	Barwin	21.77	
CL671Contig1	No Gene Ontology Hit	21.77	
CL878Contig3	Putative pollen allergen Jun o 4	21.77	
CL26Contig8	No Gene Ontology Hit	0.050	
CL350Contig1	Photosystem II reaction center Z protein	0.040	
CL185Contig1	Chloroplast 50S ribosomal protein L14	0.037	
CL120Contig2	Lipid transfer protein 1 precursor	0.030	
CL144Contig2	Alpha amylase inhibitor protein	0.026	

may be important in relieving the pressure caused by electrons accumulating in Q_B. The mobile plastocyanin molecule is a limiting factor in the electron transfer from PSII to PSI. The increased expression of plastocyanin may result in an increased activity of PSI under low temperature and may help freezing tolerant plants maintain their energy balance compared to less tolerant plants. We have previously shown that several proteins involved in improving photosynthesis, including plastocyanin, are expressed at low levels under low excitation pressure (20°C/50 μE) but markedly accumulate when transferred to 5°C under the same light regime [42]. A mutation in the PSI-E subunit was also shown to have a great impact on PSII as it becomes easily affected by photoinhibition even under low light [43]. Similarly mutants in the PSI-N subunit, which participates in the docking of PC, are impaired in PSI activity [44]. The over-representation of ESTs encoding the PSI-E and PSI-N subunits in the FGAS dataset could thus provide an integrated response to reduce photoinhibition. In order to maintain a proper NADP/NADPH ratio, the malate valve could be activated to transfer excess reducing power to the cytoplasm [45]. ESTs encoding two PSI components are less abundant in FGAS. One of these is a subunit of the chloroplastic NADH dehydrogenase equivalent to the mitochondrial enzyme. Interestingly, the *FRO1* gene was recently shown to encode the mitochondrial NADH dehydrogenase counterpart which plays a role in controlling ROS and the ability of *Arabidopsis* to respond to low temperature [46]. An excess of ROS in mitochondria was proposed to affect the induction of CBF transcription factors and cold acclimation. The chloroplastic NADH dehydrogenase may also affect the ability to induce CBF if the ROS that accumulate during photoinhibition at low temperature are not detoxified. Tolerant plants may adapt their photosystems to avoid the accumulation of ROS in chloroplasts, thus allowing a strong CBF response and a stable induction of downstream cold-regulated genes. This hypothesis may explain why tolerant plants are able to maintain a strong expression of several freezing tolerance-associated genes while less tolerant plants show transient, reduced expression of these genes at low temperature [1].

Signalling cascades and transcription factors

Among the contigs with an over-representation in FGAS ESTs, we identified several proteins involved in the synthesis or perception of different hormones. These include enzymes of the ethylene, auxin and jasmonic acid metabolism; brassinosteroid LRR receptor, receptor-like kinases CLAVATA2 and PERK1, and phytosulfokine receptor. Contigs encoding several proteins involved in signalling cascades were also found such as calcium binding proteins, diacylglycerol kinase, lipid phosphate phosphatase-2, inositol 1-monophosphatase, GTP-binding proteins, MAP kinases and MAPKK, serine/threonine kinase, CIPK-like protein-1, histidine kinase-2, and protein phosphatases 2A and 2C.

The potentially increased activity of the various signalling pathways is associated with a differential expression of many families of transcription factors (TF; Table 7). The results show that at least 220 contigs contain ESTs encoding TF that are over- or under-represented more than two-fold in the FGAS dataset. Using a more stringent cut-off excludes some TF that may not be strongly regulated, but should also reduce the number of false positives. With a 5-fold cut-off, 151 TF were identified, with 30 of them being contigs unique to FGAS. The most highly represented TF families are the zinc fingers, WRKY, AP2, Myb and NAC. Several members of these families were previously identified as being responsive to various stresses. The most studied members are those of the AP2 family, in particular the CBF/DREB subfamily. CBF members are involved in the cold/drought responses [47]. We have identified 3 different contigs, with a 5-fold over-representation in the FGAS dataset, that contain CBF-like binding factors and 5 unique FGAS contigs containing at least 3 ESTs (annotated as CBF-like, CBF1-like, CBF3-like, C-

Table 5: Contigs containing at least 5 ESTs that are unique to the FGAS dataset.

Contig name	Annotation	Number of ESTs	Reference
CL1638Contig1	No Gene Ontology Hit (no homology)	24	[76]
CL1293Contig2	Wheat cold acclimation protein Wcor80 (Wcs120 homolog, manual annotation)	19	[65]
CL386Contig3	Chitinase I	18	[31]
CL347Contig3	Cold acclimation protein WCS120 (manual annotation)	17	[65]
CL2466Contig1	Putative heat shock protein (E. Coli contaminant, manual annotation)	16	
CL3394Contig1	Nitrogen regulation protein NR(II) (EC 2.7.3.-) (E. coli contaminant, manual annotation)	12	
CL7Contig23	Aquaporin PIP1		[77]
CL40Contig14	Chitinase IV	11	[31]
CL650Contig3	Chloroplast-targeted COR protein (Wcor14c, manual annotation)		[76]
CL1239Contig3	Putative LMW heat shock protein	10	
CL2570Contig1	Hypothetical protein OJ1015F07.4		
CL125Contig7	O-methyltransferase	9	[75]
CL206Contig11	Low molecular mass early light-inducible protein HV90, chloroplast precursor (ELIP)		[60]
CL3635Contig1	No Gene Ontology Hit		
CL4047Contig1	ABA responsive protein mRNA (manual annotation)		[78]
CL52Contig12	No Gene Ontology Hit		
CL52Contig13	No Gene Ontology Hit		
CL619Contig5	WSI76 protein induced by water stress (galactinol synthase, manual annotation)		[79]
CL1228Contig3	Leaf senescence protein-like	8	
CL1293Contig1	Dehydrin (Wcs120 homolog, manual annotation)		[65]
CL2543Contig2	No Gene Ontology Hit		
CL400Contig4	Cysteine protease		[80]
CL4107Contig1	No Gene Ontology Hit		
CL4776Contig1	Probable arylsulfatase activating protein aslB (E. coli contaminant, manual annotation)		
CL1051Contig5	C repeat-binding factor 2	7	[81]
CL2204Contig1	No Gene Ontology Hit (Wheat Ice recristallization inhibitor, manual annotation)		[32]
CL3474Contig1	No Gene Ontology Hit		
CL3792Contig1	No Gene Ontology Hit		
CL4454Contig1	No Gene Ontology Hit		
CL5468Contig1	Ubiquinone/menaquinone biosynthesis methyltransferase ubiE (EC 2.1.1.-) (E. coli contaminant, manual annotation)		
CL833Contig4	Putative EREBP-like protein (putative AP2 domain transcription factor, manual annotation)		
CL1318Contig2	S-like Rnase	6	[82]
CL1368Contig4	Beta-expansin		
CL17Contig3	Type I non-specific lipid transfer protein precursor (Fragment)		[83]
CL20Contig27	No Gene Ontology Hit		
CL2425Contig2	Putative lectin		[84]
CL280Contig2	Low temperature responsive barley gene blt14 (manual annotation)		[62]
CL280Contig4	Cold regulated protein pao29 (similar to blt14 manual annotation)		[62]
CL2910Contig1	CONSTANS-like protein CO6		
CL3212Contig2	No Gene Ontology Hit		
CL3324Contig2	RING zinc finger protein-like		
CL3647Contig2	No Gene Ontology Hit		
CL3778Contig2	Putative phenylalanyl-tRNA synthetase alpha chain		
CL4292Contig1	C2H2 Zinc finger protein (manual annotation)		
CL4895Contig1	No Gene Ontology Hit		
CL5228Contig1	Putative inositol-(1,4,5) trisphosphate 3-kinase		
CL5712Contig1	Putative ABCF-type protein (anthocyanin transport)		
CL5985Contig1	Hypothetical protein P0508B05.10		
CL6056Contig1	Putative calcium binding EF-hand protein (caleosin: lipid body trafficking, manual annotation)		
CL6257Contig1	No Gene Ontology Hit		
CL6493Contig1	No Gene Ontology Hit		
CL861Contig2	No Gene Ontology Hit		
CL1051Contig2	C repeat-binding factor 2	5	[81]
CL1182Contig3	OSJNBa0043A12.18 protein (putative transcription factor)		
CL1279Contig2	Isoflavone reductase homolog (EC 1.3.1.-)		
CL1366Contig3	Putative UDP-glucose: flavonoid 7-O-glucosyltransferase		
CL206Contig6	High molecular mass early light-inducible protein HV58, chloroplast precursor (ELIP)		[60]
CL3647Contig1	No Gene Ontology Hit		

Table 5: Contigs containing at least 5 ESTs that are unique to the FGAS dataset. *(Continued)*

CL4058Contig1	Myb-related protein Hv33	
CL411Contig7	No Gene Ontology Hit (Wheat Ice recristallization inhibitor, manual annotation)	[32]
CL4350Contig2	Similarity to protein kinase	GenBank AY738149
CL4537Contig1	Putative ACT domain-containing protein	
CL4642Contig1	Chitinase 1	[31]
CL4666Contig1	Farnesylated protein 1	[85]
CL4825Contig1	Hypothetical protein P0473D02.6 (Hypothetical protein OJ1368_G08.21)	
CL6137Contig1	No Gene Ontology Hit	
CL6258Contig1	Putative sodium-dicarboxylate cotransporter	
CL6567Contig1	Putative arabinogalactan protein	
CL6634Contig1	No Gene Ontology Hit	
CL6741Contig1	Putative b-keto acyl reductase (fatty acid elongase, waxes biosyntheisis)	
CL6821Contig1	Putative strictosidine synthase (alkaloid biosynthesis)	
CL7090Contig1	No Gene Ontology Hit	
CL721Contig3	No Gene Ontology Hit	
CL7241Contig1	No Gene Ontology Hit	
CL7243Contig1	No Gene Ontology Hit	
CL7272Contig1	Early light-inducible protein	[60]
CL7415Contig1	No Gene Ontology Hit	
CL7455Contig1	ABC1 family protein-like	
CL754Contig3	Chitinase 3	[31]
CL7581Contig1	Aspartate transaminase, mitochondrial	
CL7608Contig1	Putative aspartic proteinase nepenthesin 1	
CL7617Contig1	No Gene Ontology Hit (barley Blt14 homolog, manual annotation)	[62]
CL7686Contig1	No Gene Ontology Hit	
CL7701Contig1	Putative FH protein interacting protein FIP2 (potassium channel tetramerization)	
CL7785Contig1	No Gene Ontology Hit	
CL7794Contig1	No Gene Ontology Hit	
CL807Contig3	Putative diphosphonucleotide phosphatase (calcineurin-like phosphoesterase)	
CL861Contig5	No Gene Ontology Hit	
CL963Contig4	OSJNBb0013O03.11 protein (bHLH transcription factor, manual annotation)	

repeat binding factor 3-like, C-repeat/DRE binding factor 3, CRT/DRE binding factor 2, DRE binding factor-2). Expression profiling using qRT-PCR has confirmed that transcripts corresponding to 7 of the 8 contigs are over-expressed at specific time points during cold acclimation (Sarhan *et al.* unpublished results). Expression of the CBF genes in *Arabidopsis* was shown to be regulated by members of the bHLH family [48]. We have identified 7 contigs encoding bHLH members that are over-represented by two-fold, with two of them being over-represented more than 5-fold (Table 7). However, the genes encoding the bHLH ICE proteins in *Arabidopsis* are not cold-induced. Although the expression pattern with regards to cold inducibility of the ICE genes could be different between wheat and *Arabidopsis*, the isolation of the full length genes, phylogenetic analysis and expression studies are required to determine if any of the over-represented bHLH encode ICE homologs. In addition to the CBFs and bHLH families, several other TF families may be part of other stress components associated with abiotic stress such as drought, salinity, oxidative, etc. Interestingly, several genes that control flowering have also been identified (FLT, Gigantea, MADS, CO, Aintegumenta). These genes are most likely associated with the vernalization response in wheat as was recently shown for *TaVRT1* and *TaVRT2* [49,50].

Conclusion

The large number of ESTs annotated from FGAS and NSF-DuPont datasets represents an important resource for the wheat community. Digital expression analyses of these datasets provide an overview of metabolic changes and specific pathways that are regulated under stress conditions in wheat and other cereals. The information generated will help construct network models of abiotic stress responses that will facilitate computational predictions and direct future experimental work like the development of models such as the "Metabolic pathways of the diseased potato" [51] or MapMan for the analysis of gene expression data in *Arabidopsis* [52]. The results could facilitate the understanding of cellular mechanisms involving groups of gene products that act in coordination in response to environmental stimuli.

Methods

A total of eleven different cDNA libraries were prepared from hexaploid wheat (*Triticum aestivum*) for the FGAS EST sequencing project and are summarized in Table 1.

Cultivar Norstar was used for Libraries 2 to 6 to represent various tissues, developmental stages and stress conditions. Six subtracted cDNA libraries (suppression subtractive hybridization; SSH), named TaLT2 to TaLT7, were also prepared from two different wheat lines (CI14106 and PI178383) and cv Norstar as a complementary approach to isolate differentially expressed transcripts. The "Library 1" and TaLT1 libraries were not used for the large scale EST sequencing FGAS project since the former was not prepared in a Gateway-compatible vector and the latter was generated to optimize the SSH protocol.

Preparation of the cDNA libraries

Growth conditions

For Libraries 2 and 3, the seeds were germinated in water-saturated vermiculite for 7 days at 20°C and 70% relative humidity under an irradiance of 200 μmol m^{-2} sec^{-1} and a 15-hr photoperiod. At the end of this period, the aerial parts (crowns and leaves) and roots of control plants were sampled and individually frozen. Cold acclimation was performed by subjecting germinated seedlings to a temperature of 4°C with a 12-hr photoperiod for 1, 23 and 53 days under an irradiance of 200 μmol m^{-2} sec^{-1}. Seedlings were watered with a nutrient solution (0.5 g/l 20:20:20; N:P:K). Salt stress was induced by watering with the nutrient solution containing 200 mM NaCl for 0.5, 3 and 6-hr. Aerial parts of cold-acclimated plants were sampled for Library 2 and roots of both cold-acclimated and salt-stressed plants were sampled for Library 3.

For Library 4, two different water stress conditions were used. For bench experiments, seeds were germinated for 7 days as described for Library 2. At the end of this period, plants were removed from vermiculite and left at room temperature on the table without water for 1, 2, 3 and 4 days before sampling. For growth chamber experiments, seeds were germinated in a water-saturated potting mix (50% black earth and 50% ProMix) for 7 days under an irradiance of 200 μmol m^{-2} sec^{-1}. The temperature was maintained at 20°C with a 15-hr photoperiod under a relative humidity of 70%. After this period, watering of plants was stopped. Four time points were sampled during a two weeks period; the first after wilting was observed and the last, two weeks later, and consisted of living crown and stem tissues (leaf tissue was yellow and thus not included in the sampled material).

For Library 5, seeds were germinated for 7 days and cold-treated for 49 days (full vernalization) as described for Library 2. Seedlings were then potted in water-saturated potting mix and transferred to flower inducing conditions (20°C and a 15-hr photoperiod). Tissues were sampled as follows: 1 cm crown sections after 30 days of cold treatment; 1 cm vernalized (49-day cold-treated) crown sections that were exposed to flower inducing conditions for 11 days; different developmental stages of spike formation (5 to 50 mm); and different developmental stages of spike and seed formation after the spikes had emerged from the flag leaf (visible).

For Library 6, seeds were germinated for 7 days and cold-treated as described for Library 2, except that cold treatments were performed for short time points (1, 3 and 6 hr) in the light or in the dark. Crown sections (1 cm) and green leaf tissues were harvested individually for each time point and for both exposure conditions.

For SSH libraries TaLT2 to TaLT7, plants were germinated as described for Library 2 except that the light intensity was 275 μm m^{-2} s^{-1} and the cold treatment was performed at 2°C for 1, 21 or 49 days. Crown sections (1 cm) were harvested individually for each time point.

RNA purification and cDNA synthesis

For Libraries 2 and 3, total RNA was isolated using the phenol method [53] except that the heating step at 60°C was omitted, whereas the TRI Reagent method (Sigma) was used for Libraries 4 to 6 and TRIzol (Life Technologies) was used for the TaLT libraries. For Libraries 2 to 6, poly(A)$^+$ RNA was purified from the total RNA samples using two cycles of an oligo(dT)-cellulose affinity batch-enrichment procedure [53] whereas PolyA Pure (Ambion) was used for the TaLT libraries. Total RNAs were subsequently used for cDNA synthesis. For all libraries, cDNA synthesis was initiated with a *NotI* primer-adaptor ($GCGGCCGCCCT_{15}$) using the 'SuperScript™ Plasmid System with Gateway Technology for cDNA Synthesis and Cloning' kit (Invitrogen). For Libraries 3 to 6, methylated dCTP was added to the first strand reaction mix to prevent cleavage by the NotI restriction enzyme used for directional cloning. For Library 6, the 'GeneRacer' kit (Invitrogen) was used prior to first strand synthesis to dephosphorylate truncated and non-mRNAs, remove the 5' cap structure from intact mRNA, and ligate the gene racer RNA oligo 5'-CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3'. The precipitation steps in the kit were replaced by the RNeasy Mini Protocol for RNA Cleanup (QIAGEN). For this library, the second strand cDNA was synthesized using *Pfx* DNA polymerase (Invitrogen) and the primer 5'-CGACTGGAGCACGAGGACACTGA-3' homologous to the RNA oligo. The 'SuperScript™ Plasmid System with Gateway Technology for cDNA Synthesis and Cloning' kit (Invitrogen) was used for the remaining steps of the construction of Libraries 2 to 6 except that the precipitation steps without yeast carrier tRNA were replaced by the QIAquick PCR purification procedure (QIAGEN). For the TaLT2, 3, 6 and 7 libraries, the Nitro-pyrrole anchored oligo-dT priming technique was used [54]. For TaLT4 and TaLT5 libraries, the SMART cDNA (Clontech) priming kit was used.

Table 6: Annotated contigs that are unique to the TaLT libraries (SSH).

Contig name	Annotation	Number of ESTs	Contigs with similar annotation containing ESTs over-represented in FGAS
CL1246Contig2	Putative high-affinity potassium transporter	29	
CL1122Contig2	Putative phosphoribosylanthranilate transferase	27	7-fold 7e-53 CL10525Contig1
CL1701Contig1	Potential phospholipid-translocating ATPase	23	
CL1961Contig1	Transcriptional factor B3-like	20	
CL1506Contig2	DHHC-type zinc finger domain-containing protein-like	19	
CL2126Contig1	Putative ACT domain-containing protein	19	
CL622Contig3	50S ribosomal protein L22-like	19	
CL2193Contig1	Putative DEAD/DEAH box RNA helicase protein	17	
CL1038Contig2	Pollen-specific calmodulin-binding protein	16	
CL3163Contig1	ATP synthase protein 9, mitochondrial precursor (EC 3.6.3.14) (Lipid-binding protein)	12	
CL3186Contig1	Putative pollen specific protein (Putative ascorbate oxidase)	12	
CL1986Contig1	Putative dCK/dGK-like deoxyribonucleoside kinase	10	
CL3856Contig1	Protein kinase domain	10	
CL2813Contig3	MKIAA0124 protein (Fragment)	9	
CL4654Contig1	Hypothetical protein OSJNBa0088106.19	8	
CL4703Contig1	40S ribosomal protein S7	8	
CL4937Contig1	Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12)	8	
CL1038Contig3	Hypothetical protein AT4g28600	7	
CL4812Contig1	Homeobox transcription factor-like	7	
CL4821Contig1	Agglutinin isolectin 3 precursor (WGA3) (Fragment)	7	
CL4846Contig1	Putative aldo/keto reductase family protein	7	
CL10Contig35	Ribosomal protein L10A	6	
CL4Contig25	Phytochrome B (Fragment)	6	
CL5821Contig1	Putative very-long-chain fatty acid condensing enzyme CUT1	6	7-fold 2e-57 CL5480Contig1
CL5833Contig1	Putative UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase-I	6	
CL6515Contig1	NBS-LRR disease resistance protein homologue	6	
CL823Contig3	Putative RNA splicing protein	6	
CL1392Contig2	Heat shock factor-binding protein 1	5	
CL4432Contig2'	Putative chromomethylase	5	
CL5300Contig2	Hypothetical protein	5	
CL6924Contig1	Beta-expansin (Fragment)	5	7-fold 8e-48 CL235Contig6
CL6960Contig1	Hypothetical protein OSJNBb0027B08.22 (Hypothetical protein OSJNBa0078D06.5)	5	
CL7305Contig1	Agglutinin (CCA)	5	
CL7698Contig1	Putative resistance gene analog PIC27	5	
CL1101Contig4	Putative amino acid transporter	4	
CL1531Contig2	Putative ZIP-like zinc transporter	4	
CL1739Contig3	Putative ethylene-responsive small GTP-binding protein	4	
CL18Contig7	Putative ribosomal protein L5	4	
CL2037Contig3	Protoporphyrin IX Mg-chelatase subunit precursor	4	
CL2221Contig1	Putative Ribosome recycling factor, chloroplast	4	
CL2305Contig1	Eukaryotic translation initiation factor 3 subunit 12 (eIF-3 p25) (eIF3k)	4	
CL3669Contig2	Putative ascorbate oxidase promoter-binding protein AOBP	4	
CL36Contig7	Adenosylhomocysteinase-like protein	4	
CL3840Contig2	Putative aminopropyl transferase	4	
CL6158Contig2	Cytochrome C6, chloroplast-like protein	4	
CL7225Contig1	P0076O17.10 protein	4	
CL732Contig2	OSJNBa0070C17.10 protein	4	
CL7697Contig1	Heat shock factor protein hsf8-like	4	
CL8407Contig1	Aldo/keto reductase family-like protein	4	11-fold 3e-54 CL3996Contig1
CL9543Contig1	Anthranilate N-benzoyltransferase-like protein (AT5g01210/F7J8_190)	4	
CL10751Contig1	Histone H4-like protein	3	7-fold 6e-46 CL9Contig66
CL10863Contig1	Methionine S-methyltransferase (EC 2.1.1.12) (AdoMet:Met S-methyltransferase)	3	
CL11049Contig1	Transferase family	3	
CL12283Contig1	Putative PPR-repeat containing protein	3	

Table 6: Annotated contigs that are unique to the TaLT libraries (SSH). *(Continued)*

CL12337Contig1	U3 small nucleolar RNA-associated protein 14 (U3 snoRNA-associated protein 14)	3	
CL12711Contig1	Putative lipase/acylhydrolase (Putative anther-specific proline-rich protein)	3	
CL1347Contig2	Omega-3 fatty acid desaturase	3	
CL1402Contig2	Putative VIP2 protein	3	
CL1688Contig3	Putative plastid ribosomal protein L11	3	
CL1Contig342	Protein H2A	3	15-fold 8e-74 CL1Contig113
CL1Contig350	Protein H2A	3	15-fold 2e-47 CL1Contig113
CL1Contig361	60S ribosomal protein L17-1	3	
CL2045Contig1	Cap-binding protein CBP20	3	
CL2470Contig2	Putative inorganic pyrophosphatase	3	7-fold 1e-75 CL2470Contig1
CL2890Contig3	Mak3 protein-like protein	3	7-fold 4e-91 CL2890Contig1
CL3033Contig2	Putative serine/threonine phosphatase	3	
CL3124Contig2	Putative ATP phosphoribosyl transferase	3	
CL4048Contig2	Boron transporter	3	
CL4808Contig2	Putative DNA topoisomerase II	3	
CL617Contig3	Putative calreticulin	3	5-fold 9e-152 CL617Contig1
CL7904Contig1	Hypothetical protein OSJNBb0004M10.19	3	
CL9749Contig1	Putative subtilisin-like proteinase	3	9-fold 3e-20 CL5317Contig1
CL9993Contig1	Hypothetical protein At1g78915	3	
CL4836Contig2	MtN3-like	2	

Suppression Subtractive Hybridization

For the TaLT libraries, SSH was performed on the RNAs isolated from crowns. For the TaLT2 library, RNA from CI14106 cold-acclimated for 1 day was used as tester RNA and subtracted by SSH against the driver RNA from cv Norstar cold-acclimated for 21 and 49 days (equal amounts of cDNAs were pooled together before subtraction). For TaLT3, 21 and 49-day cold-acclimated CI14106 was subtracted against cv Norstar cold-acclimated for 1 day. For TaLT4, 1 day cold-acclimated PI178383 was subtracted against 21 and 49 days cold-acclimated cv Norstar. For TaLT5, 21 and 49 days cold-acclimated PI178383 was subtracted against 1 day cold-acclimated Norstar. For TaLT6, 1 day cold-acclimated CI14106 was subtracted against non-acclimated CI14106. For TaLT7, 21 and 49 days cold-acclimated CI14106 was subtracted against non-acclimated CI14106.

Cloning into vectors

For Libraries 2 to 6, a *Sal*I adaptor (GTCGAC-CCACGCGTCCG) was ligated to the 5' end of the cDNAs synthesized with the *Not*I primer-adaptor to allow for directional cloning. The first two (for Libraries 3 to 5) or five (for Libraries 2 and 6) fractions eluting from size fractionation column chromatography and containing cDNAs larger than 0.5 kb were pooled for ligation with the vector. About 15 ng of *Sal*I-*Not*I-digested cDNAs was ligated with 50 ng of the pCMV.SPORT6 vector, which contains the attB1 and attB2 site-specific recombination sites flanking the multiple cloning sites. Therefore, clones isolated from these libraries can be rapidly transferred into Gateway™ destination vectors using site-specific recombination (Invitrogen). The libraries were then transformed into ElectroMAX™ DH10B cells (Invitrogen) for Library 2 or ElectroTen-Blue™ cells (Stratagene) for Libraries 3 to 6. For TaLT libraries, the PCR-amplified products of SSH were non-directionally cloned into the pGEM-T vector and transformed into DH5α cells.

Assessment of library quality and selection of clones for sequencing

Around 6.0×10^6 primary clones were obtained for Libraries 2 to 6. To determine the average cDNA size, 96 clones were randomly chosen from different libraries and the plasmids digested and characterized on agarose gels. Average insert sizes were estimated at 1300 bp (Library 2: 14% of inserts below 750 bp, 59% between 750 and 1500 bp, and 27% above 1500 bp), 1560 bp (Library 3: 10% below 750 bp, 44% between 750 and 1500 bp, and 46% above 1500 bp), and 1100 bp (Library 6: 17% below 750 bp, 68% between 750 and 1500 bp, and 15% above 1500 bp). Since all libraries contain an average of 6 million different clones, this collection represents an important resource to isolate full length clones for which only truncated cDNAs are available. To reduce the number of ESTs representing highly expressed genes, Libraries 2 to 6 were hybridized to ^{32}P-labelled cDNAs from non-acclimated plants. Colonies showing with the weakest hybridization signals were picked for sequencing.

Bioinformatics

Trimming high quality sequences

Sequence tracefiles were obtained from the FGAS project (110,544 ESTs) and from the NSF (82,332 ESTs; [55]) and DuPont (154,171 ESTs) collections. The latter two collections comprise EST sequences derived from many cDNA libraries prepared from various wheat RNA sources. All sequences were processed as follows. Quality score

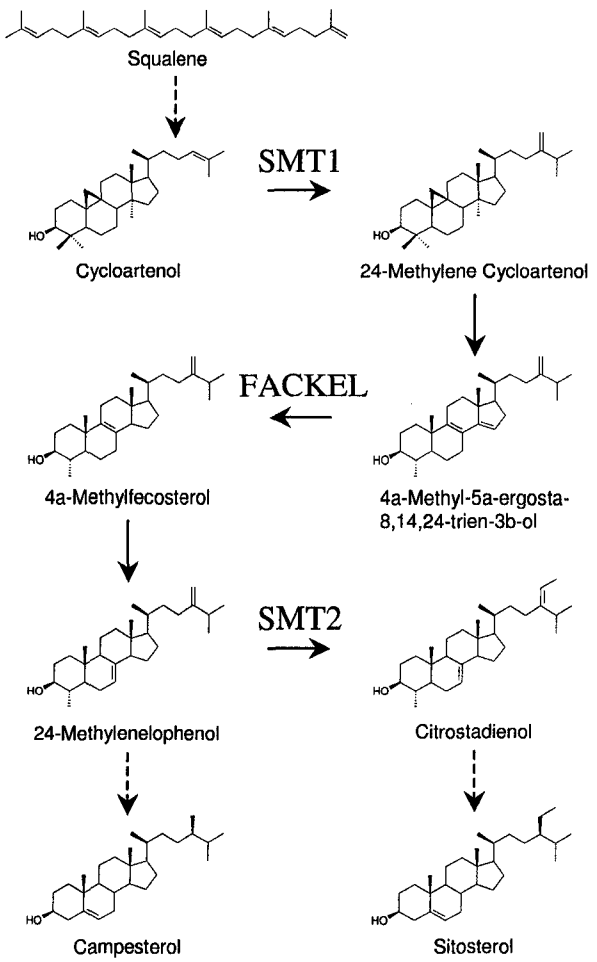

Figure 3
Plant sterols pathway. ESTs encoding several enzymes of the sterol pathways are over-represented in the FGAS dataset. Three enzymes are involved in the production of squalene from which cycloarthenol is obtained. The FACKLE and SMT2 enzymes are involved in the production of sitosterol with a concomitant decrease in campesterol.

sequences were obtained from tracefiles using PHRED [56,57]. Only sequences with mean Q≥20 were retained. Poly(A) or poly(T) regions with length = 14 (± 2 errors) were trimmed and all sequences containing more than one poly(A) and/or poly(T) sequences were flagged as putative chimeras. SeqClean3 with generic Univec DB as well as Lucy4 (using pCMV.SPORT6 and pBlueScript II splice sites) were used with the default settings in an iterative manner. This recursive approach proved more efficient in removing vector and linker sequences, and low quality regions than using either one only once. All resulting high quality sequences were then re-checked for low-complexity and all sequences containing more that 50% repeats were rejected. A repeat was defined as a minimum word size of 4 identical bases with a maximum of 1 error. RepeatMasker2 was used with Repeat DB to mask regions that could eventually bias the assembly. All information pertaining to library details, sequences and data quality scores were stored in a mySQL database. After filtering, 269,562 cleaned ESTs were retained for assembly (73,521 ESTs from FGAS, 68,886 ESTs from NSF and 127,155 ESTs from DuPont).

Clustering, assembly and annotation

Clustering was performed to reduce the redundancy of the dataset and increase the overall quality of the derived consensus sequences. When a small set of sequences (FGAS 73,521 quality-filtered sequences) was used, the clustering performed well through TGICL and d2_cluster. However, when the NSF and DuPont data (196,041 sequences) were added, aberrant large clusters were obtained. This is presumably due to undetected chimeras, multi-domain proteins and the transitive closure technique applied by these applications. These large clusters (38 k sequences for TGICL and 25 k for d2_cluster) contained many unrelated sequences and were difficult to assemble, yielding many incongruent and low quality contigs. To avoid such artifacts, a cluster breaking strategy was used. First, all sequences that could be contained in other ESTs were removed, thereby reducing the dataset to parent sequences. These sequences were then BLASTed against themselves and results were parsed to extract the e-values in order to build an adjacency matrix. The distance (d) between the sequences was calculated based on the level of similarity established using BLAST e-value where d = 100/-log (e-value). Two parent sequences were considered to be part of the same cluster when the BLASTN identity result between them was greater than or equal to 96%. GRAPH9 was used to flag bridges (articulation points where the removal of an EST breaks the link between subclusters) and manually split the large graph into distinct smaller sub-graphs. Other suspicious clusters that were not automatically detected were manually investigated and split when required (Figure 4a). Child ESTs, removed in the first stage were then incorporated into the cluster containing the parent sequence. For example, the largest cluster was broken down using the approach described above and yielded 250 sub-clusters, with the largest being of 6 k sequences (Figure 4b). TGICL and d2_cluster results were compared using randomly chosen clusters that were re-assembled using either clustering tools. It was observed that TGICL had a higher tendency of joining similar genes and falsely splitting sequences from the same gene, thus indicating that d2_cluster was a more reliable clustering tool in our case.

Both CAP3 [58] and PHRAP were tested to assemble the sequences. CAP3 was used on TGICL results using the set-

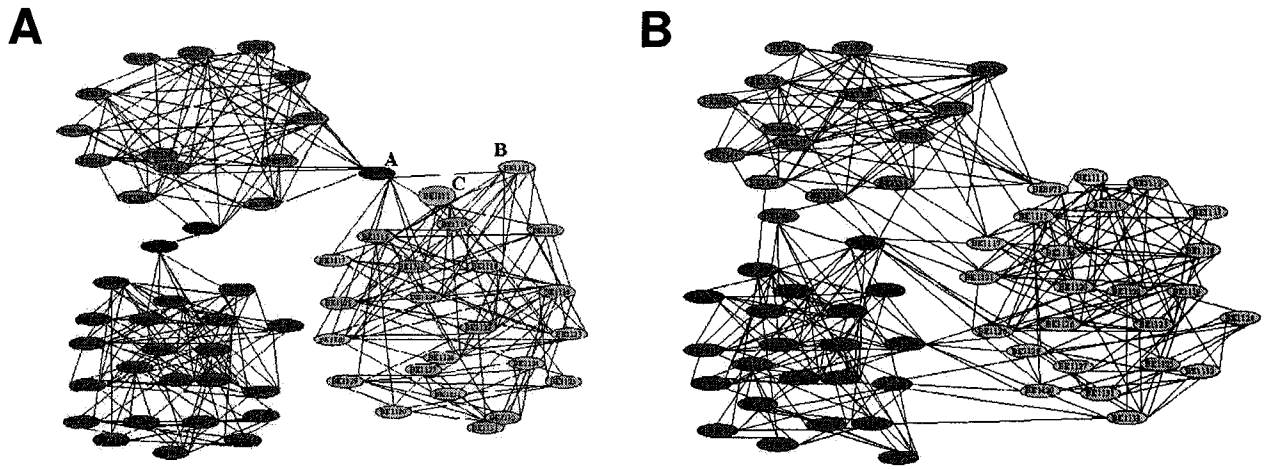

Figure 4
Breaking strategy of large clusters. A breaking strategy was used to reduce the size of large clusters. Each sequence in a cluster was BLASTed against the others and e-values were used to build an adjacency matrix (see Materials and Methods). For example, an e-100 value will result in a distance of 1 cm between two sequences. Only values below e-25 were used for graphical display. GRAPH9 was used to flag bridges (articulation points where an EST links two potential sub-clusters) and manually split a cluster into distinct sub-clusters. **A)** Example of a cluster region where specific ESTs (in red) can be manually transferred to sub-clusters (based on the smallest e-value). **B)** Example of a cluster region that could not be broken into sub-clusters due to the complex interrelations between ESTs.

tings that appeared satisfactory when assembling barley EST sequences [7] while PHRAP was used to assemble d2_cluster results using the default parameters. The first method generated ~32 k contigs while the latter produced over 50 k contigs. The first approach gave results more consistent with the Unigene and TIGR Wheat Gene Index assembly data with respect to contig number, suggesting that PHRAP was less appropriate for assembly of the large dataset used in this study. The total number of singletons and singlets in both cases was similar; 39 k for PHRAP (14% of all ESTs) vs. 42 k for CAP3 (15.5% of all ESTs) and the percentage was close to that found in TIGR (13.3% of all ESTs). Singletons are defined as unique sequences that could not be assembled in a cluster whereas singlets are unique sequences that were assembled in a cluster but could not be assembled in a contig. Based on the TGICL and d2_cluster comparison and on the number of contigs obtained with CAP3 and PHRAP, we chose d2_cluster and CAP3 as the clustering and assembly tools for this project.

We used different annotation tools to increase the number of annotated sequences. The unique assembled sequences produced in our study were annotated after translation using prot4EST and then BLASTed (BLASTX) against a GO-annotated database. All the sequences that did not show sufficient similarity to be functionally classified with this method were investigated with AutoFact where sequences are BLASTed against other complementary databases (ex. PFAM, KEGG, Ribosomal Sequences database) having GO details.

Digital expression analysis

The relative abundance (digital expression) of FGAS ESTs was analysed as follows: 1) among the contigs containing EST sequences present in both the FGAS dataset and NSF-DuPont dataset, abundance was expressed as a ratio of FGAS ESTs (without SSH ESTs) to NSF-DuPont ESTs, after correction for the size (total number of ESTs) in each dataset; 2) contigs that contained only FGAS ESTs were analyzed separately; 3) SSH EST abundance was compared between similar SSH libraries to determine if common ESTs can be identified; and 4) unique SSH contigs were identified as these could represent new genes expressed during cold acclimation.

Identification of homologous genes regulated by stress in Arabidopsis

The 2,637 putative wheat stress-regulated genes identified in our study were BLASTed (TBLASTX) against the *Arabidopsis* proteins TAIR database [12] using a cut-off e-value of e-25. The Protein ID of the homologous *Arabidopsis* proteins were used to identify those that are represented on the Affymetrix ATH1 genome array and the MWG Bio-

Table 7: Transcription factors that are differentially expressed in the FGAS dataset relative to the NSF-DuPont dataset.

Transcription factor family	over-represented 2 to 5-fold	over-represented over 5-fold	Contigs unique to FGAS with at least 3 ESTs	TOTAL
AP2 (ex. CBF1,2,3, Aintegumenta)	4	7	9	20
BHLH (Ex. AtMYC2)	5	2	0	7
BZIP (Ex. FD)	5	3	0	8
CCAAT-box transcription factor	2	1	0	3
DEAD/DEAH box helicase	4	4	0	8
F-box protein family	3	0	0	3
FLOWERING LOCUS T	1	0	1	2
GIGANTEA protein	0	1	1	2
Homeodomain Leucine zipper protein (Ex. ABF3 ABF4, ABA response)	2	2	1	5
MADS box transcription factor (Ex. TaVRT1)	2	0	0	2
MYB (Ex. AtMYB2)	14	7	2	23
NAC-domain containing protein (Ex. RD26 dehydration)	11	8	0	19
PHD finger (Ex. pollen development, chromatin-mediated transcription regulation, a variant of Zn-finger)	2	1	0	3
RING finger containing protein (Ex. HOS1 regulating cold response, A variant of Zn finger)	14	4	3	21
SCARECROW gene regulator-like (Ex. Oxidative stress)	3	1	0	4
WD-repeat containing protein	0	1	0	1
WRKY transcription factor (Ex. Drought, oxidative stress and pathogen induced)	14	7	7	28
Zinc finger protein (Ex. CO, Indeterminate-related)	30	11	6	47
Other Transcription factor-like	113	47	14	174
Other DNA-binding protein	143	46	11	200
Total	372	153	55	580

tech 25 k 50-mer oligonucleotide array. The cold- and drought-regulated genes were then identified from the available published data [14,15].

Authors' contributions

MH, FS, PG, AL and WLC conceived the study and participated in its design and coordination. MH carried out the analyses of the EST datasets and drafted the manuscript. MH, MB and AB carried out the bioinformatics analyses. FO and FS participated in the drafting and editing of the manuscript. JD constructed Libraries 2 to 6. AM, AD and PG prepared the clones from Libraries 2 to 6 for sequencing. AL constructed libraries TaLT2 to TaLT6 and prepared the clones for sequencing. ML, LMcC and WLC carried out the sequencing reactions, the bioinformatics analyses of the FGAS dataset, and submitted the data to Genbank. All authors read and approved the final manuscript.

Additional material

Additional File 1

Contigs containing ESTs that are over- or under-represented at least two-fold in the FGAS dataset compared to the NSF/DuPont dataset. *SSH ESTs are not part of this analysis. The contigs containing ESTs over-represented at least 5-fold in FGAS were analyzed by TBLASTX against the* Arabidopsis *TAIR database to find homologues (e-25 cut-off). For those that are represented on the Affymetrix and/or MGW microarrays, the expression data with respect to cold or drought regulation was obtained. U, up-regulated; D, down-regulated.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-7-149-S1.xls]

Additional File 2

Contigs containing at least three ESTs that are present only in the FGAS dataset. *SSH ESTs are not part of this analysis. The contigs were analyzed by TBLASTX against the* Arabidopsis *TAIR database to find homologues (e-25 cut-off). For those that are represented on the Affymetrix and/or MGW microarrays, the expression data with respect to cold or drought regulation was obtained. U, up-regulated; D, down-regulated.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-7-149-S2.xls]

Additional File 3

Contigs containing at least three ESTs that are present only in the TaLT libraries of the FGAS dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-149-S3.xls]

Acknowledgements

This work was funded by Genome Canada (MH, PG, AL, WLC, FS), Genome Prairie (AL, WLC), Génome Québec (MH, PG, FS) and Canarie (MH, FS). We thank the technical staff and students who participated in this study.

References

1. Sarhan F, Ouellet F, Vazquez-Tello A: **The wheat wcs120 gene family. A useful model to understand the molecular genetics of freezing tolerance in cereals.** *Physiol Plant* 1997, **101**:439-445.
2. Initiative TAG: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408**:796-815.
3. Project IRGS: **The map-based sequence of the rice genome.** *Nature* 2005, **436**:793-800.
4. Bennett MD, Leitch IJ: **Nuclear DNA amounts in angiosperms.** *Ann Bot* 1995, **73**:113-176.
5. Sasaki T: **Rice genome analysis: Understanding the genetic secrets of the rice plant.** *Breed Sci* 2003, **53**:281-289.
6. Bräutigam M, Lindlöf A, Zakhrabekova S, Gharti-Chhetri G, Olsson B, Olsson O: **Generation and analysis of 9792 EST sequences from cold acclimated oat, Avena sativa.** *BMC Plant Biol* 2005, **5**:18.
7. Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP: **A new resource for cereal genomics: 22K barley GeneChip comes of age.** *Plant Physiol* 2004, **134**:960-968.
8. Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ: **Comprehensive EST analysis of tomato and comparative genomics of fruit ripening.** *Plant J* 2004, **40**:47-59.
9. Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S: **A Populus EST resource for plant functional genomics.** *Proc Natl Acad Sci USA* 2004, **101**:13951-13956.
10. Ogihara Y, Mochida K, Kawaura K, Murai K, Seki M, Kamiya A, Shinozaki K, Carninci P, Hayashizaki Y, Shin I, Kohara Y, Yamazaki Y: **Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags.** *Genes Genet Syst* 2004, **79**:227-232.
11. Zhang D, Choi DW, Wanamaker S, Fenton RD, Chin A, Malatrasi M, Turuspekov Y, Walia H, Akhunov ED, Kianian P, Otto C, Simons K, Deal KR, Echenique V, Stamova B, Ross K, Butler GE, Strader L, Verhey SD, Johnson R, Altenbach S, Kothari K, Tanaka C, Shah MM, Laudencia-Chingcuanco D, Han P, Miller RE, Crossman CC, Chao S, Lazo GR, Klueva N, Gustafson JP, Kianian SF, Dubcovsky J, Walker-Simmons MK, Gill KS, Dvorak J, Anderson OD, Sorrells ME, McGuire PE, Qualset CO, Nguyen HT, Close TJ: **Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (Triticum aestivum L.).** *Genetics* 2004, **168**:595-608.
12. **Functional Genomics of Abiotic Stress (FGAS).** 2006.
13. Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NL, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD: **Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map.** *Genetics* 2004, **168**:585-593.
14. Hannah MA, Heyer AG, Hincha DK: **A global survey of gene regulation during cold acclimation in Arabidopsis thaliana.** *PLoS Genet* 2005, **1**:e26.
15. Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R: **When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress.** *Plant Physiol* 2004, **134**:1683-1696.
16. Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, Satou M, Akiyama K, Taji T, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y, Shinozaki K: **Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray.** *Plant J* 2002, **31**:279-292.
17. Xin Z, Browse J: **Eskimo1 mutants of Arabidopsis are constitutively freezing-tolerant.** *Proc Natl Acad Sci USA* 1998, **95**:7799-7804.
18. Breitkreuz KE, Allan WL, Van Cauwenberghe OR, Jakobs C, Talibi D, Andre B, Shelp BJ: **A novel gamma-hydroxybutyrate dehydrogenase: identification and expression of an Arabidopsis cDNA and potential role under oxygen deficiency.** *J Biol Chem* 2003, **278**:41552-41556.
19. Sirko A, Blaszczyk A, Liszewska F: **Overproduction of SAT and/or OASTL in transgenic plants: a survey of effects.** *J Exp Bot* 2004, **55**:1881-1888.
20. Uemura M, Steponkus PL: **A contrast of the plasma membrane lipid composition of oat and rye leaves in relation to freezing tolerance.** *Plant Physiol* 1994, **104**:479-496.
21. Holmberg N, Harker M, Gibbard CL, Wallace AD, Clayton JC, Rawlins S, Hellyer A, Safford R: **Sterol C-24 methyltransferase type 1 controls the flux of carbon into sterol biosynthesis in tobacco seed.** *Plant Physiol* 2002, **130**:303-311.
22. Shen H, Dowhan W: **Regulation of phospholipid biosynthetic enzymes by the level of CDP-diacylglycerol synthase activity.** *J Biol Chem* 1997, **272**:11215-11220.
23. Charron JBF, Breton G, Danyluk J, Muzac I, Ibrahim RK, Sarhan F: **Molecular and biochemical characterization of a cold-regulated phosphoethanolamine N-methyltransferase from wheat.** *Plant Physiol* 2002, **129**:363-373.
24. Takahashi M, Yamaguchi H, Nakanishi H, Shioiri T, Nishizawa NK, Mori S: **Cloning two genes for nicotianamine aminotransferase, a critical enzyme in iron acquisition (Strategy II) in graminaceous plants.** *Plant Physiol* 1999, **121**:947-956.
25. Higuchi K, Suzuki K, Nakanishi H, Yamaguchi H, Nishizawa NK, Mori S: **Cloning of nicotianamine synthase genes, novel genes involved in the biosynthesis of phytosiderophores.** *Plant Physiol* 1999, **119**:471-479.
26. Allard F, Houde M, Kröl M, Ivanov A, Huner NPA, Sarhan F: **Betaine improves freezing tolerance in wheat.** *Plant Cell Physiol* 1998, **39**:1194-1202.
27. Hincha DK, Oliver AE, Crowe JH: **Lipid composition determines the effects of arbutin on the stability of membranes.** *Biophys J* 1999, **77**:2024-2034.
28. Breton G, Danyluk J, Ouellet F, Sarhan F: **Biotechnological applications of plant freezing associated proteins.** *Biotechnol Annu Rev* 2000, **6**:59-101.
29. Knight CA, DeVries AL, Oolman LD: **Fish antifreeze protein and the freezing and recrystallization of ice.** *Nature* 1984, **308**:295-296.
30. Gaudet DA, Laroche A, Frick M, Davoren J, Puchalski B, Ergon : **Expression of plant defence-related (PR-protein) transcripts during hardening and dehardening of winter wheat.** *Physiol Mol Plant Pathol* 2000, **57**:15-24.
31. Yeh S, Moffatt BA, Griffith M, Xiong F, Yang DS, Wiseman SB, Sarhan F, Danyluk J, Xue YQ, Hew CL, Doherty-Kirby A, Lajoie G: **Chitinase genes responsive to cold encode antifreeze proteins in winter cereals.** *Plant Physiol* 2000, **124**:1251-1264.
32. Tremblay K, Ouellet F, Fournier J, Danyluk J, Sarhan F: **Molecular characterization and origin of novel bipartite cold-regulated ice recrystallization inhibition proteins from cereals.** *Plant Cell Physiol* 2005, **46**:884-891.
33. Hincha DK, Meins Jr. F, Schmitt JM: **ß-1,3-glucanase is cryoprotective in vitro and is accumulated in leaves during cold acclimation.** *Plant Physiol* 1997, **114**:1077-1083.

34. Öquist G, Huner NP: **Photosynthesis of overwintering evergreen plants.** *Annu Rev Plant Biol* 2003, **54**:329-355.
35. Jansen MA, Mattoo AK, Edelman M: **D1-D2 protein degradation in the chloroplast. Complex light saturation kinetics.** *Eur J Biochem* 1999, **260**:527-532.
36. Ishihara S, Yamamoto Y, Ifuku K, Sato F: **Functional analysis of four members of the PsbP family in photosystem II in Nicotiana tabacum using differential RNA interference.** *Plant Cell Physiol* 2005, **46**:1885-1893.
37. Ifuku K, Yamamoto Y, Ono TA, Ishihara S, Sato F: **PsbP protein, but not PsbQ protein, is essential for the regulation and stabilization of photosystem II in higher plants.** *Plant Physiol* 2005, **139**:1175-1184.
38. Makino A, Sakashita H, Hidema J, Mae T, Ojima K, Osmond B: **Distinctive responses of ribulose-1,5-bisphosphate carboxylase and carbonic anhydrase in wheat leaves to nitrogen nutrition and their possible relationships to CO2 transfer resistance.** *Plant Physiol* 1992, **100**:1737-1743.
39. Moreno JI, Martin R, Castresana C: **Arabidopsis SHMT1, a serine hydroxymethyltransferase that functions in the photorespiratory pathway influences resistance to biotic and abiotic stress.** *Plant J* 2005, **41**:451-463.
40. Scheibe R, Backhausen JE, Emmerlich V, Holtgrefe S: **Strategies to maintain redox homeostasis during photosynthesis under changing conditions.** *J Exp Bot* 2005, **56**:1481-1489.
41. Schöttler MA, Kirchhoff H, Weis E: **The role of plastocyanin in the adjustment of the photosynthetic electron transport to the carbon metabolism in tobacco.** *Plant Physiol* 2004, **136**:4265-4274.
42. N'Dong C, Danyluk J, Huner NP, Sarhan F: **Survey of gene expression in winter rye during changes in growth temperature, irradiance or excitation pressure.** *Plant Mol Biol* 2001, **45**:691-703.
43. Varotto C, Pesaresi P, Meurer J, Oelmuller R, Steiner-Lange S, Salamini F, Leister D: **Disruption of the Arabidopsis photosystem I gene psaE1 affects photosynthesis and impairs growth.** *Plant J* 2000, **22**:115-124.
44. Haldrup A, Naver H, Scheller HV: **The interaction between plastocyanin and photosystem I is inefficient in transgenic Arabidopsis plants lacking the PSI-N subunit of photosystem I.** *Plant J* 1999, **17**:689-698.
45. Scheibe R: **Malate valves to balance cellular energy supply.** *Physiol Plant* 2004, **120**:21-26.
46. Lee BH, Lee H, Xiong L, Zhu JK: **A mitochondrial complex I defect impairs cold-regulated nuclear gene expression.** *Plant Cell* 2002, **14**:1235-1251.
47. Gilmour SJ, Fowler SG, Thomashow MF: **Arabidopsis transcriptional activators CBF1, CBF2, and CBF3 have matching functional activities.** *Plant Mol Biol* 2004, **54**:767-781.
48. Chinnusamy V, Ohta M, Kanrar S, Lee BH, Hong X, Agarwal M, Zhu JK: **ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis.** *Genes Dev* 2003, **17**:1043-1054.
49. Danyluk J, Kane NA, Breton G, Limin AE, Fowler DB, Sarhan F: **TaVRT-1, a putative transcription factor associated with vegetative to reproductive transition in cereals.** *Plant Physiol* 2003, **132**:1849-1860.
50. Kane NA, Danyluk J, Tardif G, Ouellet F, Laliberté JF, Limin AE, Fowler DB, Sarhan F: **TaVRT-2, a member of the StMADS-11 clade of flowering repressors, is regulated by vernalization and photoperiod in wheat.** *Plant Physiol* 2005, **138**:2354-2363.
51. **Metabolic pathways of the diseased potato** 2006 [http://www.scri.sari.ac.uk/TiPP/pps/Chart.pdf].
52. **MapMan** 2006 [http://gabi.rzpd.de/projects/MapMan/].
53. Danyluk J, Sarhan F: **Differential mRNA transcription during the induction of freezing tolerance in spring and winter wheat.** *Plant Cell Physiol* 1990, **31**:609-619.
54. Guo Z, Liu Q, Smith LM: **Enhanced discrimination of single nucleotide polymorphisms by artificial mismatch hybridization.** *Nat Biotechnol* 1997, **15**:331-335.
55. **Index of /NSF/curator/quality** 2006 [http://wheat.pw.usda.gov/nsf/curator/quality].
56. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
57. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
58. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
59. Breton G, Danyluk J, Charron JB, Sarhan F: **Expression profiling and bioinformatic analyses of a novel stress-regulated multispanning transmembrane protein family from cereals and Arabidopsis.** *Plant Physiol* 2003, **132**:64-74.
60. Shimosaka E, Sasanuma T, Handa H: **A wheat cold-regulated cDNA encoding an early light-inducible protein (ELIP): its structure, expression and chromosomal location.** *Plant Cell Physiol* 1999, **40**:319-325.
61. Castillo J, Rodrigo MI, Márquez JA, Zúñiga , Franco L: **A pea nuclear protein that is induced by dehydration belongs to the vicilin superfamily.** *Eur J Biochem* 2000, **267**:2156-2165.
62. Cattivelli L, Bartels D: **Molecular cloning and characterization of cold-regulated genes in barley.** *Plant Physiol* 1990, **93**:1504-1510.
63. Tsuda K, Tsvetanov S, Takumi S, Mori N, Atanassov A, Nakamura C: **New members of a cold-responsive group-3 Lea/Rab-related Cor gene family from common wheat (Triticum aestivum L.).** *Genes Genet Syst* 2000, **75**:179-188.
64. Koike M, Okamoto T, Tsuda S, Imai R: **A novel plant defensin-like gene of winter wheat is specifically induced during cold acclimation.** *Biochem Biophys Res Commun* 2002, **298**:46-53.
65. Houde M, Danyluk J, Laliberté JF, Rassart E, Dhindsa RS, Sarhan F: **Cloning, characterization and expression of a cDNA encoding a 50-kilodalton protein specifically induced by cold acclimation in wheat.** *Plant Physiol* 1992, **99**:1381-1387.
66. Shih MD, Lin SC, Hsieh JS, Tsou CH, Chow TY, Lin TP, Hsing YI: **Gene cloning and characterization of a soybean (Glycine max L.) LEA protein, GmPM16.** *Plant Mol Biol* 2004, **56**:689-703.
67. Livingston III DP, Henson CA: **Apoplastic sugars, fructans, fructan exohydrolase, and invertase in winter oat: responses to second-phase cold hardening.** *Plant Physiol* 1998, **116**:403-408.
68. Chauvin LP, Houde M, Sarhan F: **A leaf-specific gene stimulated by light during wheat acclimation to low temperature.** *Plant Mol Biol* 1993, **23**:255-265.
69. Gong Z, Dong CH, Lee H, Zhu J, Xiong L, Gong D, Stevenson B, Zhu JK: **A DEAD box RNA helicase is essential for mRNA export and important for development and stress responses in Arabidopsis.** *Plant Cell* 2005, **17**:256-267.
70. De Santis A, Landi P, Genchi G: **Changes of mitochondrial properties in maize seedlings associated with selection for germination at low temperature. Fatty acid composition, cytochrome c oxidase, and adenine nucleotide translocase activities.** *Plant Physiol* 1999, **119**:743-754.
71. Phillips JR, Dunn MA, Hughes MA: **mRNA stability and localisation of the low-temperature-responsive barley gene family blt14.** *Plant Mol Biol* 1997, **33**:1013-1023.
72. Massonneau A, Condamine P, Wisniewski JP, Zivy M, Rogowsky PM: **Maize cystatins respond to developmental cues, cold stress and drought.** *Biochim Biophys Acta* 2005, **1729**:186-199.
73. Kim KN, Lee JS, Han H, Choi SA, Go SJ, Yoon IS: **Isolation and characterization of a novel rice Ca2+-regulated protein kinase gene involved in responses to diverse signals including cold, light, cytokinins, sugars and salts.** *Plant Mol Biol* 2003, **52**:1191-1202.
74. Kowalski LR, Kondo K, Inouye M: **Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae.** *Mol Microbiol* 1995, **15**:341-353.
75. N'Dong C, Anzellotti D, Ibrahim RK, Huner NP, Sarhan F: **Daphnetin methylation by a novel O-methyltransferase is associated with cold acclimation and photosystem II excitation pressure in rye.** *J Biol Chem* 2003, **278**:6854-6861.
76. N'Dong C, Danyluk J, Wilson KE, Pocock T, Huner NP, Sarhan F: **Cold-regulated cereal chloroplast late embryogenesis abundant-like proteins. Molecular characterization and functional analyses.** *Plant Physiol* 2002, **129**:1368-1381.
77. Gao YP, Young L, Bonham-Smith P, Gusta LV: **Characterization and expression of plasma and tonoplast membrane aquaporins in primed seed of Brassica napus during germination under stress conditions.** *Plant Mol Biol* 1999, **40**:635-644.
78. Liu JH, Luo M, Cheng KJ, Mohapatra SS, Hill RD: **Identification and characterization of a novel barley gene that is ABA-inducible**

and expressed specifically in embryo and aleurone. *J Exp Bot* 1999, **50**:727-728.
79. Zhao TY, Martin D, Meeley RB, Downie B: **Expression of the maize galactinol synthase gene family: II) Kernel abscission, environmental stress and myo-inositol influences transcript accumulation in developing seeds and callus.** *Physiol Plant* 2004, **121**:647-655.
80. Campalans A, Pages M, Messeguer R: **Identification of differentially expressed genes by the cDNA-AFLP technique during dehydration of almond (Prunus amygdalus).** *Tree Physiol* 2001, **21**:633-643.
81. Kume S, Kobayashi F, Ishibashi M, Ohno R, Nakamura C, Takumi S: **Differential and coordinated expression of Cbf and Cor/Lea genes during long-term cold acclimation in two wheat cultivars showing distinct levels of freezing tolerance.** *Genes Genet Syst* 2005, **80**:185-197.
82. Salekdeh GH, Siopongco J, Wade LJ, Ghareyazie B, Bennett J: **Proteomic analysis of rice leaves during drought stress and recovery.** *Proteomics* 2002, **2**:1131-1145.
83. White AJ, Dunn MA, Brown K, Hughes MA: **Comparative analysis of genomic sequence and expression of a lipid transfer protein gene family in winter barley.** *J Exp Bot* 1994, **45**:1885-1892.
84. Potter E, Beator J, Kloppstech K: **The expression of mRNAs for light-stress proteins in barley: inverse relationship of mRNA levels of individual genes within the leaf gradient.** *Planta* 1996, **199**:314-320.
85. Barth O, Zschiesche W, Siersleben S, Humbeck K: **Isolation of a novel barley cDNA encoding a nuclear protein involved in stress response and leaf senescence.** *Physiol Plant* 2004, **121**:282-293.

APPENDIX B

Large scale EST analysis

Links, M. G., Nowak, J. J., and Crosby, W. L. (2004) Large Scale EST analysis.

Chapter 6

Large scale EST analysis

Matthew Links, J.J. Nowak and W.L. Crosby

Abstract

Expressed Sequence Tag (EST) collections serve as an initial foundation for investigation into the gene expression of a given organism. While there have been advances in the acquisition of genomic sequence data this has been limited to a select group of organisms. Thus for many organisms EST data will continue to be a major focus of inquiry. The data present in an EST collection consists of varying quality, redundant sequences and incomplete transcripts. Presented here is a discussion of the approaches commonly undertaken to deal with the imperfect nature of EST data and thereby glean insight into the gene expression of a given organism.

1. INTRODUCTION

For organisms without a finished genome, Expressed Sequence Tag (EST) collections are a crucial source of information representative of the genes actively being transcribed. In combination with a genome sequence, ESTs can be used to determine the chromosomal location of genes through sequence homology. Furthermore, ESTs can be invaluable for validating computational predictions of exon boundaries and splice sites when examining predicted gene structure at the chromosomal level. Several international projects are dedicated to the archiving and exploitation of EST collections through clustering and assembly of the EST sequences including dbEST (a division of GenBank; Boguski *et al.*, 1993), STACKdb™ (Christoffels *et al.*, 2001; Miller *et al.*, 1999) and the TIGR Gene Indices (Liang *et al.*, 2000; Quackenbush *et al.*, 2000; 2001; Kasukawa *et al.*, 2003).

As of November 2003, dbEST contained over 19 million EST sequences including more than 5 million for *Homo sapiens*, nearly 4 million for Mus musculus and Mus domesticus, as well as one half million each for Triticum aestivum and Gallus gallus . DbEST is the major international repository for EST sequences. The EST sequences present in dbEST are clustered into non-redundant sets and distributed within a GenBank division called UniGene (Pontius *et al.*, 2003). Through clustering, UniGene aims to provide a non-redundant representation of the EST sequences in dbEST. The goal of STACKdb™ and the TIGR Gene Indices (TGIs) is analogous to UniGene in that each of these projects produces a non-redundant representation of the underlying EST information. As such, using the projects together presents advantages for maximizing the utility of the underlying EST sequence information.

Each of the non-redundant EST sequence sets is an attempt to computationally derive the most complete representation of the actual messenger RNAs present in the original cell, tissue or organism. To accomplish this, each project or suite of software applications proceeds through four major steps: filtering of the primary sequencing reads for quality, partitioning of the sequences into clusters, assembly of the clusters into contigs, and finally

the addition of annotation to the contigs. Initial filtering of the sequences for quality includes attempts to remove portions of the sequencing read which are not actually part of the original messenger RNA sequence, such as the cloning vector. The initial clustering of sequences is done to segregate the sequences into groups based on a minimum similarity level. Clusters are subsequently assembled into representative contigs which then form the non-redundant representation of the EST collection. Lastly, the representative contigs are used in further analyses in order to ascribe them an annotation and a putative function.

Lewis et al.

2. FILTERING SEQUENCES FOR QUALITY

2.1 Assessing the Quality of Sequences

The first annotation an EST sequence receives is whether the sequencing reaction meets some minimum quality threshold in order to exclude, up front, those sequences which are inherently of low quality. If allowed to continue forward in the analysis, low quality sequences may taint the subsequent assembly and annotation processes. Sequence assembly performed in the presence of quality information greatly enhances the accuracy with which assembly programs identify the correct assembly. Likewise, the filtering of sequences for quality serves to improve the accuracy of downstream annotation, thereby reducing inaccuracies inherent to the prediction of function.

There are two levels at which a sequence is typically assigned a quality value: as a whole (globally) and at the nucleotide level (locally). The global sequence quality is strictly a measure of the sequencing reaction itself; in other words, "Was the sequencing reaction successful?" Local sequence quality, on the other hand, provides statistical support for each nucleotide call, where the most commonly used program for determining local quality information is Phred (Ewing *et al.*, 1998; Ewing and Green, 1998).

Phred reads and analyzes trace data directly to produce nucleotide calls and corresponding quality information. This local quality information is then used when assembling sequences to resolve base-pair mismatches and, in some cases, to identify single nucleotide polymorphisms(Marth *et al.*, 1999). Phred attempts to estimate where the peaks in sequence trace data should occur and, using these locations, dynamically map them to the actual observed centers of the peaks for each channel giving rise to a base, or nucleotide position. Finally, sampling the peak intensity at each base position provides a quantitative measurement for each nucleotide call such that:

$$Q(n) = -10 * \log_{10}(\text{Perror}(n))$$

Where n is the specific nucleotide (Green, 1999)).

A quality value of 10 corresponds to a 1/10 chance that the nucleotide called in this position is in error, whereas a quality value of 20 equates to a 1/100 chance of an erroneous base call at a given position in the sequence.

Global sequence quality is usually a qualitative measure of the behavior of the local sequence quality along the entire length of the sequence. The objective of performing a qualitative measure is to determine whether a sequence should be eliminated or should continue through subsequent analysis steps. Another way to think about the question of global sequence quality is to ask "is there a region within this sequence that should be analyzed further?" In practice this would mean that the identification of a single high quality subsequence is enough to proceed with the analysis of the sequence.

The simplest way to look at global sequence quality is to look for a consecutive stretch of nucleotides which score above some arbitrary threshold (perhaps Q(n) >= 15). This type of quality measure is limited because, if a single base falls below the selected quality level, then the sequence may fail the qualitative test. A similar problem would arise if the qualitative test were the mean of quality values over a region. In the case of a mean, the test might fail if there were more low quality bases than high quality ones. In practice a more complex windowed calculation tends to be used. For example the test might consist of finding the maximum sustained quality value.

To do this one could take the average Phred value within a window of length 20 and for a consecutive set of windows (*e.g.* 5), calculate the maximum sustained quality level for all the consecutive windows. Finally, the maximal sustained average quality should be assessed against the qualitative measure to determine whether to proceed with the analysis of this sequence. In example given, the qualitative measure is based on the maximum sustained average over 100 base pairs.

2.2 Masking Vector sequence

EST sequencing data often contain some portion of the vector sequence. Failure to remove this portion of the sequence may result in significant problems when it comes to clustering and assembling the raw sequences. Since clustering and assembly programs identify regions of similarity between sequences, the presence of common vector sequence among multiple independent sequence reads could create an artificial similarity between them, resulting in poor quality assemblies or sequences being placed in incorrect clusters.

There are three options for dealing with vector or other contaminating sequences: they can be removed, masked, or simply their boundaries identified. In the case of masking, nucleotides are typically converted to Xs. If simply identifying the boundaries of the EST insert is desired, then a program like Lucy (Chou and Holmes, 2001) can be used to identify the recombinant portion of the sequence. Determining which course to follow depends on the goals of the analysis, the component programs used, and the way in which the analysis results will be subsequently interpreted.

Cross_match (Green, 1999) is a commonly used masking program that is based on a specialized version of the Smith-Waterman algorithm that limits the search for an optimal local alignment to a region around one or more matching words. Masking requires that the nucleotide sequence of the vector is known and yields a FASTA sequence with the corresponding vector sequence replaced, or masked with Xs. Consideration should be given to whether any linker-adapter sequences were used in the cloning and whether they are included in the vector sequence used for masking. If not, downstream analysis would require the creation of a non-standard vector sequence that includes extraneous cloning sequences, leading to a more stringent masking.

3. REDUCING REDUNDANCY

3.1. Clustering sequences

The problem of deriving consensus sequence(s) from an EST collection en masse can be so computationally intensive that it is not feasible to undertake the analysis in real time. Clustering serves to reduce to a tractable number the number of sequences used to derive a consensus sequence. The criterion used to determine cluster membership is normally defined by the level of sequence similarity, where the similarity could be global (a percentage of

the total sequence required to be identical) or local. In the latter case, the input sequences may exhibit a window of similarity and overlap, but may nevertheless be globally very distinct.

TGICL (Pertea *et al.*, 2003) is an automated solution for the clustering and assembly of diverse EST collections while providing for a relatively low computational time. As with all clustering systems, TGICL should be used on EST sequence data that has been filtered for vector sequence, any cloning adapters-linkers used and should be masked for repetitive elements. TGICL is based on a two-stage model: compartmentalizing the sequences into clusters followed by assembling the ESTs for each cluster to create a longer and more complete consensus sequence.

Clustering in TGICL is accomplished using mgblast which is a modified version of megablast (Zhang *et al.*, 2003). Mgblast identifies overlaps, based on some level of similarity, between the input EST sequences. Overlaps are identified between each sequence pair in the EST database and, since each pairwise comparison is independent of every other one, it is possible to perform the pairwise comparisons in parallel. TGICL is capable of partitioning the pairwise comparisons into slices, which are subsequently allocated to multiple processes running on the same machine (for SMP architectures) or across a distributed supercomputer using PVM. This parallelization strategy reduces computational time and makes tractable the clustering of very large datasets.

Upon obtaining all of the overlaps present in an EST database, the TGICL software sorts the overlaps into a list based on the pairwise alignment score. Clusters are derived by applying transitive closure starting from the highest scoring pair and iterating over all overlaps. For example:

EST sequences in the database:
{ A, B, C, D, E, F, G, H, I }

Sorted Overlaps (decreasing pairwise alignment scores):
{ (A, B), (G, H), (B, C), (E, I), (A, D), (F, G) }

Leading to the following transitive closure:
{ (A, B), (B, C), (A, D) }
{ (G, H), (F, G) }
{ (E, I) }

This gives rise to the clusters
{ (A, B, C, D), (F, G, H), (E, I) }

For each cluster, TGICL builds a corresponding FASTA file containing the sequences of the component ESTs. The FASTA file is subsequently passed to an assembly program such as CAP3 (the default assembler) although this may be changed by the user (CAP3 is discussed in detail later in this chapter). In the case of the TIGR Gene Indices (TGIs), contigs resulting from the CAP3 assembly phase are referred to as Tentative Contigs (TCs), given a provisional functional assignment and then released publicly in the TGI for the appropriate organism.

The assembly of each cluster is undertaken independent from every other cluster in a manner analogous to the pairwise alignments in the clustering phase, thereby allowing for parallelization in the assembly process. TGICL uses a forking model to achieve parallelization for SMP architectures and uses PVM to allow parallelization in a distributed environment.

One additional advantage provided by TGICL is the optional usage of seeds in the clustering phase. TGICL reserves a sequence name prefix "etl" that is used to identify sequences which are full length. These full length EST sequences are used to anchor clusters and thereby limit their extension via overlaps. This becomes important if a full length EST exists since TIGCL will use the full length sequence to identify and avoid the incorporation of chimeric sequences. In some cases the use of seeds can be warranted but consideration must be made for the consequences of bounding the clusters, since the whole purpose of using a full length sequence is to apply fixed boundaries to the cluster.

The STACK project is an initiative of the South African National Bioinformatics Institute (SANBI) that aims to provide a comprehensive representation of every expressed human gene. The project has given rise to two major components: STACKdb™ which is the set of contigs and stackPACK™ which is the software used to generate the contigs. In many ways the STACK project is similar to the TIGR Gene Indices in that they both computationally derive clusters of sequences and subsequently assemble the sequences to give rise to a set of representative contigs. However, the algorithms used differ between stackPACK™ and TGICL and may lead to different representative contigs.

StackPACK™ is an automated analysis pipeline consisting of four major computational components: masking, clustering, assembly, and analysis of the EST data. The masking step involves the masking-out of vector sequence and other contaminants from the EST sequences using either RepeatMasker (Smit and Green, 2004) or cross_match (Green, 1999), and is entirely analogous to the masking performed by TGICL.

The clustering engine employed in stackPACK™ is provided through the use of d2_cluster (Burke *et al.*, 1999). D2_cluster is an agglomerative clustering method in which every sequence begins in its own cluster and clusters are subsequently merged if some similarity criteria is met or exceeded. The criteria for merger are a measure of identity equal to or in excess of some "stringency" value over a window of at least "Window_size" nucleotides where typical parameter set would be a Stringency of 90% and a Window_size of 150 nucleotides.

D2_cluster iterates over the EST sequences and merges two clusters if any two component sequences (one from each cluster) have an alignment meeting the minimal criteria. The first iteration is to take the first sequence in the database and compare it to all the others, adding them to the first cluster if they pass the similarity criteria. The next iteration takes the first sequence (which is a single sequence cluster) and compares it to the remaining single sequence clusters, subsequently merging the clusters if they pass the criteria.

Upon completion, d2_cluster obtains transitive closure on the grounds that clusters will have been joined if any sequence provides an alignment that meets or exceeds the similarity criteria. Both d2_cluster and the approach used in TGICL are forms of transitive closure. As such both d2_cluster and TGICL should produce similar, if not identical clusters, given equal stringency parameters. Where the two approaches differ is in the pairwise alignment of sequences: TGICL uses BLAST (Altschul *et al.*, 1999) in the form of megablast whereas d2_cluster uses a specialized version of the d2 algorithm (Hide *et al.*, 1994).

Malde *et al.* (2003) proposed a distinctively different and efficient algorithm for clustering EST sequences using a suffix array. Suffix arrays are used to represent suffixes which are shared between sequences. In order for two sequences to have a "matching block" they must contain contiguous regions of identity between the two sequences. Suffix arrays are used to represent the maximal matching blocks, meaning that the sequences differ immediately beyond each end point of the block.

The algorithm presented by Malde *et al.* uses the matching blocks between sequences to determine the pairwise score. Clusters are subsequently assembled by examining each sequence less similar, on the basis of matching blocks. The process specifically involves:

(1) Identify all matching blocks of length k:
   (a) Construct all suffixes from the data.
   (b) Sort the suffixes into a suffix array.
   (c) Group the suffixes that share a prefix of length at least k into cliques.
   (d) For each clique, generate the maximal matching blocks between each pair of suffixes in the clique.

(2) Score the resulting sequence pairs:
   (a) For each pair of sequences sharing at least one matching block, collect all matching blocks between two sequences
   (b) Calculate the largest consistent set of matching blocks and score them

(3) Generate the clustering:
   (a) Starting with the highest scoring sequence pair and working downward, build clusters hierarchically by connecting sequences
   (b) Split the clusters according to the clustering threshold

The major advantage of using suffix arrays is that they reduce the number of pairwise sequence comparisons which need to be performed by eliminating some that would produce low scores. Clustering approaches such as BLAST and d2_cluster require $O(m^2)$ comparisons to be performed (where m is the number of sequences) whereas the Malde algorithm is capable of using less than m scores in a single linkage clustering.

There are trade offs made when using a suffix array based algorithm. While the number of BLAST or d2_cluster pairwise comparisons may be linear as opposed to quadratic (with respect to the number of sequences) , the memory usage is usually higher. Malde *et al.* proposed a way to reduce the amount of memory used in creating the suffix array by eliminating the persistence of intermediate data structures. The authors argue they will be able to save 4^l, where l is the length of a prefix and, since the selection of the prefix applies globally, it is fixed relative to the size of the dataset. Therefore the memory savings will not scale with the number of sequences in the EST library. Thus, for very large EST data sets the memory requirements may grow beyond a tractable level. However, for smaller datasets the use of a suffix array algorithm may be desirable because of its faster computational time.

3.2 Assembling clusters

The primary goal of EST analysis is to produce a set of contigs which represent the total diversity in the original messenger RNA that the EST collection derives from. While clustering serves to segment the EST collection into groups of related sequences, it does not provide a mechanism for determining the identity of the representative sequence(s) for that cluster. The creation of a final, representative contig is typically derived through the application of an assembler. Performing EST assemblies differs from the assembly of genomic sequences in that EST sequences tend to present more noise. Furthermore the presence of vector contaminants, single pass reads, polymorphisms, and sequencing errors

make the assembly of EST sequences more tenuous (Liang *et al.*, 2000). For example one of the most common issues arising in an assembly is the resolution of base pair mismatches. If quality values are available most assembly programs can use the quality values to determine which nucleotide call is correct, however the absence of quality information from most public repositories such as GenBank make the incorporation of quality information impractical if not impossible. Thus it is in the context of these complications that assembly software must determine canonical contig(s) for a cluster.

Phrap is an assembly program which uses a banded implementation of the Smith-Waterman algorithm that relies on the identification of identical subsequences, or "word matches" between two sequences. A "word match" is a maximal subsequence that cannot be extended in either direction without a base pair mismatch. Each "word match" is then used to identify a band, centered on the diagonal of the dot matrix for the two sequences, within which the optimal alignment is identified. In the case where there are two or more "word matches" found, the union of bands identified by all the "word matches" is searched for the optimal alignment. The assembly of sequences is performed by progressively merging sequence pairs of decreasing score, giving rise to contigs which are the consensus of the sequencing reads.

In performing assemblies, Phrap gives special consideration to identifying homo-polymeric regions at the start and end of each sequence. Such regions are usually of poor quality and would mislead the alignment if used. Phrap also attempts to detect chimeras in which the "confirmed" region (the region of a sequence which aligns to another sequence) can be separated into two or more pieces. Chimeric sequences may derive from chimeric clones, but can also arise from incomplete masking of vector sequence. Perhaps most importantly, Phrap will use base quality information when performing alignments if the quality data is present.

CAP3

The CAP3 (Huang and Madan, 1999) assembly process is similar to Phrap in that 5' and 3' regions of poor quality are first identified and then removed. Overlaps are calculated for sequence pairs, producing a list of pairwise scores representative of the shared identity between reads. Sequences become joined on the basis of decreasing overlap score.

CAP3, unlike Phrap, inspects potential overlaps by examining the optimal global alignment of two "clean" reads where a "clean" read refers to the region of the sequencing read devoid of the 5' and 3' low quality regions. Huang and Madan explained (Huang and Madan, 1999) that examining the optimal global alignment between two clean reads can reveal whether there are regions of good quality sequence that are not shared between the two sequences. Such cases may be indicative of false overlaps.

For each pair of reads containing a true overlap, CAP3 searches a band of diagonals centered at the optimal local alignment and in this respect is analogous to Phrap. CAP3, like Phrap, also uses base quality values to aid in the calculation of alignments. Furthermore, CAP3 uses quality values to examine each overlap with respect to its sequencing error rate by comparing the sequencing error rate of the overlap against sequencing error rate of each of the reads contributing to it. If there is a discrepancy between the observed sequence error rate of the overlap and either of the component reads, the overlap is deemed false.

An added feature of CAP3 is the application of forward and reverse constraints. Constraints are used to bind a particular sequence to the forward or reverse strand of the assembly based on prior knowledge about which primer site was used in the sequencing

reaction. One application of directional constraints is the case where there are only two sequencing reads, one using the forward primer site and the other using the reverse site. In this case the constraints are useful in limiting the sequence alignment.

Potential problems arise in the use of directional constraints, particularly when the constraint is simply false. For EST libraries cloned using a directional cloning method, constraints can be readily applied in the assembly due to the consistency of orientation throughout the library. A forward constraint in every sub-clone of a directionally cloned library has exactly the same meaning, whereas in the case of a non-directionally cloned library the use of constraints is much more tenuous since a forward constraint in one sub-clone may be the same as a reverse constraint in another.

In cases where there is some ambiguity in the directional nature of the cloning (Huang and Madan, 1999) constraints can still be used. For example, CAP3 attempts to deal with such ambiguity by reliance on the assertion:

"a majority of the constraints are correct and wrong constraints usually occur randomly." (Huang and Madan, 1999)

Thus CAP3 operates under a model where a few constraints can be violated but most will be upheld. This approach may be suitable if the cloning methodology was predominately directional. The application of the directional constraints allows CAP3 to make corrections to the initial contigs from which CAP3 performs a multiple sequence alignment that yields the final consensus sequence.

In performing the multiple sequence alignment, CAP3 aligns the next read to the current alignment according to their position in the contig. Again, the base quality values are used to validate the sequence alignments. Base quality assignment to the consensus sequence is determined by the column-wise contribution of the bases within the multiple sequence alignment. This approach to quality assignment can lead to a very low quality scores if a specific nucleotide is found to be polymorphic.

Liang *et al.* (2000), have sought to objectively examine the performance of four assembly applications: Phrap, CAP3, TIGR Assembler (for genomic sequence assemblies), and TA-EST (TIGR Assembler for EST sequences), resulting in illustration of many of the issues that arise in the assembly of EST sequences.

The authors point out that sequencing errors arising from automated DNA sequencing tend to occur at either end of the read, where these are more problematic for EST sequence assemblies than for genomic sequence assemblies. In genomic sequence assembly the origin of a sequencing read is effectively random. However, EST sequencing reactions tend to line up due to the commonality of EST sequences derived from oligo(dT) primed reverse transcription reactions. Therefore the positional localization of sequencing errors poses an additional problem for EST assembly software in that all sequencing reads are likely to have low quality information at the extremities of the assembly.

Liang *et al.* (2000) also demonstrated the effect that sequencing errors have on EST assemblies by performing assemblies on sequences with imposed error rates between 1 and 8%, and lengths ranging between 450 -550 base pairs. The authors were able to show that these error rates were large enough to mislead some assembly programs - both TIGR Assembler and TA-EST split the sequences into two contigs while Phrap and CAP3 both generated single consensus sequences, although the consensus sequence produced by Phrap was of consistently lower fidelity than that produced by CAP3. The authors also examined the tendency for assembly programs to perpetuate sequencing errors and ambiguity into

the consensus sequence they produce. Of particular note was the observation that Phrap tended to perpetuate sequence ambiguity in the form of nucleotide errors (insertions or substitutions) known to be wrong, and which actually increased as the number of sequences in the assembly increased. The work presented in the paper was in part the basis for adoption of CAP3 as the assembly software used to create the TIGR Gene Indices.

4. ANNOTATING SETS OF CONTIGS

The ability to create a representative set of contigs allows for a reduction of redundancy within an EST collection. However, without also adding annotation to the contig set, the utility of an EST collection is greatly limited. A typical first order annotation is to examine what shared similarities exist between the transcripts under study and those found in the same or unrelated organisms. Asking deeper and more rigorous questions about individual transcript function often involves a second order annotation. Mapping the representative contigs to genomic locations is also done in those instances where a genomic sequence is present. Most assignments of annotation for molecular function rely, at least in some part, on the assumption that:

"sequences that have similar sequence have similar biological function"

This assumption touches on one of the most difficult issues for Bioinformatics and Computational Biology: "When are two sequences similar enough to be considered functionally the same?" The unfortunate and seemingly trite answer is, when they are the same! This poses a problem in the assignment of annotations based on sequence homology in that the assignment of annotation should be conceptually "weighted" by the similarity of the two sequences. The matter is further complicated by the realization that an EST sequence is only the transcribed portion of the gene and that similar sequences found in databases may be regulated differently so as to effect an entirely different biological function.

This distinction between sequences that are similar versus the same is only meant to serve as a reasonable warning for accepting computational annotations as definitive, which they should never be. Instead, computational annotations provide nothing more and nothing less than a limited insight into the sequence in hand.

4.1 BLAST

By far the most common element of annotation to associate with a contig would be the similarity (that the contig has with other sequences) based on BLAST analysis. Contigs can be BLASTed against protein databases such as trEMBL, SWISSPROT, Brookhaven Protein Data Bank and the non-redundant protein database of GenBank. Such BLAST searches are typically done by translating the contig sequence and comparing against the amino acid sequences in the database.

There are several points to consider when performing this type of multi-organism BLAST search. Due to inherent redundancy in the genetic code, it is more advantageous to translate the given nucleotide sequence to an amino acid sequence and use this as a query against the desired database, rather than back-translating the known proteins in the database to DNA.

Furthermore, comparisons should be made against different taxonomic divisions of the protein databases. This enables the annotation to be more specific to the organism from which the EST collection is derived. For example, if the EST sequences are derived from

a plant species, these are best compared against a plant protein database as well as the total collection resident in the non-redundant protein database of GenBank. The resulting annotations can then be used in a progressive manner, looking primarily at similarities to other plants and then widening the examination to include vertebrates and other data sets.

It is important to understand and respect the limitations of the tools at hand. In the case of BLAST, one of the major limitations is that, in the general case, there is no controlled vocabulary imposed upon the contents of a sequence description. This problem becomes significant when a researcher wishes to perform keyword searching of BLAST reports. For example, suppose it was desirable to identify phosphatases in a given EST collection. This could be accomplished by BLASTing each contig against the non-redundant protein database of GenBank. The resulting BLAST reports could then be computationally screened for the keyword "phosphatase", although herein lies a significant complication: the lack of a controlled vocabulary means that special consideration must be paid to consider such common problems as alternative spellings, literal spelling mistakes and alternate meanings. In the "phosphatase" example the following record exists in GenBank (As of November 14, 2003).

```
LOCUS     NP_521001     246 aa     linear     BCT 09-DEC-2002
DEFINITION PROBABLE PHOSPHOGLYCOLATE PHOSPHATAS PROTEIN [Ralstonia solanacearum].
ACCESSION     NP_521001
VERSION       NP_521001.1 GI:17547599
DBSOURCE      REFSEQ: accession NC_003295.1
```

portion of the GenPept record for NP521001

This record contains a simple spelling mistake (phosphatas as opposed to phosphatase), yet it is this type of simple error which would, in the case of a keyword search for "phosphatase", have resulted in this record being invisible to the computer query and thereby the investigator.

There exists no concise way to entirely avoid these types of problems. Instead, one must anticipate them and understand the limitations they impose on the results. In the case of keyword searching of BLAST reports, one must realize that the BLAST algorithm pays exclusive attention to sequence similarity, so that interpreting BLAST results on anything other than sequence similarity is ill-advised.

4.2 The Gene Ontology

In the previous example of keyword searching, it is easy to appreciate the utility that such a keyword search offers - particularly since BLAST reports lack a controlled vocabulary for annotation terms. The use of a controlled vocabulary for the annotation of gene products, ideally for function and localization, would provide a way to rigorously compare gene products on the basis of these annotation terms.

The Gene Ontology Consortium (Ashburner and Lewis, 2002) is just such a project, endeavoring to create consistent descriptions for gene products in diverse and distributed databases. The GO Consortium grew out of collaboration between three model organism databases:

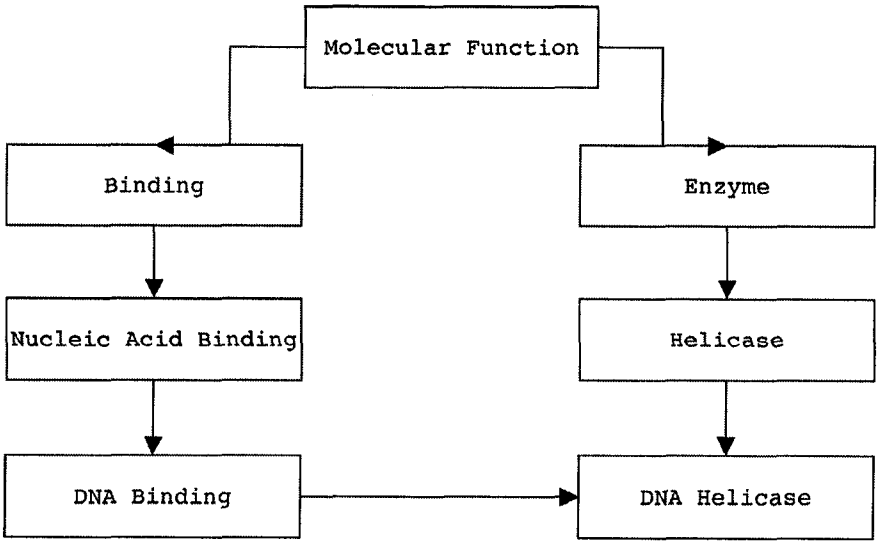

Figure 1. Sample of a portion of the Molecular Function DAG.

FlyBase http://flybase.bio.indiana.edu/
Saccharomyces Genome Database http://genome-www.stanford.edu/Saccharomyces/
Mouse Genome Database http://www.informatics.jax.org/

The GO Consortium has since grown to include major international repositories of biological information. GO is comprised of three ontologies which describe and classify a gene product's molecular function, the biological process it is part of, and what cellular components it is associated to. Each ontology is represented by a directed acyclic graph (DAG).

Figure 1 represents a small portion of the Molecular Function ontology using a DAG to show the hierarchical lineage of terms. Allowing and formalizing how terms relate to one another is a significant advantage of GO. For example, the Molecular Function term "Helicase" is a refinement, or specialization of, the term "Enzyme". The use of a DAG allows a term to be the child of multiple terms. In the example given "DNA Helicase" is a refinement of "Helicase" which is in turn a refinement of "Enzyme, DNA Helicase" is also a refinement of "DNA Binding" which is a refinement of "Nucleic Acid Binding", which is in turn a refinement of "Binding". The real power and utility of GO comes from the very rigorous meanings that ontology terms imply and how these are applied through annotation.

The GO database is based on the representation inherent to three ontology categories: Biological Process, Cellular Component, and Molecular Function. Gene products are associated to the component terms of each ontology through an evidence code (Table 1). The purpose of the evidence code is to identify why a gene product was associated to a particular GO term. An aim of GO is to provide a rigorous mechanism for the association of a GO term to a particular gene product.

Table 1. GO Evidence codes.	
Code	Meaning
IC	inferred by curator
IDA	inferred from direct assay
IEA	inferred from electronic annotation
IEP	inferred from expression pattern
IGI	inferred from genetic interaction
IMP	inferred from mutant phenotype
IPI	inferred from physical interaction
ISS	inferred from sequence or structural similarity
NAS	non-traceable author statement
ND	no biological data available
TAS	traceable author statement N.B. used to be known as ASS author said so
NR	not recorded

As of November 2003 the central GO database had in excess of 100,000 annotated gene products. To continue the aforementioned phosphatase example, a search for the GO term phosphatase using AmiGO (Amigo, 2004) can be performed. Of note is that the keyword searching is performed by first identifying those terms containing the term "phosphatase"; only then is there a progression to find gene products associated to each of the specific terms. The search for phosphatase brings up several terms containing phosphatase and if the interest was in "nucleotide phosphatase activity" then 4 gene product annotations are retrieved along with the full lineage of the term "nucleotide phosphatase activity". The gene products are presented along with the corresponding gene symbol, datasource, and evidence code responsible for assigning the gene product to this term.

```
CET1 SGD TAS RNA 5'-triphosphatase, mRNA capping enzyme beta subunit (80 kDa)
```

Example gene product annotated to "nucleotide phosphatase activity".

The GO database offers a means to traverse directly to the datasource that contains further information on a specific gene product. The utility of GO is that it allows a rigorous framework in which to annotate gene products, thereby enabling inter-database and inter-organism comparisons and queries.

4.3 How to 'GO' from an EST sequence?

One of the most desirable annotations for an EST sequence is the assignment of a putative function. A very simplistic way to do that would be to perform a BLAST search and then apply the description of the most similar sequences to the original EST. This is not, however, a rigorous method in that it relies on the assumption that similarity revealed by BLAST analysis is definitive for the function of the sequence in the database. InterPro (2004) is a collection of tools and databases built on known proteins, domains, and functional sites that can be applied to novel protein sequences in an attempt to glean some insight into their function.

The InterPro member databases include SWISS-PROT, prosite, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIR SuperFamily, and SuperFamily. These databases

range in information from primary sequence information with annotation, to protein fingerprints, hidden Markov Models and conserved motifs. The tools within InterPro are useful for identifying components of an unknown sequence which match entries in the member databases. However, InterPro is particularly valuable for the integration of the member databases into a single interface: the InterPro database (Interpro, 2004).

As with GO, InterPro can be searched on the basis of specific terms. InterPro is another example of a controlled vocabulary however, unlike GO, InterPro terms are not formally curated in a hierarchy, and therefore lack the lineage information of a GO term. Thus, the use of an InterPro term is less fluid and less amenable to refinement relative to use of a GO term. While this may be considered as a limitation of InterPro, InterPro is nevertheless very usefully deployed via an InterPro to GO mapping relationship, which readily allows the connection of InterPro and GO terms. This relationship makes it possible to derive a "putative" GO annotation for each sequence or contig in an EST collection. Using InterProScan, the annotation tool of InterPro, a sequence can be annotated to InterPro and the annotation term may be in turn mapped from the InterPro term to a GO term. This feature is desirable by the fact that it provides a rigorous way to derive functional classes by enabling the broad categorization of the EST sequences in the collection.

Pie charts are sometimes improperly used to represent the diversity of EST collections. The inherent problem lies in the fact that the determination of categories represented by a pie chart does not always correspond to the type of annotation derived for the collection. This becomes particularly acute when rationalized over all the annotation associated with an EST collection. As described previously, resorting to simple BLAST descriptions is difficult and prone to significant errors. However, if the annotations adopt the use of a controlled vocabulary, such as InterPro, then at least those sequences with the same annotation terms can be grouped together. Furthermore if an ontology of annotation terms is used, as in the case of GO, then the hierarchical relationship between terms provides broader categories suitable for the creation of pie charts.

4.4 When and what to annotate?

It is important to consider a schedule for the annotation of an EST collection. For example, in the case of large EST collections, it may be desirable to perform intermediate clustering and assembly analyses in order to assess sequence redundancy while the EST program is underway. It is therefore worth considering when to perform certain types of annotation. A well-designed annotation schedule will consider which annotations are immediately required versus those that are best left until the final assembled contigs are complete.

If the annotation depends heavily on the deduced amino acid sequence of the gene product, then the annotation will likewise be heavily dependent on the final assembly. This is particularly true for protein-domain finding tools (such as those in InterPro) which require viewing the entire, or predominately full length domain in order to recognize it. Accordingly, annotation of large features in a gene product should generally be left to the final annotation of the assembled contig set.

Protein annotation can be further complicated if the EST collection contains a content biased to the 3' end of messenger RNA since the N-terminus of the corresponding protein will be under-represented in the collection. If there is a canonical protein domain model in InterPro for a protein family of interest, and if that domain is predominately found in N-terminal regions, then it is unlikely that any contig or EST in the collection will ever

be annotated to that protein family. In such cases it would be advantageous to leave the annotation to the final assembled contigs on the possibility that the final contig assembly will extend sufficiently 5' to present a recognizable domain.

Annotations such as those revealed by BLAST analyses are universally desirable and as such they tend to be performed at both the individual EST sequence as well as the assembled contig levels. Another consideration is the computational resources available since it is usually advantageous to perform all of the possible analyses at every step. However, by performing the annotation on the final assembled contigs the number of sequences to annotate will be reduced and the length of the contigs may make it possible to identify features that would otherwise not be visible among individual EST sequences.

5. ITERATIVE CLUSTERING

5.1 Measuring redundancy of the EST library to limit sequencing efforts

EST collections where the library construction involved normalization for transcript abundance can lead to situations where the probability that the next sequencing read will belong to any specific gene are approximately equal. Therefore, if the objective of the EST sequencing program is to maximize the number of genes revealed with the fewest sequencing reactions, then measuring the redundancy of the collection can indicate when to cease sequencing efforts. As the number of genes covered by the EST sequences in the collection approaches saturation, the number of new sequencing reads that are unique will diminish. This can be measured on an ongoing basis by iteratively clustering the EST collection as sequencing progresses and observing the degree of redundancy between successive iterations.

To characterize the EST differences between libraries representing two time points in a series, one can assess whether all new sequencing reads cluster and assemble with previously sequenced ESTs. If the new sequences do assemble with old sequences, then the new reactions are redundant. If the distribution of the transcripts is random, then the behavior of the redundancy over time should be a sigmoidal curve that eventually plateaus, indicating that further sequencing of clones may not be warranted.

5.2 Maintaining sequence lineage across different clusterings and assemblies

A significant problem with iterative clustering, apart from the computational time to perform it, is that the coordinates of a given sequence within a contig must be re-evaluated and recorded at each iteration. In the case of clustering and assembly iterations, a sequence's coordinates are derived from its position in a given assembly. Each time a clustering is performed each sequence may potentially receive a new cluster assignment. Likewise for each assembly of that cluster, each component sequence will be assigned to some contig whose coordinates are simply the corresponding cluster and contig position. Suppose that there are the sequences:

{ A, B, C, D, E, F, G, H, I, J }

This clusters into:

Singletons	{ E }	this is a true singleton
Cluster 1	{ A, B, C, G, H, I, J }	
Cluster 2	{ D, F }	

Assembling into:

Singletons	{ E }	
Cluster 1	Contig 0 { H }	this is an assembly singleton
Cluster 1	Contig 1 { A, B, I, J }	
Cluster 1	Contig 2 { C, G}	
Cluster 2	Contig 1 { D, F }	

Giving the sequence coordinates

Sequence	Cluster	Contig
A	1	1
B	1	1
C	1	2
D	2	1
E	0	0
F	2	1
G	1	2
H	1	0
I	1	1
J	1	1

The advantage of retaining the sequence coordinates is that the coordinates are the minimal representation of the assembly, which in turn readily allows for the full reconstruction of the assembly. Therefore, as iterative rounds of clustering and assembly are undertaken it is permissible to eliminate the alignments so long as the sequence coordinates are retained. The retention of sequence coordinates also allows for the subsequent calculation of redundancy between iterations.

6. TAKING EST DATA FURTHER

6.1 Mapping into and out of the collection

By far the most common approach to mapping into and out of an EST collection is through a simple homology search program such as BLAST. Creating a BLAST database of all the ESTs present in a collection provides a means to retrieve information about the EST collection itself. For example, if a researcher has a canonical form of a gene - perhaps from a different but related organism - they can quickly assess if there is an EST in the collection worthy of further investigation. In this example the BLAST database contains each individual EST sequence, which would be appropriate for questions where the identification of a single EST is essential to the investigation.

In an analogous way, a BLAST database could be created from the assembled contigs of an EST collection. Such a database would be useful for asking questions regarding potential full-length transcripts present in the collection.

6.2 Comparing clustering and assembly results from distinct software

Performing comparisons between assembled sets of ESTs from two sources is analogous to the aforementioned situation of comparing results between iterations of assembly. Performing the comparisons requires that the sequence coordinates for every sequence under each condition being compared be known. As an example, assume that two conditions being compared (I and II) are represented by the following sequence coordinate sets:

Condition I

Sequence	Cluster	Contig
A	1	1
B	1	1
C	1	2
D	2	1
E	0	0
F	2	1
G	1	2
H	1	0
I	1	1
J	1	1

Condition II

Sequence	Cluster	Contig
A	1	1
B	1	1
C	1	2
D	2	1
E	3	1
F	2	1
G	1	2
H	3	1
I	1	3
J	1	3

In order to compare the two conditions, iterate over the sequences to see if the same sequence in each condition has the same sibling sequences (sequences within the same contig). If the sequences are examined exclusively at the level of clusters then the comparisons are analogous to those illustrated by Burke *et al.* (1999). In this example, sequences I and J are in a separate contig in condition II, as opposed to in condition I where they are with sequences A and B. This is an example of contig splitting between the two assemblies. Similarly, sequences H and E have merged in condition II. The result is an equal number of total unique transcripts, although the composition of the transcripts is distinct between conditions.

7. CONCLUSION

The creation of a non-redundant set of contigs representative of an EST collection provides a powerful tool for functional genomics. By searching the annotation of an EST collection, one can identify potential genes of interest. EST collections can also be used to investigate quantitative aspects of gene expression. In the case of non-normalized EST libraries, simple transcript counts for each contig may relate to transcript abundance in the original biological source cells or tissue. The non-redundant set of contigs for a collection can in turn be used in the design of microarray resources and experiments in order to directly measure expression of their corresponding genes.

DNA arrays can be constructed using the EST sequences representative for each contig. Common constraints on sequence selection include length, melting temperature and 3' bias. Oligonucleotide arrays based on synthetic 70-mer oligonucleotide probes designed to be specific to a single contig or EST sequence are increasingly being deployed for functional genomics studies (Wang *et al.*, 2003a). An important consideration in designing such oligonucleotide arrays is the identification of a unique oligonucleotide for each contig.

as well as physical constraints placed on the oligonucleotide design including melting temperature and the lack of secondary structure formation (Wang *et al.*, 2003b).

EST collections provide insight into the composition and nature of the expressed sequences present in an organism or tissue at a specific time and place in development. The utility of an EST collection derives from the ability to determine a non-redundant set of contigs which fully represent the diversity of the collection. By reducing the EST collection to a representative set, further investigation into questions surrounding gene expression and function in the biological source organism can be undertaken. In a reverse sense, individual EST sequences can be used as an initial approximation of gene expression in the organism. Perhaps, of greater interest is the potential for functionally meaningful annotation at both the single sequence and non-redundant contig levels, so as to enable the discovery of genes and to empower their application in functional genomics studies.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

AmiGO (2004). http://www.godatabase.org/cgi-bin/go.cgi

Ashburner, M., and Lewis, S.E. (2002). On ontologies for biologists: the Gene Ontology - uncoupling the web. In: *In Silico* Biology. Novartis Found Symp. 247, 66-80; discussion 80-3, 84-90, 244-252.

Boguski, M.S., Lowe, T.M., and Tolstoshev, C.M. (1993). dbEST-- database for "expressed sequence tags". Nat. Genet. 4, 332-3.

Burke, J., Davison, D., and Hide, W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res. 9, 1135-42.

Christoffels, A., van Gelder, A., Greyling, G., Miller, R., Hide, T., and Hide, W. (2001) STACK: Sequence tag alignment and consensus knowledgebase. Nucl. Acids Res. 29, 234-8.

Chou, H., and Holmes, M. (2001). DNA sequence quality trimming and vector removal. Bioinformatics 17, 1093-1104.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. (1998). Genome Res. 8, 175-185.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186-194.

Green, P. (1999). http://www.phrap.org/phrap.docs/phrap.html

Hide, W., Burke, J., and Davison, D.B. (1994). Biological evaluation of d2, an algorithm for high-performance sequence comparison. J. Comp. Biol. 1, 199-215.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Research 9, 868-77.

InterPro database (2004). http://www.ebi.ac.uk/interpro

Kasukawa, T., Furuno, M., Nikaido, I., Bono, H., Hume, D.A., Bult, C., Hill, D.P., Baldarelli, R., Gough, J., Kanapin, A., Matsuda, H., Schriml, L.M., Hayashizaki, Y., Okazaki, Y., and Quackenbush, J. (2003). Development and evaluation of an automated annotation pipeline and cDNA annotation system. Genome Res. 13, 1542-1551.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. (2000). An optimized protocol for analysis of EST sequences. Nucl. Acids Res. 28, 3657-65.

Malde, K., Coward, E., Jonassen, I. (2003). Fast sequence clustering using a suffix array algorithm. Bioinformatics. 19, 1221-1226

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. (1999). A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 23, 452-6.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R., and Hide, W.A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res. 9, 1143-55.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J.A., Vaughan, R, Zdobnov, E.M. (2003). The InterPro Database, 2003 brings increased coverage and new features. Nucl. Acids Res. 31, 315-318.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003). TIGR Gene Indices clustering tools (TIGCL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651-652.

Pontius, J.U., Wagner, L., and Schuler, G.D. (2003). UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucl. Acids Res.h 28, 141-145.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. 29, 159-164.

Smit, A.F.A., and Green, P. (2004). http://repeatmasker.genome.washington.edu/RM/RepeatMasker.html

Wang, H.Y., Malek, R.L., Kwitek, A.E., Greene, A.S., Luu, T.V., Behbahani, B, Frank, B., Quackenbush J., and Lee N.H. (2003a). Assessing unmodified 70-mer oligonucleotide probe performance on glass-slide microarrays. Genome Biol. 4, R5.

Wang, X., and Seed, B. (2003b). Selection of Oligonucleotide Probes for Protein Coding Sequences. Bioinformatics 19, 796-802.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J. Comp. Biol. 7, 203-214.

APPENDIX C

Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequenced-based methods

Hill, J. E., Goh, S. H., Money, D. M., Doyle, M., Li, A., Crosby, W. L., Links, M., Leung, A., Chan, D., and Hemmingsen, S. M. (2005) *Am. J. Obstet. Gynecol.* **193,** 682-692

DNA sequencing tracefiles for *CPN60* libraries were obtained for a group of patients (n=16) presenting as healthy and non pregnant at an outpatient Sexually Transmitted Diseases clinic, British Columbia Centre for Disease Control. Libraries were prepared by colleagues as described previously [27]. Sequencing reactions represented randomly selected *CPN60* PCR products from vaginal swabs of asymptomatic, non pregnant, sexually active females. In order to create a profile of the microbes present in each *CPN60* library contigs were identified by BLAST and FASTA [127] similarity searching against reference databases from cpnDB [28]. Similarity criteria for taxonomic assignment were based on a similarity of at least 80% identity over at least 200 nucleotides. With contigs assigned taxonomically the number of sequencing reads belonging to each contig was used to numerically weight the contribution of a given organism within the complex community. Percentage composition of the microbial community was calculated by the percentage of sequencing reactions belonging to specific organisms per library. Results were integrated into a web portal and delivered over the internet using SSL and Apache.

As mentioned elsewhere there are potential problems underlying the depth of the sequence data in each library [27,33]. In particular the recognition that the use of PCR based

profiling of the universal target for *CPN60* has sensitivity of the order of 1 in 1125. Therefore given that these libraries represent 480 sequencing reactions per patient these data are likely not comprehensive for the entire distribution of micro-organisms in the samples. However the methodology employed here for profiling microbial communities has a significant advantage in its ability to profile microbial communities in a culture independent manner. Most clinically applied techniques to profile microbial communities related to human health require specialized culture based assays. When performing culture based assays for micro-organisms the assays are inherently biased. The methodology shown here and elsewhere [26,27,30-33] has a distinct advantage over culture based assays because it is an unbiased approach. In principle, the sequencing of universal target sequences for *CPN60* is blind to constraints about biological niche whereas culture based assays intentionally exploit information about biological niche. Therefore in cases where organisms are not screened for, as is the case with *C. psittaci* shown here, culture based assays have the potential to miss entirely (false negative) the presence of certain organisms. Shown here is a system which automatically analyzes DNA sequencing reactions from complex microbial communities and determines the distribution of organisms within the community.

American Journal of Obstetrics and Gynecology (2005) 193, 682–92

ELSEVIER

American Journal of
Obstetrics & Gynecology
www.ajog.org

Characterization of vaginal microflora of healthy, nonpregnant women by *chaperonin-60* sequence-based methods

Janet E. Hill, PhD,[a] Swee Han Goh, PhD,[c] Deborah M. Money, MD, FRCSC,[d,*] Melissa Doyle,[a] Andra Li, BSc,[a] William L. Crosby, PhD,[b] Matthew Links, BSc,[b] Amy Leung,[c] Debbie Chan,[c] Sean M. Hemmingsen, PhD[a]

National Research Council of Canada, Plant Biotechnology Institute[a]; Department of Computer Science, University of Saskatchewan,[b] Saskatoon, Saskatchewan, Canada; Department of Pathology and Laboratory Medicine and UBC Centre for Disease Control, University of British Columbia[c]; and Department of Obstetrics and Gynecology, Children's and Women's Health Centre of British Columbia,[d] Vancouver, British Columbia, Canada

Received for publication May 28, 2004; revised January 6, 2005; accepted February 14, 2005

KEY WORDS
Vaginal flora
Chaperonin-60
Molecular method

Objective: The purpose of this study was to use a novel method that was based on the application of chaperonin-60 sequencing to describe the vaginal microflora of 16 healthy women.
Study design: Asymptomatic women consented for vaginal swabs to be collected at the time of a clinical pelvic examination. Total genomic DNA was isolated from the vaginal swabs. Degenerate, universal polymerase chain reaction primers were used to amplify an approximately 555 base pair region of the universal chaperonin-60 gene, which is found in all eubacteria and eukaryotes, from the total genomic DNA and libraries of cloned polymerase chain reaction products were constructed. Library clones were sequenced, and the resulting sequences were assigned to taxonomic groups on the basis of similarity to reference sequence data. Presence of *Chlamydophila psittaci* sequences in the samples was confirmed by species-specific polymerase chain reaction.
Results: Sixteen of the 23 women who were enrolled had normal flora by Nugent's score of <4 and had adequate polymerase chain reaction product for assessment. Vaginal flora libraries were dominated by a variety of sequences with similarity to *Lactobacillus* spp *L crispatus*, *L iners*, *L gasseri*, *L jensenii*, and *L buchneri*. Other sequences that were identified included representatives of *Gardnerella* spp, sequences with similarity to *Porphyromonas* spp and *Megasphaera* spp and sequences identical to *C psittaci*.

* Reprint requests: Deborah M. Money, MD, FRCSC, Assistant Professor & Head, Division of Maternal Fetal Medicine, University of BC, BC Women's Hospital, Rm 2H30, 4500 Oak St, Vancouver, BC V6H 3N1.
E-mail: dmoney@cw.bc.ca

0002-9378/$ - see front matter
doi:10.1016/j.ajog.2005.02.094

Conclusion: Culture-independent, chaperonin-60 sequence-based molecular methods can lead to the identification of greater diversity within defined taxa compared with those that are identified by standard culture-based methods and to the identification of novel organisms that were not previously associated with vaginal flora.

Human vaginal flora plays a profound role in reproductive health and disease. However, our primitive understanding of the complex microbial ecosystem of the genital tract greatly hampers our ability to develop appropriate, focused therapies for genital infections. Given the current limitations in our diagnostic abilities, it is naive to assume that we know all of the organisms that are involved in genital tract health. It is likely that the microorganisms that are responsible for reproductive health and disease remain to be discovered. To date, no exhaustive, culture-independent survey has been done of this important microbial community.

The use of conventional culture repeatedly has been found to be unhelpful, because approximately 5% of normal flora is comprised of multiple organisms that are implicated as genital pathogens; thus, their presence alone is insufficient information. In addition, many are extremely difficult or impossible to culture routinely, and there may be organisms that have yet to be detected.[1] The organisms that are associated most often with bacterial vaginosis include anaerobic bacteria, in particular *Bacteroides* spp, *Peptostreptococcus* spp, *Gardnerella vaginalis,* and *Mycoplasma hominis.*[2] The fastidious nature of these organisms makes culture-based methods impractical. This major limitation of culture-based methods has been described as "the great plate count anomaly"[3] because only a small fraction of microorganisms that are present in a population can be cultured. Studies that are based on culture have characterized vaginal lactobacilli[4] or other specific organisms of interest, such as *G vaginalis.*[5] A few new organisms have been recognized,[6-8] but these generally are associated with disease, not with healthy flora. There is a reasonable expectation that organisms that are important to reproductive health may have evaded detection with standard methods.

The development of culture-independent, gene-based methods has facilitated small-scale studies of a wide variety of complex microbial communities.[9] Molecular methods have been applied in previous studies to identify and enumerate vaginal organisms. However, the relatively small scale and often generally descriptive nature of these studies leaves us with a somewhat superficial understanding of vaginal flora.[4,10-12] Recently, results of a larger study demonstrated the potential usefulness of the application of high throughput molecular methods to the characterization of vaginal microflora.[13]

Chaperonin-60 is a molecular chaperone essential for the folding and assembly of proteins and protein complexes in all eubacteria and in the plastids and mitochondria of eukaryotes. The gene encoding chaperonin-60 (*cpn*60) offers several advantages over the widely used 16S recombinant RNA (rRNA) gene as a target for microbial species identification and phylogenetics.[14] A robust molecular method for the identification of microorganisms, which is based on the amplification of a 549– to 567–base pair (bp) portion of the *cpn*60 gene (the "universal target") with universal degenerate polymerase chain reaction (PCR) primers,[15] and the comparison of amplified sequences to a reference database of *cpn*60 sequences has been applied previously to phylogenetic studies,[16] the identification of clinical isolates,[17-19] and studies of the microbial ecologic condition of the animal gastrointestinal tract.[20,21] The *cpn*60 universal target region generally provides more discriminating and phylogenetically informative data than the 16S rRNA target, particularly between closely related species.[16] Sequence variation extends quite uniformly throughout the *cpn*60 coding region, whereas variable regions of 16S rRNA genes are dispersed between regions of highly conserved sequence that result in stable secondary structure and facilitates PCR artifacts. *Cpn*60 genes generally are present in a single copy in prokaryotic genomes, which makes an attractive target for quantitative methods. The relatively small size of the universal target facilitates high throughput sequencing approaches. Finally, a reference database of *cpn*60 sequences is available. [14]

We describe the application of *cpn*60 sequence-based methods to the characterization of vaginal flora of 16 healthy nonpregnant women. The sample size was chosen in consideration of the cost of the sequencing of a large number of clones and our previous experience with *cpn*60 sequence-based microbial ecologic studies. Our previous results demonstrate that the size of the current pilot-scale study will be sufficient to identify potentially interesting targets that can then be monitored in a larger group of subjects.[20,21]

Material and methods

Subjects

This project received University of British Columbia Research Ethics Board approval. Women who attended

an outpatient Sexually Transmitted Diseases clinic, at the British Columbia Centre for Disease Control, Vancouver, British Columbia, who self-identified as not having symptoms were offered enrollment. Written informed consent was obtained; at the time of the standard speculum examination, an additional swab for vaginal secretions was taken from the posterior fornix of the vagina. All women had clinical specimens that were evaluated for bacterial vaginosis by standardized Nugent scoring at the British Columbia Centre for Disease Control clinical laboratory. As part of routine clinical assessments, samples were taken for routine light microscopy assessment for yeast and *Trichomonas vaginalis*, *Chlamydia trachomatis* PCR, and *Neisseria gonorrhoeae* culture.

Total genomic DNA isolation

Each clinical Dacron swab was processed within 24 hours of receipt from the clinic. The contents of each swab were extracted in 800 μL of DNAzol (MRC Inc, Cincinnati, Ohio). The extract was vortexed vigorously, with pulsing for 2 minutes. Ethanol (600 μL) was added, mixed, and incubated at room temperature for 5 minutes. Precipitated nucleic acid was pelleted by centrifugation and washed twice with 1 mL of 75% ethanol. One hundred microliters of 8 mmol/L NaOH was used to solubilize the nucleic acid, which was followed by the addition of 3 μL of 1 mol/L HEPES to neutralize the purified DNA solution.

PCR and creation of PCR product libraries

Each DNA sample (2 μL) was used as template in PCR reactions with 0.5 μg of each universal *cpn*60 PCR primer (H279 5′-GAI III GCI GGI GAY GGI ACI ACI AC-3′ and H280 5′-YKI YKI TCI CCR AAI CCI GGI GCY TT-3′), 50 mmol/L KCl, 10 mmol/L Tris (pH 8.3), 1.5 mmol/L $MgCl_2$, 200 mmol/L of each deoxynucleoside triphosphate, 2 U Taq DNA polymerase in a final volume of 50 μL. After the addition of paraffin oil, PCR amplification with a robocycler (Stratagene, La Jolla, Calif) was carried out for 3 minutes at 95°C for 1 cycle, followed by 40 cycles of 1 minute at 94°C, 2 minutes at 40°C, 5 minutes at 72°C, and completed with 1 cycle of 10 minute at 72°C. PCR products from each template were agarose gel purified and ligated into T-A cloning vector pCR2.1-TOPO (Invitrogen, Carlsbad, Calif). Ligation mixtures were used to transform *Escherichia coli* strain JM109. The 16 resulting libraries were plated on Luria Broth (LB)/ampicillin/X-gal, and 480 white colonies were picked for each library. Colonies were picked into 96-well plates that contained 100 μL of LB with 100 μg/mL ampicillin. After overnight incubation at 37°C, 100 μL of 30% (vol/vol) glycerol was added to each well, and the cultures were stored at −80°C.

Table I Numbers of clones and unique sequences from 16 vaginal flora *cpn*60 libraries

Library	Number of clones sequenced	Number of unique nucleotide sequences
hvf3233	430	5
hvf3238	410	1
hvf3244	436	1
hvf3245	440	3
hvf3246	441	2
hvf3247	439	6
hvf3257	431	9
hvf3258	446	5
hvf3265	434	12
hvf3266	449	1
hvf3267	338	3
hvf3268	386	1
hvf3269	419	1
hvf3271	451	3
hvf3272	466	17
hvf3273	367	4

Template preparation and sequencing

The sequencing template was prepared by the TempliPhi system (Amersham Pharmacia, Piscataway, NJ) at a 1 to 8 recommended reaction scale in 384-well plates (total reaction volume 2.5 μL). Sequencing reactions were performed by direct addition of ET terminator reaction mix (Amersham Pharmacia) to 384-well plates that contained the TempliPhi products. Sequencing reactions were thermocycled according to the manufacturer's recommended protocol. Sequencing reactions were resolved on an ABI PRISM 3730XL DNA Analyzer system (Applied Biosystems, Foster City, Calif) at the McGill University and Genome Québec Innovation Centre.

Sequence data processing and bioinformatics

Raw sequence data was processed using the Phred[22] program, which assigns quality values to the bases and trims poor quality regions. The resulting sequences were clustered on the basis of sequence identity with the d2_cluster.[23] Clusters of identical sequences were assembled, with the use of Phrap,[24] which incorporates the base quality information from Phred. Manual confirmation of contig assembly was done with Gap4 (version 4.6) in the Staden software package (release 2000.0; J. Bonfield, K. Beal, M. Betts, M. Jordan, R. Staden, 2000). Sequence data, template information and similarity results were placed in a mySQL database (mySQL AB, Uppsala, Sweden) for storage and further analysis. Sequence manipulations, such as format changes and amino acid translations, were done with the European Molecular Biology Open Software Suite software.[25] Sequence

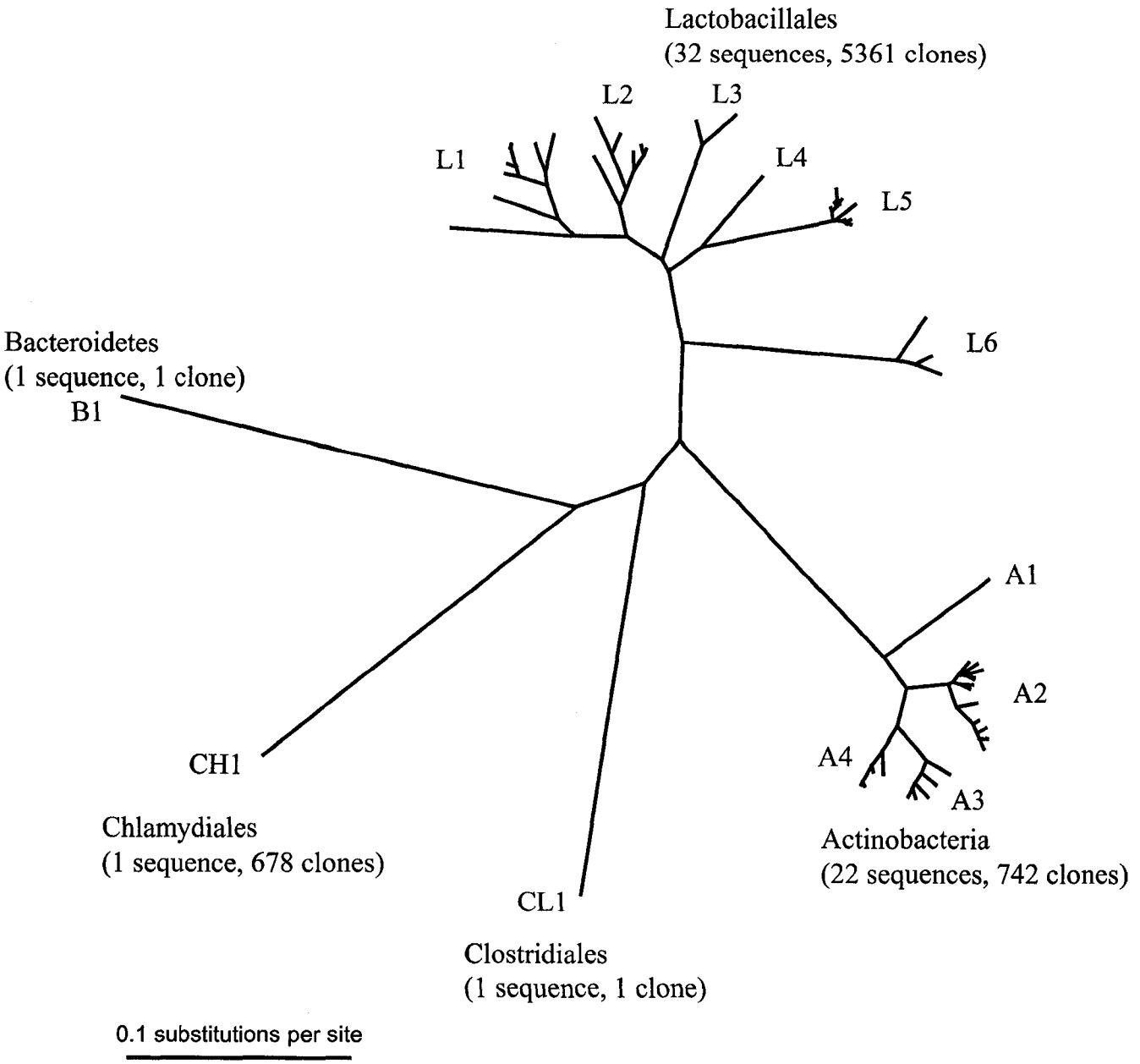

Figure 1 Phylogenetic relationships of unique *cpn*60 sequences from 16 vaginal flora libraries. The tree was calculated with the use of a maximum likelihood distance calculation, followed by neighbor-joining. Phylogenetic clusters L1-L6, A1-A4, B1, CH1, and CL1 are indicated within the major taxonomic groups.

alignments were done using Clustal W.[26] To determine the putative taxonomy of each contig and singleton that arises from the assembly step, each sequence was compared with a reference set of *cpn*60 sequences with the sequence alignment program FASTA.[27]

Phylogenetic analysis was done with programs in the Phylogeny Inference Package software package (version 3.5c; Distributed by the author [Felsenstein J], 1993, Department of Genetics, University of Washington, Seattle, Wash). Specifically, alignments were sampled for bootstrap analysis with *seqboot*; distances were calculated with the maximum likelihood option of *dnadist*. Dendrograms were constructed from distance data by *neighbor*-joining with neighbor. Consensus trees were calculated with *consense*, and branch lengths were superimposed on consensus trees using *fitch*.

Species-specific PCR primer design and amplification of *C psittaci cpn*60

Primers H1520 (5′- GCT CAG GTA GCC ACC ATT TC –3′) and H1521 (5′- GCT AGA AAG GTA TCC

Table II Taxonomic identification of cloned *cpn60* sequences, based on comparison to reference sequence data

Taxonomic group	Nearest reference sequence neighbor	Range of sequence identity (%)	Number of sequences recovered	Number of clones recovered
A1	*G vaginalis* ATCC14018	89	1	1
A2	*G vaginalis* ATCC14018	88-95	12	629
A3	*G vaginalis* ATCC14018	90-100	5	100
A4	*G vaginalis* ATCC14018	94-97	4	11
B1	*Porphyromonas levii* ATCC29147	72	1	1
CH1	*C psittaci* 6BC ATCCVR-125	100	1	678
CL1	*Megasphaera elsdenii* ATCC25940	70	1	1
L1	*L crispatus* A3M75	93-100	8	2393
L2	*L gasseri* ATCC9857	93-97	6	393
L3	*L jensenii* ATCC25258	93-96	2	65
L4	*L gasseri* ATCC9857	92	1	2
L5	*L iners* A3M7	98-99	12	2503
L6	*L buchneri* ATCC4005	83-84	3	5

TCG GTT G –3′) were designed with Oligo6 (Molecular Biology Insights Inc, Cascade, Colo) and Signature Oligo (LifeIntel, Port Moody, BC, Canada). Amplifications were performed in a 50 μL reaction that contained 20 mmol/L Tris-HCl pH 8.0, 50 mmol/L KCl, 1.5 mmol/L $MgCl_2$, 0.2 mmol/L deoxyribonucleoside triphosphate, 20 pmol of H1520, 20 pmol of H1521, and 2 units Taq DNA polymerase. Reactions were incubated at 95°C for 5 minutes, followed by 40 cycles of 30 seconds at 95°C, 30 seconds at 62°C, and 30 seconds at 72°C. A final extension of 10 minutes at 72°C followed the last cycle. Amplifications were performed with a thermocycler instrument (BioRad iCycler; Bio-Rad Laboratories, Hercules, Calif).

Results

Clinical samples

Twenty-three nonpregnant, sexually active women who were being seed for a sexually transmitted infection screening examination, without genital symptoms, were enrolled after written informed consent. Their ages ranged from 19 to 35 years. All of the women underwent the clinic's standardized questionnaire and examination. Two women had abnormal findings on pelvic examination and were excluded immediately. All remaining women tested negative for *N gonorrheae,* and *C trachomatis.* Yeast and *T vaginalis* were not detected by light microscopy in any subjects. One sample was "intermediate" for bacterial vaginosis, with a Nugent's score of 4, and was excluded. Two samples were eliminated because they had been delayed in transit to the research laboratory by >24 hours, and 3 samples did not yield sufficient PCR product in the initial reactions to generate an adequate library. The remaining 16 women for whom samples were used had completely normal flora by all standard clinical testing.

Cpn60 gene sequences amplified from vaginal swabs

High-quality sequence data were obtained for 6869 of the 7680 clones that were picked randomly from 16 libraries. Data from the remaining 811 clones were excluded from the analysis because of incomplete or partially ambiguous sequences. As summarized in Table I, pairwise comparisons of the sequences that were determined for each library indicated that from 1 to 17 different nucleotide sequences were identified in each of the 16 libraries. Pooling of the sequence data from all 16 libraries resulted in the identification of a total of 57 different nucleotide sequences (GenBank accessions AY581720-AY581776). Phylogenetic analysis of the 57 different nucleotide sequences that were recovered from the pooled library data resulted in the identification of 13 distinct clusters of sequences (Figure 1). Clusters were classified on the basis of similarity of the sequences to reference sequence data (Table II). Most clones that were analyzed (5361 clones, representing 32 distinct sequences) fell into clusters L1 to L6. Sequences in L1 to L5 were at least 92% identical to *Lactobacillus* spp. Sequences in L6 were consistent with the Lactobacillales family but had weaker sequence similarity (83%-84%) to any reference *Lactobacillus* sp. Four distinct clusters (A1-A4) of Actinobacteria-like sequences were identified. These sequences (22 sequences, represented by 742 clones) were all at least 88% identical to *G vaginalis* ATCC 14018 (American Type Culture Collection, Manassas, Va). Single sequences were identified with similarity to the Clostridiales family (cluster CL1), the Bacteroidetes family (cluster B1), and the Chlamydiales

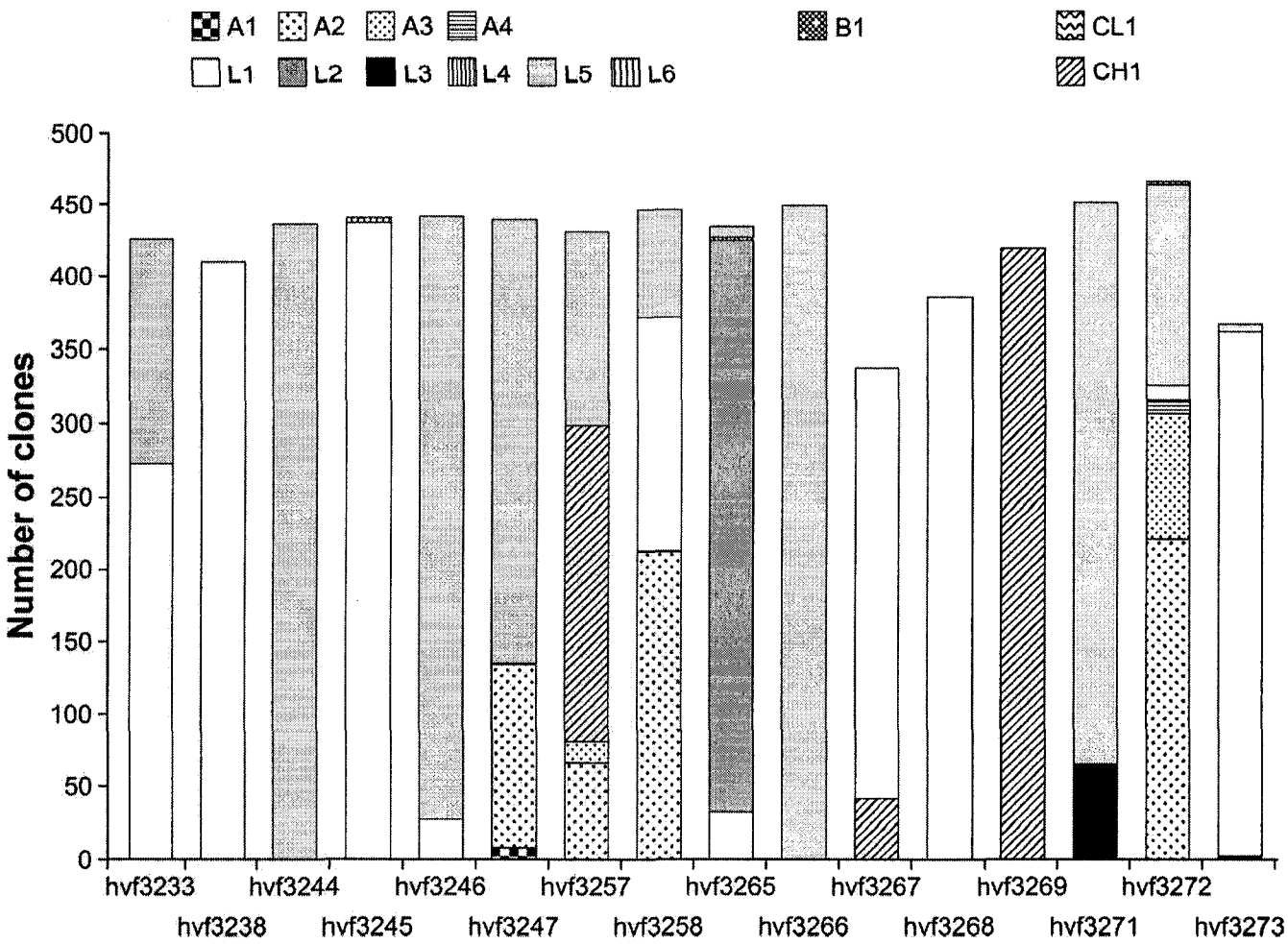

Figure 2 Composition of 16 vaginal flora libraries. The number of clones that corresponded to each phylogenetic group are presented. Phylogenetic groups are described in **Figure 1** and in the text.

(cluster CH1, 100% identical to *Chlamydophila psittaci* ATCC VR-125).

Comparison of sequence profiles of individual libraries

Figure 2 shows the taxonomic composition of the 16 vaginal flora libraries. Nine of the libraries were composed exclusively of *Lactobacillus*-like sequences. Library hvf3273 contained one sequence with similarity to the Bacteroidetes family and one sequence with similarity to the Clostridiales family and was otherwise composed completely of *Lactobacillus*-like sequences. The major *Lactobacillus* constituent in 6 libraries (hvf3233, hvf3238, hvf3245, hvf3267, hvf3268 and hvf3273) was the L1 cluster (93% identical to *L crispatus*). The L5 group (98%-99% identical to *L iners*) was most abundant in hvf3244, hvf3246, hvf3266 and hvf3271. Library hvf3265 was the only library to be dominated by L2 type sequences (93%-97% identical to *L gasseri* ATCC 9857). Libraries hvf3247, hvf3257, hvf3258, and hvf3272 contained sequences in the A1 to A4 clusters (88%-100% identical to *G vaginalis* ATCC 14018). Sequences identical to the type strain of *C psittaci* were identified in libraries hvf3257, hvf3267 and hvf3269.

Variation within *G vaginalis* sequence cluster

Twenty-two sequences (represented by 742 clones) with strong similarity to *G vaginalis* were identified. A phylogenetic analysis of these sequences and other related Actinobacteria showed that these sequences reliably cluster with *G vaginalis* (Figure 3). Pairwise nucleotide sequence identities within this group of sequences ranged from 86% to 99%. A multiple sequence alignment of the 22 *G vaginalis*-like sequences resulted in the identification of 107 positions of difference in the 552-bp alignment (data not shown). Seven of these differences were found in codon position 1; 1 difference was found in position 2, and 99 differences were found in the third codon position.

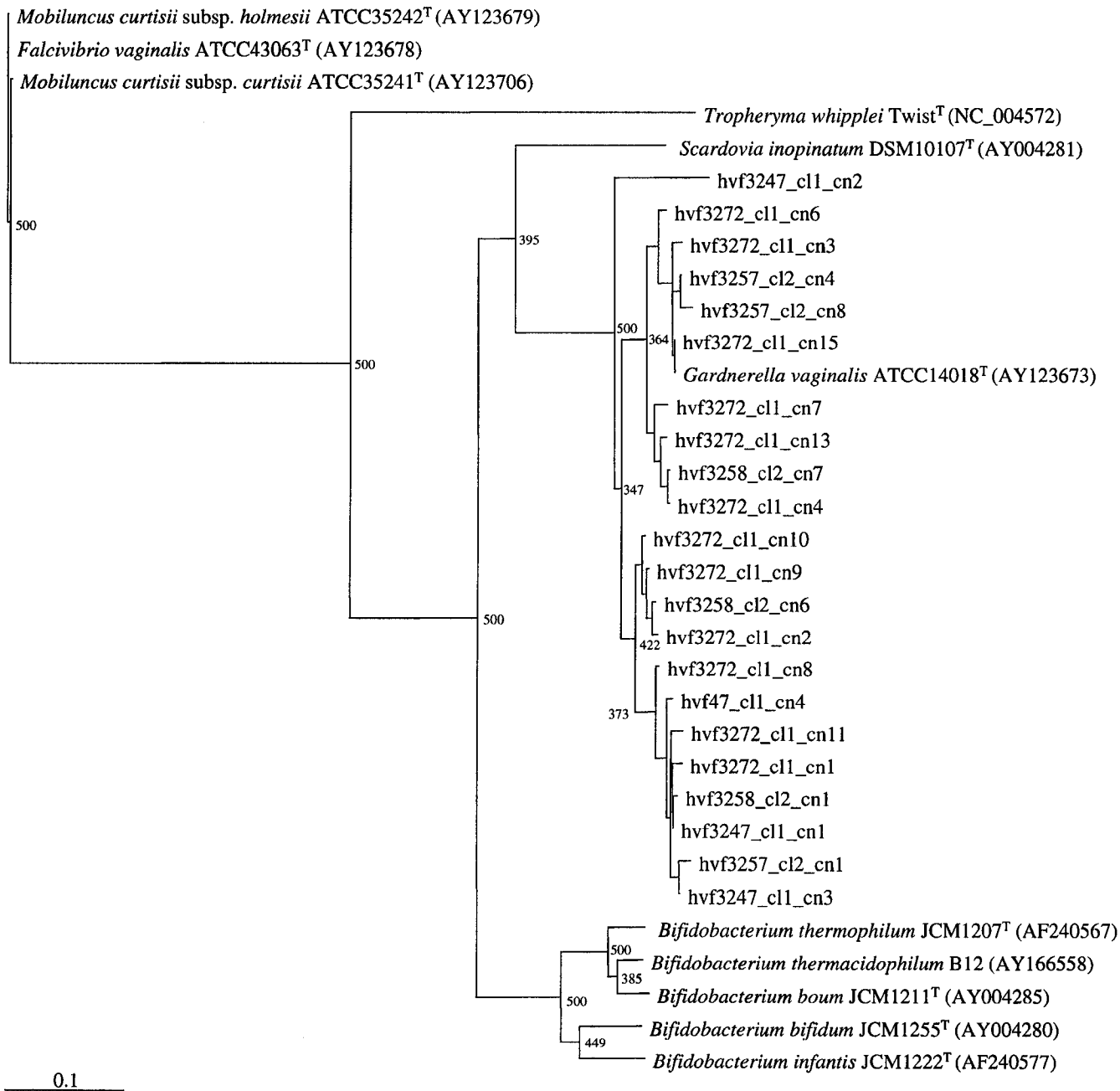

Figure 3 Phylogenetic tree shows the relationships of *Gardnerella vaginalis*–like library sequences to closely related actinobacteria. The tree was calculated from 500 bootstrap iterations with maximum likelihood distance calculation and neighbor-joining. Bootstrap values (of 500) for major branch points in the tree are indicated. *Scale bar* indicates 0.1 substitutions per site.

Specific amplification of *C psittaci*-like sequences

Phylogenetic analysis of the *C psittaci*-like sequence that were derived from the vaginal flora libraries and reference sequence data from additional Chlamydiales family members was performed to confirm that *C psittaci* formed a distinct taxon based on the 555-bp region of *cpn*60 (Figure 4) and could be discriminated from closely related species such as *C abortus* (95% identical to *C psittaci* in the 555-bp amplified region of *cpn*60). Primers for the specific amplification of a 174-bp region of the *C psittaci cpn*60 gene (from positions 160-333 of the 555-bp *cpn*60 universal target) were designed on the basis of a multiple sequence alignment of partial *cpn*60 sequences from *C pneumoniae* J138 (Genbank accession NC_002491), *C muridarum* (NC_002620), *C trachomatis* D/UW-3/CX (NC_000117), *C pneumoniae* CWL029

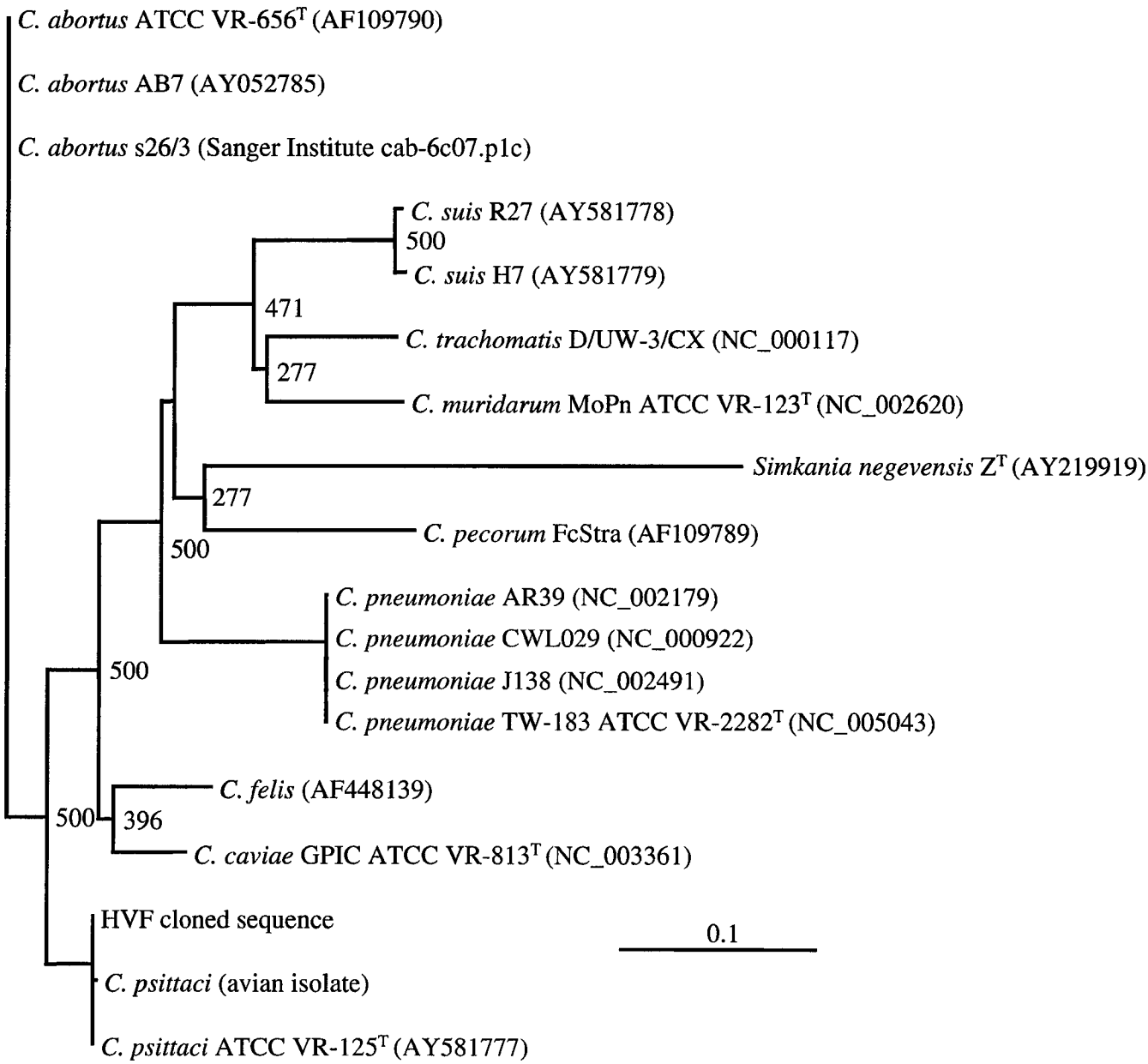

Figure 4 Phylogenetic tree shows the relationships of *Chlamydophila psittaci*–like library sequences to closely related *Chlamydiaceae*. The tree was calculated from 500 bootstrap iterations with maximum likelihood distance calculation and neighbor-joining. Bootstrap values (of 500) for major branch points in the tree are indicated. *Scale bar* indicates 0.1 substitutions per site.

(NC_000922), *C pneumoniae* AR39 (NC_002179), *C felis* FEIS (AF448139), *C pecorum* (AF109789), *C abortus* B577 (AF109790), *C caviae* GPIC (NC_003361), *C abortus* AB7 (AY052785), *C suis* R27 (AY581778), *C suis* H7 (AY581779) and *C psittaci* ATCC VR-125 (AY581777). Figure 5 shows the results of PCR reactions that were performed with *C psittaci*-specific primers on total genomic DNA samples that were used to generate the vaginal flora libraries. The expected product size was obtained from samples 3247, 3267, and 3269. Product was also detected in template 3246. The identity of amplified products from these 4 templates was confirmed by sequence analysis (data not shown). All sequences were 100% identical to library clone sequences.

Comment

Previous studies of the vaginal flora of healthy individuals that were based on culture or sequence-based methods have led to the understanding of this microbial community as relatively homogenous and dominated by a small subset of the *Lactobacillus acidophilus* complex, particularly *L crispatus*, *L gasseri*, *L jensenii*, and *L iners*. The shortcomings of exclusively culture-based studies

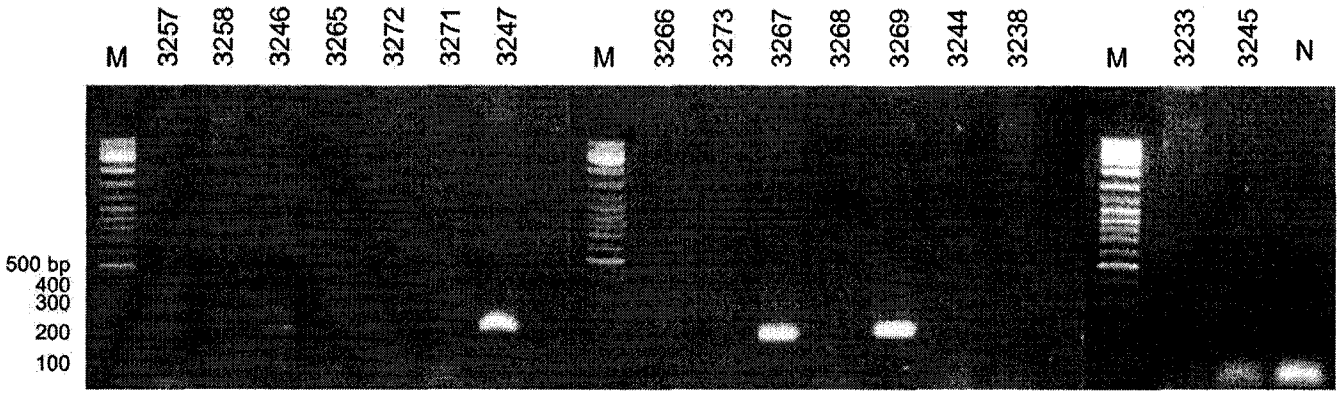

Figure 5 *Chlamydia psittaci*–specific PCR. PCRs performed with *C psittaci*–specific primers on total genomic DNA samples that were used to generate the vaginal flora libraries. A 174-base pair product was obtained from samples 3246, 3247, 3267, and 3269. *Lane N* is a negative control reaction that contains no template DNA. *M*, Size markers.

are illustrated by the case of *L iners*. This organism, unlike other *Lactobacillus* spp, can be cultured only on blood agar and was thus overlooked in most studies of vaginal lactobacilli that rely on Man, Rogosa, and Sharpe agar. It is now apparent that *L iners* is a major component of the vaginal flora.[4]

Consistently with previous findings, we found that most of the *cpn*60 sequence libraries that were constructed in the current study were dominated by sequences with strong similarity to the *L acidophilus* complex, specifically *L crispatus, L gasseri, L jensenii* and *L iners*. Most of the libraries were found to be composed of representatives of 1 or 2 of the *Lactobacillus* clusters, frequently L1 (*L crispatus*) and L5 (*L iners*; Figure 2). A similar result was obtained in a study of the vaginal *Lactobacillus* flora of 23 healthy Swedish women with the use of randomly amplified polymorphic DNA analysis, where most individual samples were found to contain only 1 or 2 randomly amplified polymorphic DNA patterns, which indicated the dominance of 1 or 2 species.[4] The results of a recent 16S rRNA sequence-based study of the vaginal flora of 3 healthy subjects support the observation that healthy vaginal flora is dominated by 1 or 2 *Lactobacillus* spp.[13]

Five of the libraries that we examined contained only 1 sequence each (Table I). We do not suggest that these subjects are colonized only by 1 organism but rather consider that this apparent lack of diversity is the result of a combination of factors that are related to the application of a PCR-based method to a microbial community that is dominated largely by a small number of species. The relative abundances of organisms in any complex microbial community can vary over many orders of magnitude so that, in a total DNA preparation from the community, genomes of the most abundant organisms far outnumber those of rare organisms and will be over represented correspondingly in the PCR product pool. Given that we sequenced only a few hundred clones from each library, it is not surprising that only 1 sequence was recovered in some libraries. In a much more diverse community (pig feces), most sequences were detected at a frequency of approximately 0.1% (1 occurrence in 1125 sequences),[20] a level that would make the sequences likely undetectable in a smaller study. It is also likely that there is bias in the PCR reaction, in which some templates are favored on the basis of composition (especially the guanine-cytosine content) or priming efficiency. To address these issues and identify rarer, potentially unculturable organisms in vaginal flora, technical advances that include subtraction methods and modified PCR protocols are being pursued.

In addition to identifying gross taxonomic clusters, we also identified a large amount of variation within these defined taxa. The 6 identified *Lactobacillus*-like clusters, L1 to L6, each contained from 1 to12 distinct sequences (Table II; Figure 1). This potentially biologically significant "intraspecies" diversity would not be apparent with the use of culture-based methods or molecular methods (such as denaturing gradient gel electrophoresis) in which banding patterns are often identical for closely related species of *Lactobacillus*.[11]

Similar "intraspecies" variation was observed in the *G vaginalis*-like taxa, clusters A1 to A4. *G vaginalis* can be isolated from the vaginal flora of individuals with healthy vaginal ecosystems and individuals who receive a diagnosis with bacterial vaginosis, although the reported proportions of healthy individuals harboring *G vaginalis* varies widely.[5,28,29] Results of previous studies of *G vaginalis* isolates from the vagina suggest that the "biotypes" of *G vaginalis* that are associated bacterial vaginosis are distinct from the *G vaginalis* that are found to varying degrees in healthy individuals.[5] The data presented in Figure 3 certainly support the idea that there is tremendous variability within the *G vaginalis* taxon and that a quantitative assessment of the occurrence of these organisms in healthy and nonhealthy vaginal ecosystems certainly would provide clues to the significance of the variability.

The most surprising finding in the current study was the detection of sequences that are identical to *C psittaci* in 3 of the libraries. Of the 8 characterized serovars of *C psittaci*, serovars A, C, D, and E have been identified as human pathogens.[28] However, *C psittaci* has not been found previously in the vaginal mucosal flora. Two other species of the family Chlamydiaceae have been isolated from human vaginal flora. *Chlamydia trachomatis* causes trachoma, sexually transmitted disease, some types of arthritis, neonatal conjunctivitis, and pneumonia; *Chlamydophila abortus* has been found in sporadic zoonotic infections that cause abortion in women who work with sheep.[30] However, all of these species can be distinguished readily by *cpn*60 sequence (Figure 4). The detection of *C psittaci*-like *cpn*60 sequences raises the question of the role of *C psittaci* in the vagina and also suggests that the application of molecular methods will likely lead to the identification of other organisms that have not been associated previously with vaginal flora in culture-based studies.

This study presents an evaluation of a small number of healthy women and demonstrates the usefulness of *cpn*60 sequence-based methods to enhance detailed the evaluation of human vaginal flora. The application of this method resulted in the identification of "intraspecies" sequence variation that likely would be undetectable with the use of culture-based or sequence-based methods that use the 16S rRNA target. Clearly, further studies of larger populations of women with and without normal flora in concert with technical developments to improve the detection of rarer sequences will be needed to expand our understanding of this complex microbial community. The specific sequence data that were generated in these studies can be used to quantify and monitor individual population members so that their contributions to the function of vaginal microflora can be assessed and understood.

Acknowledgments

We thank Jennifer Town for excellent technical assistance and Tony Rees, RN, for clinical support and subject recruitment.

References

1. Hillier SL. Diagnostic microbiology of bacterial vaginosis. Am J Obstet Gynecol 1993;169:455-9.
2. Thorsen P, Jensen IP, Jeune B, Ebbesen N, Arpi M, Bremmelgaard A, et al. Few microorganisms associated with bacterial vaginosis may constitute the pathologic core: a population-based microbiologic study among 3596 pregnant women. Am J Obstet Gynecol 1998;178:580-7.
3. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 1985;39:321-46.
4. Vasquez A, Jakobsson T, Ahrne S, Forsum U, Molin G. Vaginal lactobacillus flora of healthy Swedish women. J Clin Microbiol 2002;40:2746-9.
5. Aroutcheva AA, Simoes JA, Behbakht K, Faro S. *Gardnerella vaginalis* isolated from patients with bacterial vaginosis and from patients with healthy vaginal ecosystems. Clin Infect Dis 2001;33:1022-7.
6. Rodriguez Jovita M, Collins MD, Sjoden B, Falsen E. Characterization of a novel *Atopobium* isolate from the human vagina: description of *Atopobium vaginae* sp nov. Int J Syst Bacteriol 1999;49:1573-6.
7. Collins MD, Jovita MR, Hutson RA, Ohlen M, Falsen E. *Aerococcus christensenii* sp nov, from the human vagina. Int J Syst Bacteriol 1999;3(49 Pt):1125-8.
8. Shukla SK, Bernard KA, Harney M, Frank DN, Reed KD. *Corynebacterium nigricans* sp nov: proposed name for a black-pigmented corynebacterium species recovered from the human female urogenital tract. J Clin Microbiol 2003;41:4353-8.
9. Morris CE, Bardin M, Berge O, Frey-Klett P, Fromin N, Girardin H, et al. Microbial biodiversity: approaches to experimental design and hypothesis testing in primary scientific literature from 1975 to 1999. Microbiol Mol Biol Rev 2002;66:592-616.
10. Pavlova SI, Kilic AO, Kilic SS, So JS, Nader-Macias ME, Simoes JA, et al. Genetic diversity of vaginal lactobacilli from women in different countries based on 16S rRNA gene sequences. J Appl Microbiol 2002;92:451-9.
11. Burton JP, Reid G. Evaluation of the bacterial vaginal flora of 20 postmenopausal women by direct (Nugent score) and molecular (polymerase chain reaction and denaturing gradient gel electrophoresis) techniques. J Infect Dis 2002;186:1770-80.
12. Tarnberg M, Jakobsson T, Jonasson J, Forsum U. Identification of randomly selected colonies of lactobacilli from normal vaginal fluid by pyrosequencing of the 16S rDNA variable V1 and V3 regions. APMIS 2002;110:802-10.
13. Verhelst R, Verstraelen H, Claeys G, Verschraegen G, Delanghe J, Van SL, et al. Cloning of 16S rRNA genes amplified from normal and disturbed vaginal microflora suggests a strong association between *Atopobium vaginae*, *Gardnerella vaginalis* and bacterial vaginosis. BMC Microbiol 2004;4:16.
14. Hill JE, Penny SL, Crowell KG, Goh SH, Hemmingsen SM. CpnDB: a chaperonin sequence database. Genome Res 2004;14:1669-75.
15. Goh SH, Potter S, Wood JO, Hemmingsen SM, Reynolds RP, Chow AW. HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. J Clin Microbiol 1996;34:818-23.
16. Brousseau R, Hill JE, Prefontaine G, Goh SH, Harel J, Hemmingsen SM. *Streptococcus suis* serotypes characterized by analysis of chaperonin 60 gene sequences. Appl Environ Microbiol 2001;67:4828-33.
17. Goh SH, Santucci Z, Kloos WE, Faltyn M, George CG, Driedger D, et al. Identification of *Staphylococcus* species and subspecies using the chaperonin-60 gene identification method and reverse checkerboard hybridization. J Clin Microbiol 1997;35:3116-21.
18. Goh SH, Driedger D, Gillett S, Low DE, Hemmingsen SM, Amos M, et al. *Streptococcus iniae*, a human and animal pathogen: specific identification by the chaperonin-60 gene identification method. J Clin Microbiol 1998;36:2164-6.
19. Goh SH, Facklam RR, Chang M, Hill JE, Tyrrell GJ, Burns EC, et al. Identification of *Enterococcus* species and phenotypically similar *Lactococcus* and *Vagococcus* species by reverse checkerboard hybridization to chaperonin 60 gene sequences. J Clin Microbiol 2000;38:3953-9.
20. Hill JE, Seipp RP, Betts M, Hawkins L, Van Kessel AG, Crosby WL, et al. Extensive profiling of a complex microbial community by high-throughput sequencing. Appl Environ Microbiol 2002;68:3055-66.
21. Hill JE, Hemmingsen SM, Goldade B, Zijlstra R, Goh SH, Klassen J, et al. Characterization and quantification of ileum microflora of pigs fed corn, wheat or barley-based diets using

chaperonin-60 sequencing and quantitative PCR. Appl Environ Microbiol 2005;71:867-75.
22. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred I: accuracy assessment. Genome Res 1998;8:175-85.
23. Burke J, Davison D, Hide W. D2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res 1999;9:1135-42.
24. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res 1998;8:195-202.
25. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16:276-7.
26. Thompson JD, Higgins DG, Gibson TJ. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673-80.
27. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85:2444-8.
28. Burton JP, Dixon JL, Reid G. Detection of *Bifidobacterium* species and *Gardnerella vaginalis* in the vagina using PCR and denaturing gradient gel electrophoresis (DGGE). Int J Gynaecol Obstet 2003;81:61-3.
29. Totten PA, Amsel R, Hale J, Piot P, Holmes KK. Selective differential human blood bilayer media for isolation of *Gardnerella (Haemophilus) vaginalis*. J Clin Microbiol 1982;15: 141-7.
30. Everett KD, Bush RM, Andersen AA. Emended description of the order chlamydiales, proposal of *Parachlamydiaceae* fam nov and *Simkaniaceae* fam nov, each containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. Int J Syst Bacteriol 1999;49:415-40.

REFERENCES

1. Franca, L. T., Carrilho, E. & Kist, T. B. A review of DNA sequencing techniques. *Q. Rev. Biophys.* **35**, 169-200 (2002).

2. Sanger, F., Nicklen, S. & Coulsen, A. R. DNA sequencing with chain terminating inhibitors. Proc.Natl.Acad.Sci.U.S.A 74, 5463-5467. 1977.
Ref Type: Generic

3. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441-448 (1975).

4. Neela Grani Cooper. DNA Sequencing. Los Alamos Science 20, 151-159. 1992.
Ref Type: Generic

5. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A* **74**, 560-564 (1977).

6. Almira, E. C., Panayotova, N. & Farmerie, W. G. Capillary DNA sequencing: maximizing the sequence output. *J. Biomol. Tech.* **14**, 270-277 (2003).

7. Ahmadian, A. *et al.* Single-nucleotide polymorphism analysis by pyrosequencing. *Anal. Biochem.* **280**, 103-110 (2000).

8. Commercial pyrosequencing www.biotagebio.com. 4-15-0005.
Ref Type: Internet Communication

9. Ronaghi, M. & Elahi, E. Pyrosequencing for microbial typing. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **782**, 67-72 (2002).

10. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630-634 (2000).

11. Carninci, P. *et al.* High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.* **4**, 61-66 (1997).

12. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics.* **37**, 327-336 (1996).

13. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967-974 (1998).

14. Woese, C. R., Maniloff, J. & Zablen, L. B. Phylogenetic analysis of the mycoplasmas. *Proc. Natl. Acad. Sci. U. S. A* **77**, 494-498 (1980).

15. Woese, C. R. *et al.* Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res.* **8**, 2275-2293 (1980).

16. Woese, C. R. *et al.* Conservation of primary structure in 16S ribosomal RNA. *Nature* **254**, 83-86 (1975).

17. Bonen, L. & Doolittle, W. F. Partial sequences of 16S rRNA and the phylogeny of blue-green algae and chloroplasts. *Nature* **261**, 669-673 (1976).

18. Ludwig, W. & Schleifer, K. H. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **15**, 155-173 (1994).

19. Hemmingsen, S. M. *et al.* Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature* **333**, 330-334 (1988).

20. Hill, J. E. Microsporidians lack cpn60 genes. 9-28-2006.
Ref Type: Personal Communication

21. Peyretaillade, E. *et al.* Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. *Mol. Biol. Evol.* **15**, 683-689 (1998).

22. Hill, J. E. *et al.* Biochemical analysis, cpn60 and 16S rDNA sequence data indicate that Streptococcus suis serotypes 32 and 34, isolated from pigs, are Streptococcus orisratti. *Vet. Microbiol.* **107**, 63-69 (2005).

23. Saibil, H. R. & Ranson, N. A. The chaperonin folding machine. *Trends Biochem. Sci.* **27**, 627-632 (2002).

24. Maguire, M., Coates, A. R. & Henderson, B. Chaperonin 60 unfolds its secrets of cellular communication. *Cell Stress. Chaperones.* **7**, 317-329 (2002).

25. Goh, S. H. *et al.* HSP60 gene sequences as universal targets for microbial species identification: studies with coagulase-negative staphylococci. *J. Clin. Microbiol.* **34**, 818-823 (1996).

26. Dumonceaux, T. J. *et al.* Molecular characterization of microbial communities in Canadian pulp and paper activated sludge and quantification of a novel Thiothrix eikelboomii-like bulking filament. *Can. J. Microbiol.* **52**, 494-500 (2006).

27. Hill, J. E. *et al.* Characterization of vaginal microflora of healthy, nonpregnant women by chaperonin-60 sequence-based methods. *Am. J. Obstet. Gynecol.* **193**, 682-692 (2005).

28. Hill, J. E., Penny, S. L., Crowell, K. G., Goh, S. H. & Hemmingsen, S. M. cpnDB: a chaperonin sequence database. *Genome Res.* **14**, 1669-1675 (2004).

29. Fischer, G., Rocha, E. P., Brunet, F., Vergassola, M. & Dujon, B. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS. Genet.* **2**, e32 (2006).

30. Dumonceaux, T. J., Hill, J. E., Hemmingsen, S. M. & Van Kessel, A. G. Characterization of intestinal microbiota and response to dietary virginiamycin supplementation in the broiler chicken. *Appl. Environ. Microbiol.* **72**, 2815-2823 (2006).

31. Dumonceaux, T. J. *et al.* Enumeration of specific bacterial populations in complex intestinal communities using quantitative PCR based on the chaperonin-60 target. *J. Microbiol. Methods* **64**, 46-62 (2006).

32. Hill, J. E. *et al.* Comparison of ileum microflora of pigs fed corn-, wheat-, or barley-based diets by chaperonin-60 sequencing and quantitative PCR. *Appl. Environ. Microbiol.* **71**, 867-875 (2005).

33. Hill, J. E. *et al.* Extensive profiling of a complex microbial community by high-throughput sequencing. *Appl. Environ. Microbiol.* **68**, 3055-3066 (2002).

34. NCBI. Genome Completion Status. 2007.
Ref Type: Unpublished Work

35. NCBI. GenBank feature definition. 2007.
Ref Type: Unpublished Work

36. Guarro, J., GeneJ & Stchigel, A. M. Developments in fungal taxonomy. *Clin. Microbiol. Rev.* **12**, 454-500 (1999).

37. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC. Genomics* **7**, 57 (2006).

38. Everett, K. D., Bush, R. M. & Andersen, A. A. Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. *Int. J. Syst. Bacteriol.* **49 Pt 2**, 415-440 (1999).

39. Gall, J. G. Chromosome structure and the C-value paradox. *J. Cell Biol.* **91**, 3s-14s (1981).

40. MIRSKY, A. E. & RIS, H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol* **34**, 451-462 (1951).

41. Cavalier-Smith, T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot. (Lond)* **95**, 147-175 (2005).

42. Adami, C., Ofria, C. & Collier, T. C. Evolution of biological complexity. *Proc. Natl. Acad. Sci. U. S. A* **97**, 4463-4468 (2000).

43. Ofria, C., Adami, C. & Collier, T. C. Selective pressures on genomes in molecular evolution. *J. Theor. Biol.* **222**, 477-483 (2003).

44. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414**, 450-453 (2001).

45. Keeling, P. J. & Fast, N. M. Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.* **56**, 93-116 (2002).

46. Gill, E. E. & Fast, N. M. Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes. *Gene* **375**, 103-109 (2006).

47. Peyretaillade, E. *et al.* Microsporidian Encephalitozoon cuniculi, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. *Nucleic Acids Res.* **26**, 3513-3520 (1998).

48. Keeling, P. J. & McFadden, G. I. Origins of microsporidia. *Trends Microbiol.* **6**, 19-23 (1998).

49. Cavalier-Smith, T. Archamoebae: the ancestral eukaryotes? *Biosystems* **25**, 25-38 (1991).

50. Williams, B. A., Hirt, R. P., Lucocq, J. M. & Embley, T. M. A mitochondrial remnant in the microsporidian Trachipleistophora hominis. *Nature* **418**, 865-869 (2002).

51. Hirt, R. P., Healy, B., Vossbrinck, C. R., Canning, E. U. & Embley, T. M. A mitochondrial Hsp70 orthologue in Vairimorpha necatrix: molecular evidence that microsporidia once contained mitochondria. *Curr. Biol.* **7**, 995-998 (1997).

52. Germot, A., Philippe, H. & Le, G. H. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in Nosema locustae. *Mol. Biochem. Parasitol.* **87**, 159-168 (1997).

53. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611-1618 (2002).

54. Fedorov, A. & Hartman, H. What does the microsporidian E. cuniculi tell us about the origin of the eukaryotic cell? *J. Mol. Evol.* **59**, 695-702 (2004).

55. GenBank. Genome sequencing status. GenBank . 5-8-2007.
Ref Type: Electronic Citation

56. Johnson, L., Cao, X. & Jacobsen, S. Interplay between two epigenetic marks. DNA methylation and histone H3 lysine 9 methylation. *Curr. Biol.* **12**, 1360-1367 (2002).

57. Zhang, X. *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126**, 1189-1201 (2006).

58. Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74**, 481-514 (2005).

59. Chan, S. W., Henderson, I. R. & Jacobsen, S. E. Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nat. Rev. Genet.* **6**, 351-360 (2005).

60. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**, 931-949 (2006).

61. Schaefer, M. & Lyko, F. DNA methylation with a sting: an active DNA methylation system in the honeybee. *Bioessays* **29**, 208-211 (2007).

62. Feil, R. & Berger, F. Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet.* **23**, 192-199 (2007).

63. Mhanni, A. A. & McGowan, R. A. Global changes in genomic methylation levels during early development of the zebrafish embryo. *Dev. Genes Evol.* **214**, 412-417 (2004).

64. Haig, D. & Westoby, M. An earlier formulation of the genetic conflict hypothesis of genomic imprinting. *Nat. Genet.* **38**, 271 (2006).

65. Haig, D. & Westoby, M. Parent-Specific gene expression and the triploid endosperm. 134, 147-155. 1989.
Ref Type: Generic

66. McGowan, R. A. & Martin, C. C. DNA methylation and genome imprinting in the zebrafish, Danio rerio: some evolutionary ramifications. *Biochem. Cell Biol.* **75**, 499-506 (1997).

67. Martienssen, R. A. & Colot, V. DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**, 1070-1074 (2001).

68. Lio, P. & Vannucci, M. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics.* **16**, 932-940 (2000).

69. Aissani, B. *et al.* The compositional properties of human genes. *J. Mol. Evol.* **32**, 493-503 (1991).

70. Mouchiroud, D. *et al.* The distribution of genes in the human genome. *Gene.* **100:181-7.**, 181-187 (1991).

71. Tsonis, A. A., Elsner, J. B. & Tsonis, P. A. Periodicity in DNA coding sequences: implications in gene evolution. *J. Theor. Biol.* **151**, 323-331 (1991).

72. Tsonis, A. A., Kumar, P., Elsner, J. B. & Tsonis, P. A. Wavelet analysis of DNA sequences. *PHYSICAL. REVIEW. E. STATISTICAL. PHYSICS. , PLASMAS. , FLUIDS, AND RELATED. INTERDISCIPLINARY. TOPICS.* **53**, 1828-1834 (1996).

73. Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C. & Marcourt, L. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J. Theor. Biol.* **206**, 323-326 (2000).

74. Lio, P. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics.* **19**, 2-9 (2003).

75. Audit, B. *et al.* Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys. Rev. Lett.* **86**, 2471-2474 (2001).

76. Audit, B., Vaillant, C., Arneodo, A., d'Aubenton-Carafa, Y. & Thermes, C. Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.* **316**, 903-918 (2002).

77. Wen, S. Y. & Zhang, C. T. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochem. Biophys. Res. Commun.* **311**, 215-222 (2003).

78. Links, M. G. Application of Wavelet analysis to the Arabidopsis thaliana genome. 2002. University of Saskatchewan.
Ref Type: Thesis/Dissertation

79. Schaefer, M. & Lyko, F. DNA methylation with a sting: an active DNA methylation system in the honeybee. *Bioessays* **29**, 208-211 (2007).

80. Grewal, S. I. & Jia, S. Heterochromatin revisited. *Nat. Rev. Genet.* **8**, 35-46 (2007).

81. Schaefer, M. & Lyko, F. DNA methylation with a sting: an active DNA methylation system in the honeybee. *Bioessays* **29**, 208-211 (2007).

82. Schaefer, M. & Lyko, F. DNA methylation with a sting: an active DNA methylation system in the honeybee. *Bioessays* **29**, 208-211 (2007).

83. Schaefer, M. & Lyko, F. DNA methylation with a sting: an active DNA methylation system in the honeybee. *Bioessays* **29**, 208-211 (2007).

84. Field, L. M., Lyko, F., Mandrioli, M. & Prantera, G. DNA methylation in insects. *Insect Mol. Biol.* **13**, 109-115 (2004).

85. Marhold, J. *et al.* Conservation of DNA methylation in dipteran insects. *Insect Mol. Biol.* **13**, 117-123 (2004).

86. Xiao, W. *et al.* DNA methylation is critical for Arabidopsis embryogenesis and seed viability. *Plant Cell* **18**, 805-814 (2006).

87. Sato, S. & Tabata, S. [The complete genome sequence of Arabidopsis thaliana]. *Tanpakushitsu Kakusan Koso* **46**, 61-65 (2001).

88. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science* **301**, 653-657 (2003).

89. Martienssen, R. A. Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 2021-2026 (1998).

90. Salk Institute Genomic Analysis Laboratory. Salk Institute Genomic Analysis Laboratory . 9-5-2001.
Ref Type: Electronic Citation

91. Kong, H. *et al.* Patterns of gene duplication in the plant SKP1 gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J.* ., (2007).

92. Risseeuw, E. P. *et al.* Protein interaction analysis of SCF ubiquitin E3 ligase subunits from Arabidopsis. *Plant J.* **34**, 753-767 (2003).

93. Ni, W. *et al.* Regulation of flower development in Arabidopsis by SCF complexes. *Plant Physiol* **134**, 1574-1585 (2004).

94. Yang, M., Hu, Y., Lodhi, M., McCombie, W. R. & Ma, H. The Arabidopsis SKP1-LIKE1 gene is essential for male meiosis and may control homologue separation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11416-11421 (1999).

95. Liu, F. *et al.* The ASK1 and ASK2 genes are essential for Arabidopsis early development. *Plant Cell* **16**, 5-20 (2004).

96. Zhao, D., Han, T., Risseeuw, E., Crosby, W. L. & Ma, H. Conservation and divergence of ASK1 and ASK2 gene functions during male meiosis in Arabidopsis thaliana. *Plant Mol. Biol.* **53**, 163-173 (2003).

97. Wang, Y. & Yang, M. The ARABIDOPSIS SKP1-LIKE1 (ASK1) protein acts predominately from leptotene to pachytene and represses homologous recombination in male meiosis. *Planta* **223**, 613-617 (2006).

98. Dellaert, L. M. W. Eceriferum mutants in Arabidopsis thaliana (L.) Heynh: I. Induction by X-rays and Fast Neutrons. *AIS* **16**, 1-9 (1979).

99. Dellaert, L. M. W. Comparison of X-Ray and Fast Neutron-induced mutant spectra. Experiments in Arabidopsis Thaliana (l.) Heynh. *AIS* **18**, 16-36 (1981).

100. Mednik, I. G. On methods evaluating the frequncies of induced mutations in Arabidopsis based on embry-test data. *AIS* **26**, 67-72 (1988).

101. Ramulu, K. S. & Sybenga, J. Comparison of Fast Neutrons nd X-rays in respect to genetic effects accompanying induced chromosome abberrations: Induction and analysis of translocations in Arabidopsis thaliana. *AIS* **16**, 27-34 (1979).

102. Li, X. *et al.* A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant J.* **27**, 235-242 (2001).

103. Li, X. & Zhang, Y. Reverse genetics by fast neutron mutagenesis in higher plants. *Funct. Integr. Genomics* **2**, 254-258 (2002).

104. Shalitin, D. *et al.* Regulation of Arabidopsis cryptochrome 2 by blue-light-dependent phosphorylation. *Nature.* **417**, 763-767 (2002).

105. Ingram, G. C. *et al.* Parallels between UNUSUAL FLORAL ORGANS and FIMBRIATA, genes controlling flower development in Arabidopsis and Antirrhinum. *Plant Cell* **7**, 1501-1510 (1995).

106. Wilkinson, M. D. & Haughn, G. W. UNUSUAL FLORAL ORGANS Controls Meristem Identity and Organ Primordia Fate in Arabidopsis. *Plant Cell* **7**, 1485-1499 (1995).

107. Samach, A. *et al.* The UNUSUAL FLORAL ORGANS gene of Arabidopsis thaliana is an F-box protein required for normal patterning and growth in the floral meristem. *Plant J.* **20**, 433-445 (1999).

108. An, Y. Q. *et al.* Strong, constitutive expression of the Arabidopsis ACT2/ACT8 actin subclass in vegetative tissues. *Plant J.* **10**, 107-121 (1996).

109. Han, L., Mason, M., Risseeuw, E. P., Crosby, W. L. & Somers, D. E. Formation of an SCF(ZTL) complex is required for proper regulation of circadian timing. *Plant J.* **40**, 291-301 (2004).

110. Somers, D. E., Kim, W. Y. & Geng, R. The F-box protein ZEITLUPE confers dosage-dependent control on the circadian clock, photomorphogenesis, and flowering time. *Plant Cell.* **16**, 769-782 (2004).

111. Somers, D. E. Clock-associated genes in Arabidopsis: a family affair. *Philos. Trans. R. Soc. Lond B Biol. Sci.* **356**, 1745-1753 (2001).

112. Somers, D. E., Schultz, T. F., Milnamow, M. & Kay, S. A. ZEITLUPE encodes a novel clock-associated PAS protein from Arabidopsis. *Cell.* **101**, 319-329 (2000).

113. Wang, Y. *et al.* A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance in vitro. *Nucleic Acids Res.* **32**, 1197-1207 (2004).

114. Paez, J. G. *et al.* Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71 (2004).

115. Amersham Biosciences. Use of amplified DNA as a substrate for uniplex PCR. 1-4. 4-1-2003.
Ref Type: Generic

116. V & P Scientific. Data for Disposable Pin Tools. 8-31-2006.
Ref Type: Personal Communication

117. Zhang, X. *et al.* Freezing-sensitive tomato has a functional CBF cold response pathway, but a CBF regulon that differs from that of freezing-tolerant Arabidopsis. *Plant J.* **39**, 905-919 (2004).

118. Zarka, D. G., Vogel, J. T., Cook, D. & Thomashow, M. F. Cold induction of Arabidopsis CBF genes involves multiple ICE (inducer of CBF expression) promoter elements and a cold-regulatory circuit that is desensitized by low temperature. *Plant Physiol* **133**, 910-918 (2003).

119. Gilmour, S. J. *et al.* Low temperature regulation of the Arabidopsis CBF family of AP2 transcriptional activators as an early step in cold-induced COR gene expression. *Plant J.* **16**, 433-442 (1998).

120. Fowler, S. & Thomashow, M. F. Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *Plant Cell* **14**, 1675-1690 (2002).

121. Cook, D., Fowler, S., Fiehn, O. & Thomashow, M. F. A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A* **101**, 15243-15248 (2004).

122. Martienssen, R. A., Doerge, R. W. & Colot, V. Epigenomic mapping in Arabidopsis using tiling microarrays. *Chromosome. Res.* **13**, 299-308 (2005).

123. Lippman, Z., Gendrel, A. V., Colot, V. & Martienssen, R. Profiling DNA methylation patterns using genomic tiling microarrays. *Nat. Methods* **2**, 219-224 (2005).

124. Gendrel, A. V., Lippman, Z., Martienssen, R. & Colot, V. Profiling histone modification patterns in plants using genomic tiling microarrays. *Nat. Methods* **2**, 213-218 (2005).

125. Houde, M. *et al.* Wheat EST resources for functional genomics of abiotic stress. *BMC. Genomics* **7**, 149 (2006).

126. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. dbEST--database for "expressed sequence tags". *Nat. Genet.* **4**, 332-333 (1993).

127. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

VITA AUCTORIS

Matthew Graham Links was born March 15, 1978 in Saskatoon Saskatchewan Canada. He graduated from Walter Murray High School's Advanced Program in 1996. From there he received a B.Sc. in Biochemistry while simultaneously completing all the requirements of a B.Sc. in Computer Science at the University of Saskatchewan. Matthew has worked in academic research roles with the National Research Council, Genome Prairie, and University of Saskatchewan prior to pursuing his Masters degree in Biological Sciences at the University of Windsor. He plans to graduate fall 2007. Presently Matthew is the lead Bioinformatician for the Agriculture and Agri-Food Canada research station in Saskatoon.