Bioinformatics approaches to analysing RNA mediated regulation of gene expression

Dissertation

zur Erlangung des akademischen Grades

"doctor rerum naturalium" (Dr. rer. nat.)

eingereicht im

Institut für Biochemie und Biologie an der

Mathematisch-Naturwissenschaftlichen Fakultät der

Universität Potsdam

LIAM CHILDS

Arbeitsgruppe Bioinformatik

Max-Plank-Institut für Molekulare Pflanzenphysiologie

Potsdam, den 13.07.2009

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2010/4128/
URN urn:nbn:de:kobv:517-opus-41284
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-41284

Acknowledgements

Dirk Walther is certainly one of the best supervisors I've ever had. He gave me the freedom to pursue my own ideas and goals whilst providing all the necessary help and support when needed. I have a sneaking suspicion that, if I didn't play the drums, I may not have been chosen to fill this PhD position. Nevertheless, studying under Dirk and playing in a kick-arse rock/pop/punk/funk/blues/jazz band together has made the past three years, although so very far from home, a very happy and fulfilling experience. To work at such an institute, in the company of great scientific minds, has helped my scientific thinking through a colossal improvement. I still have far to go, but I start my post-doctoral efforts from a very solid basis.

I would also like to thank my official university supervisor Joachim Selbig. Thanks to the whole of AG Bioinformatics; Christian Schudoma, Zoran Nikoloski and Patrick May for all the help provided, and also to Manuela Hische, Sergio Grimbs, Sebastian Klie, Georg Basler, Jan-Ole Christian, Tanja Gärtner, Henning Redestig, Pawel Durek, Jan Hummel and Xiaoliang Sun for the good times and great company.

Thanks also to Thomas Altmann and Hanna Witucka-Wall for their help and for my involvement in their project along with Karl Schmid and Torsten Günther.

Quick thanks to Marco Ende and Peter Krüger for technical support.

Thanks to two very special friends Jan Lisec and Sandra Witt for the company, both wonderfully immature and entertaining, at lunchtime and during the jamming sessions.

I thank my family for their support; Mum, Dad and Owen, and Sean for shedding a little brilliance on my projects whenever I was stuck. 16,110 km is no small distance and I miss you all very much. And finally, to my girlfriend Julia, my thanks to you can not be written down on paper ;).

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und unter Verwendung keiner anderen als den von mir angegebenen Quellen und Hilfsmitteln verfasst habe.

Ferner erkläre ich, dass ich bisher weder an der Universität Potsdam noch anderweitig versucht habe, eine Dissertation einzureichen oder mich einer Doktorprüfung zu unterziehen.

Potsdam, 13 July 2009

_____

Liam Childs

Contents

Abstract

Until relatively recently, the realm of RNA was considered to be almost completely explored. RNA molecules were thought to be involved purely in protein translation in the forms of mRNA, tRNA and rRNA. The discovery of ribozymes, miRNA and riboswitches soon revealed that RNA apparently carries broader functionality than realised, plays a more active role than a passive messenger in protein synthesis, and a large number of new non-coding RNA (ncRNA) families were discovered not long after. The new RNA world is growing fast and many of the new RNA functions are shown to be regulatory in nature. Non-coding RNA and RNA-related phenomena are under high scrutiny and new experimental, as well as bioinformatics tools and methods, are constantly being developed to completely explore RNA functionality and their role in the regulation of genes.

Currently, there exist several bioinformatics tools that can predict RNA function. These tools often leverage homology at either the sequence or structural levels to infer classifications through similarity. While many of these tools are shown to have high prediction performance, a common limitation for the majority of them is the inability to discover novel RNA function. We sought to address these problems by implementing methods that do not require sequence homology to pre-existing ncRNA families.

Two methods, one experimental and one bioinformatics, are presented in this thesis. The experimental method was part of a larger genotyping effort for *Arabidopsis thaliana*. In the first method, approximately one million genetic markers were discovered using tiling arrays and associated with various phenotypic and metabolic traits. Markers with significant associations within non-coding genomic regions were then characterised using GRAPPLE, a non-coding RNA prediction tool developed in the bioinformatics method.

The bioinformatics approach to ncRNA classification was developed based on graph theoretic properties. By using abstract representations of predicted RNA structure, a fast and accurate algorithm capable of predicting whether a given RNA molecule has a function and, if so, the possible Rfam family it could belong to was developed. The development process is considered in light of user accessibility, usability and scalability.

The methods developed in the thesis will form the basis for the creation of further analytical tools and experimental investigation, the details of which are presented in the final chapter.

Lay Abstract

The genome can be considered the blueprint for an organism. Composed of DNA, it harbours all organism-specific instructions for the synthesis of all structural components and their associated functions. The role of carriers of actual molecular structure and functions was believed to be exclusively assumed by proteins encoded in particular segments of the genome, the genes. In the process of converting the information stored genes into functional proteins, RNA – a third major molecule class – was discovered early on to act a messenger by copying the genomic information and relaying it to the protein-synthesizing machinery. Furthermore, RNA molecules were identified to assist in the assembly of amino acids into native proteins. For a long time, these - rather passive - roles were thought to be the sole purpose of RNA. However, in recent years, new discoveries have led to a radical revision of this view. First, RNA molecules with catalytic functions - thought to be the exclusive domain of proteins - were discovered. Then, scientists realized that much more of the genomic sequence is transcribed into RNA molecules than there are proteins in cells begging the question what the function of all these molecules are. Furthermore, very short and altogether new types of RNA molecules seemingly playing a critical role in orchestrating cellular processes were discovered. Thus, RNA has become a central research topic in molecular biology, even to the extent that some researcher dub cells as “RNA machines”.

This thesis aims to contribute towards our understanding of RNA-related phenomena by applying Bioinformatics means. First, we performed a genome-wide screen to identify sites at which the chemical composition of DNA (the genotype) critically influences phenotypic traits (the phenotype) of the model plant *Arabidopsis thaliana*. Whole genome hybridisation arrays were used and an informatics strategy developed, to identify polymorphic sites from hybridisation to genomic DNA. Following this approach, not only were genotype-phenotype associations discovered across the entire Arabidopsis genome, but also regions not currently known to encode proteins, thus representing candidate sites for novel RNA functional molecules. By statistically associating them with phenotypic traits, clues as to their particular functions were obtained. Furthermore, these candidate regions were subjected to a novel RNA-function classification prediction method developed as part of this thesis.

While determining the chemical structure (the sequence) of candidate RNA molecules is relatively straightforward, the elucidation of its structure-function relationship is much more challenging. Towards this end, we devised and implemented a novel algorithmic approach to

predict the structural and, thereby, functional class of RNA molecules. In this algorithm, the concept of treating RNA molecule structures as graphs was introduced. We demonstrate that this abstraction of the actual structure leads to meaningful results that may greatly assist in the characterization of novel RNA molecules. Furthermore, by using graph-theoretic properties as descriptors of structure, we indentified particular structural features of RNA molecules that may determine their function, thus providing new insights into the structure-function relationships of RNA. The method (termed Grapple) has been made available to the scientific community as a web-based service.

RNA has taken centre stage in molecular biology research and novel discoveries can be expected to further solidify the central role of RNA in the origin and support of life on earth. As illustrated by this thesis, Bioinformatics methods will continue to play an essential role in these discoveries.

Zusammenfassung

Das Genom eines Organismus enthält alle Informationen für die Synthese aller strukturellen Komponenten und deren jeweiligen Funktionen. Lange Zeit wurde angenommen, dass Proteine, die auf definierten Abschnitten auf dem Genom – den Genen – kodiert werden, die alleinigen Träger der molekularen - und vor allem katalytischen - Funktionen sind. Im Prozess der Umsetzung der genetischen Information von Genen in die Funktion von Proteinen wurden RNA Moleküle als weitere zentrale Molekülklasse identifiziert. Sie fungieren dabei als Botenmoleküle (mRNA) und unterstützen als Trägermoleküle (in Form von tRNA) die Zusammenfügung der einzelnen Aminosäurebausteine zu nativen Proteine. Diese eher passiven Funktionen wurden lange als die einzigen Funktionen von RNA Molekülen angenommen. Jedoch führten neue Entdeckungen zu einer radikalen Neubewertung der Rolle von RNA. So wurden RNA-Moleküle mit katalytischen Eigenschaften entdeckt, sogenannte Ribozyme. Weiterhin wurde festgestellt, dass über proteinkodierende Abschnitte hinaus, weit mehr genomische Sequenzbereiche abgelesen und in RNA Moleküle transkribiert werden als angenommen. Darüber hinaus wurden sehr kleine und neuartige RNA Moleküle identifiziert, die entscheidend bei der Koordinierung der Genexpression beteiligt sind. Diese Entdeckungen rückten RNA als Molekülklasse in den Mittelpunkt moderner molekularbiologischen Forschung und führten zu einer Neubewertung ihrer funktionellen Rolle.

Die vorliegende Promotionsarbeit versucht mit Hilfe bioinformatorischer Methoden einen Beitrag zum Verständnis RNA-bezogener Phänomene zu leisten. Zunächst wurde eine genomweite Suche nach Abschnitten im Genom der Modellpflanze *Arabidopsis thaliana* vorgenommen, deren veränderte chemische Struktur (dem Genotyp) die Ausprägung ausgewählter Merkmale (dem Phänotyp) entscheidend beeinflusst. Dabei wurden sogenannte Ganz-Genom Hybridisierungschips eingesetzt und eine bioinformatische Strategie entwickelt, Veränderungen der chemischen Struktur (Polymorphismen) anhand der veränderten Bindung von genomischer DNA aus verschiedenen Arabidopsis Kultivaren an definierte Proben auf dem Chip zu detektieren. In dieser Suche wurden nicht nur systematisch Genotyp-Phänotyp Assoziationen entdeckt, sondern dabei auch Bereiche identifiziert, die bisher nicht als proteinkodierende Abschnitte annotiert sind, aber dennoch die Ausprägung eines konkreten Merkmals zu bestimmen scheinen. Diese Bereiche wurden desweiteren auf mögliche neue RNA Moleküle untersucht, die in diesen Abschnitten kodiert sein könnten. Hierbei wurde ein

neuer Algorithmus eingesetzt, der ebenfalls als Teil der vorliegenden Arbeit entwickelt wurde.

Während es zum Standardrepertoire der Molekularbiologen gehört, die chemische Struktur (die Sequenz) eines RNA Moleküls zu bestimmen, ist die Aufklärung sowohl der Struktur als auch der konkreten Funktion des Moleküls weitaus schwieriger. Zu diesem Zweck wurde in dieser Arbeit ein neuer algorithmischer Ansatz entwickelt, der mittels Computermethoden eine Zuordnung von RNA Molekülen zu bestimmten Funktionsklassen gestattet. Hierbei wurde das Konzept der Beschreibung von RNA-Sekundärstrukturen als Graphen genutzt. Es konnte gezeigt werden, dass diese Abstraktion von der konkreten Struktur zu nützlichen Aussagen zur Funktion führt. Des weiteren konnte demonstriert werden, dass graphen-theoretisch abgeleitete Merkmale von RNA-Molekülen einen neuen Zugang zum Verständnis der Struktur-Funktionsbeziehungen ermöglichen. Die entwickelte Methode (Grapple) wurde als web-basierte Anwendung der wissenschaftlichen Welt zur Verfügung gestellt.

RNA hat sich als ein zentraler Forschungsgegenstand der Molekularbiologie etabliert und neue Entdeckungen können erwartet werden, die die zentrale Rolle von RNA bei der Entstehung und Aufrechterhaltung des Lebens auf der Erde weiter untermauern. Bioinformatische Methoden werden dabei weiterhin eine essentielle Rolle spielen.

Chapter 1.

Introduction

Until relatively recently, the RNA molecule was thought to be a simple "disposable copy" of genomic DNA that provided a middle step between the genome and the translation machinery in a cell (Crick, 1958). With the discovery of ribozymes (Altman et al., 1982; Kruger et al., 1982) and the lack of amino acids at the active site of bacterial ribosomes (Nissen et al., 2000), the accepted roles of RNA broadened to include functions once thought solely to be in the domain of proteins. More recently, additional regulatory roles have been discovered, such as microRNAs (miRNAs) and riboswitches. These new discoveries, combined with recent large scale transcriptomic analyses, which reveal that larger portions of many genomes are transcribed than previously thought (Claverie, 2005; Manak et al., 2006; Weinstock, 2007), has propelled the study of non-coding RNA (ncRNA) onto the scientific scene, transforming it from a curiosity to a central research topic.

The discovery of new ncRNA, the particular biological role they assume and their contribution toward cell regulatory networks remains largely unexplored and new ncRNA and ncRNA functions are being frequently discovered. There are many efforts underway to understand ncRNA and what they can do through the use of both experimental methods and bioinformatics. The development of novel approaches to investigate ncRNA has been pivotal in discovering its capabilities and many new algorithms, such as covariance models (Eddy and Durbin, 1994), have become first point-of-call tools in ncRNA research.

1.1. What is RNA?

RNA is a biopolymer consisting of a chain of nucleotide monomers. Each nucleotide consists of a nitrogenous base, a ribose sugar, and a phosphate. The ribose sugar and the phosphate form the backbone. The carbon atoms in the sugar are numbered from 1' through to 5' and the phosphate binds the 3' carbon of one sugar in a chain with the 5' carbon of the next. Transcription and translation travel along the RNA strand from the nucleotide with the free 5' carbon to the nucleotide with the free 3' carbon. There are four possible nitrogenous bases; adenine (A), cytosine (C), guanine (G) and uracil (U). These bases are able to form base-pairs through hydrogen bonds. The two most common base pairs are A and U, which form two hydrogen bonds with each other, and G and C, which form three. Other base pair combinations are also possible but are much rarer. RNA is very similar to DNA but is usually present natively as a single strand whereas DNA forms a more stable double strand. Due to this, RNA nucleotides are free to form bonds with other nucleotides on the same molecule and are indeed likely to do so, thereby creating substantial structural diversity.

1.1. The Central Dogma

Early understanding of RNA lies within Crick's proposal of the Central Dogma (Crick, 1958), a framework that indicates the direction of sequence information flow among the three main biopolymers; DNA, RNA and protein. The initial version of the Central Dogma stated that sequence information can only flow from DNA to DNA (replication), DNA to RNA (transcription) and RNA to protein (translation). These were known as the general transfers. This version was based upon the understanding that RNA had only three primary families; messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA).

mRNA acts as an intermediate between DNA and protein. A gene encoded in the DNA strand is first copied into a single, short RNA strand, the mRNA. This strand is then passed from the 5' end to the 3' end through the translational machinery of the cell (the ribosome) and a protein is synthesised according to the sequence of amino acids encoded in nucleotide triplets on the mRNA strand.

tRNA performs the function of decoding the codon triplets that comprise a gene. This is enabled by the anticodon, which is a reverse complement nucleotide triplet of the codon to be recognised. For each different amino acid there is at least one tRNA capable of recognising and binding both the amino acid and the codons that encode it. A protein called aminoacyl tRNA synthetase first recognises the molecular shape of the amino acid and the anticodon on

the tRNA and binds the two together. The bound molecules are then recruited by the ribosome, where the anticodon of the tRNA is used to recognise the codon on the mRNA.

rRNA comprises a major component of the ribosome and performs the critical role of catalysing the ligation of the amino acids into long chains of polypeptides. Each codon on the mRNA is fed into the ribosome along with the matching tRNA. The amino acids bound to the tRNAs are then chemically joined together and thus attached to the growing amino acid chain.

Later, the Central Dogma was revised to include special transfers. The special transfers stipulate that sequence information can flow from RNA to DNA, RNA to RNA, and (under special laboratory conditions) DNA direct to protein. The revisions to the Central Dogma were made in light of the discovery of more modes of information transfer that are mainly virus related (Baltimore, 1970; Temin and Mizutani, 1970). The remaining three possible transfers (protein to DNA, protein to RNA, protein to protein) are termed the unknown transfers and have not yet been observed. Under the Central Dogma, RNA plays a relatively simple role. However, recent discoveries reveal that RNA can perform roles that are much more complex and could also shed light upon the early history of RNA and life itself.

1.2. The RNA world hypothesis

The phrase “The RNA World” was coined by Walther Gilbert in 1986 and referred to the primordial beginnings of life where RNA was hypothesised to be both information storage and catalyst for all the necessary reactions required for early life (Gilbert, 1986). RNA has four critical properties that make it a potential candidate for the original molecule of life.

1. RNA could potentially form under early Earth conditions.
2. RNA can store information.
3. RNA is self replicating.
4. RNA can have catalytic and enzymatic properties.

The famous Miller-Urey experiments (Miller and Urey, 1959) simulated a potential early Earth atmosphere by mixing methane, hydrogen and ammonia in a closed environment and passed electrical currents through the mixture (representing the lightning present in the early Earth atmosphere) to produce basic organic molecules. In another chamber containing water, these simple organic molecules interacted to produce more complex molecules such as amino acids and other organic acids. The experiments never actually created RNA, and only very recently has a process been discovered that could possibly produce RNA under early Earth

conditions (Powner et al., 2009). Although there are serious doubts about whether the Miller-Urey experiments truly did replicate early Earth atmospheric conditions, they certainly did demonstrate that complex organic molecules are able to form through purely natural means.

The Earth is estimated to be about 4.5 billion years old (Dalrymple, 2001) and some estimates place the beginning of life at 3.8 billion years ago (Schopf et al., 2002), just 700 million years after the formation of the Earth. If RNA were able to form at this time, it would provide an ideal molecule for evolution to work upon. RNA is both able to store and replicate information, meaning that the genetic information held by one generation can be passed onto the next and any mutations introduced, through physical or chemical means, would also be inherited. Competition for resources such as external nucleic acids, and perhaps amino acids to stabilise RNA molecules, would provide a mechanism for early natural selection and thus kick start the evolution of early biopolymers into the complex cellular machinery is observed today (Hanczyc and Szostak, 2004; Szostak, 2009).

We are only now beginning to truly understand the functional and catalytic properties of RNA that are, perhaps, the final piece of the puzzle needed for early life on Earth. Theoretically, the ability to form complex molecular shapes paves the way for RNA to perform any function that proteins are able to, provided that the required molecular shape can be formed. The discovery of ribozymes and riboswitches (Guerrier-Takada et al., 1983; Kruger et al., 1982; Mironov et al., 2002; Nahvi et al., 2002; Winkler et al., 2002; Winkler et al., 2002) show that this is a tantalising possibility. It has been shown that the catalytic sites of bacterial ribosomes are 100% RNA and are able to function perfectly without the presence of any amino acids (Nissen et al., 2000), and the ability for scientists to synthesise RNA enzymes that can perform better than their protein counterparts lends strength to this argument (Bashkin et al., 1995). Throughout time, the preferred molecule for information storage has been superseded by the more stable DNA and many structural and functional roles have been superseded by the more flexible proteins, but perhaps some of the functions established by RNA almost 4 billion years ago survived till today.

1.3. Discovery of new ncRNA

We may now be discovering some of these potentially ancient RNA functions. With the explosion of different ncRNA types discovered in an assortment of different species, the need to manage all this data has arisen. By far the largest and most comprehensive database is Rfam (Griffiths-Jones et al., 2003). At the time of writing, Rfam contained 1,372 families of

ncRNA and 1,148,236 ncRNA sequences across 1,138 species. Rfam is heavily biased towards viruses and bacteria with 531 and 515 species of each respectively. Archaea have and Eukaryota both have 46 species each.

Each of the families of ncRNA presented in this section show a progression of ncRNA functionality from simple sequence targeting to a fully independent enzyme, thus demonstrating the ability for RNA to fulfil the role of the original biopolymer, and represents some of the RNA families available in the Rfam database. miRNA and snoRNA show the incorporation of ncRNA into protein machinery for the targeting of sequences through antisense complementarity. In these two families, the ncRNA itself does not appear to do any catalytic activity. Riboswitches takes us one step further in that direction by showing that ncRNA is able to bind and target molecules other than nucleic acid sequences. Such binding is no longer reliant on base pair complementarity and instead depends on the molecular surface presented by the folded ncRNA sequence. Ribozymes, both artificial and natural, make the final step by showing ncRNA as a catalytic agent capable of function without the need for proteins to perform structural or co-enzymatic activities.

1.3.1. miRNA

miRNA was first discovered in 1993 in *Caenorhabditis elegans* (Lee et al., 1993). The investigators discovered that the early larval development gene *lin-4,* was responsible for regulating the gene *lin-14,* was unlikely to be protein coding and acted instead through antisense complementarity. After this discovery, miRNA (as they came to be called) were discovered in a huge variety of different organisms ranging from plants and animals to bacteria hinting at an early evolutionary origin (Figure 1.1).

The miRNA machinery potentially evolved from a cellular defence mechanism targeting double stranded viral RNA and now plays a large variety of roles in gene regulation (Lu et al., 2008). They have been implicated in mRNA cleavage, inhibition of protein synthesis, degradation of target mRNA and degradation of viral RNA.

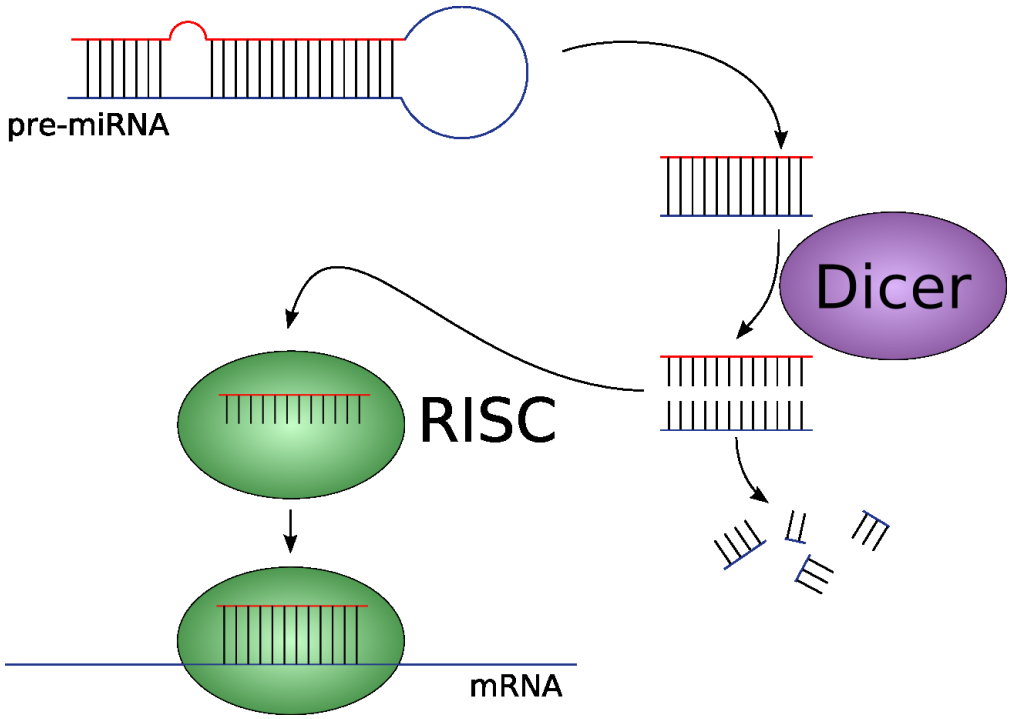

Figure 1.1: Synthesis and mechanism of miRNA

Double stranded RNA, either viral or generated through the expression of genomic anti-sense complimentary sequences, are digested into small ~21 nucleotide fragments by a protein called Dicer. A single strand from the fragment is then bound to a protein complex called RNA-induced Silencing Complex (RISC) that then targets any other RNA molecule which shows complementarity with the bound fragment. The behaviour upon binding to the target can vary.

After the initial discovery of miRNA, a variety of experimental and bioinformatics methods have been developed to aid in further discovery. Because the fully mature miRNA sequences are not generated through random digestion and the position of the cleavage sites show remarkable conservation, in some experimental methods, this conservation has been exploited to detect potential miRNA (Friedlander et al., 2008). Other bioinformatics tools seek sequences that show homology and anti-sense complementarity with other sequences in the same genome (Lu et al., 2006).

1.3.2. snoRNA

Small nucleolar RNA (snoRNA) are a class of small RNA that guide the chemical modifications, such as methylation and pseudouridylation, of rRNA, tRNA and other RNA genes. They bind to proteins that carry out the modifications to form small nucleolar ribonucleoproteins (snoRNP). To bind with the target RNA, they contain a 10-20 nucleotide antisense sequence. There are two different classes of snoRNA currently identified. C/D box snoRNAs contain two short conserved sequence motifs; C (UGAUGA) and D (CUGA)

located near the 5' and 3' ends of the snoRNA respectively. H/ACA box snoRNAs have a common secondary structure consisting of two hairpins and two single stranded regions termed a hairpin-hinge-hairpin-tail structure. They also have two conserved sequences, H (consensus ANANNA) and ACA (ACA). As can be inferred by the number of ambiguous nucleotides in the H box of the H/ACA snoRNA, the sequence is not well conserved and the structure of the folded RNA is probably more important. These motifs, combined with the use of Hidden Markov Models (described later), form the basis for further discovery of snoRNA (Chen and Wu, 2009).

1.3.3. Riboswitches

Riboswitches were initially discovered in 2002 in *Escherichia coli* (Mironov et al., 2002; Nahvi et al., 2002; Winkler et al., 2002; Winkler et al., 2002). The investigators were intrigued by reports that the 5' untranslated region (5'UTR) of a transport protein for cobalamin (*btuB*) had important genetic control structures embedded within them, but did not appear to be under the regulatory control of protein factors. Further investigation revealed that the translation levels of *btuB* appeared to respond to different levels of cobalamin concentration and that cobalamin bound directly to the mRNA.

Riboswitches are clear evidence that the structure of a single strand of folded RNA presents an interaction surface, much like a protein, that can bind metabolites and other molecules. They are composed of two parts; the aptamer and the expression platform. The aptamer is usually highly conserved and binds directly to the metabolite while the expression platform undergoes a structural change that regulates the translation of the protein (Figure 1.2).

Riboswitches can be discovered through homology although there are more sophisticated methods for riboswitch discovery. Covariance models (described later under the introductory section 1.5.5) can be used to search sequences and genomes for currently discovered riboswitches and there exist some bioinformatics methods that attempt to predict riboswitches by finding characteristics of the expression platform mechanism such as overlapping stems (Freyhult et al., 2007). Most methods of *ab initio* riboswitch discovery are generally unsuccessful and further research is required to improve the prediction accuracy.

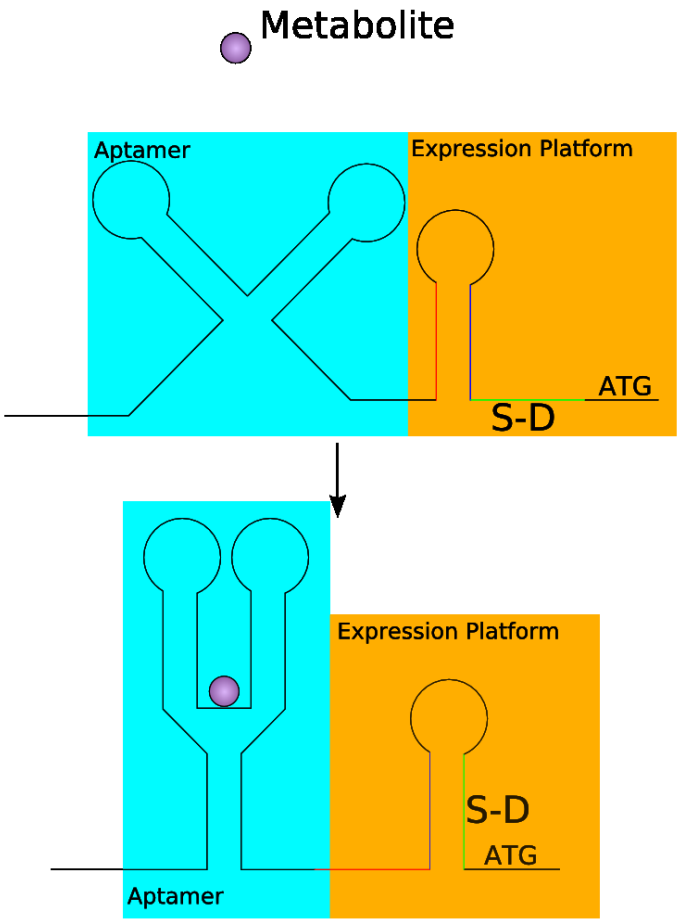

Figure 1.2: TPP riboswitch mechanism

Riboswitches act in a variety of ways. Presented in this diagram is the action of the TPP riboswitch. The aptamer is initially unbound and the expression platform folds into an energetically stable stem that exposes the Shine-Dalgarno (shown in green) sequence. Upon binding with TPP, the aptamer changes into a conformation that pulls apart the stem shown in red and blue allowing another stable hairpin to form that hides the Shine-Dalgarno sequence. This prevents further translation of this mRNA transcript.

1.3.4. Ribozymes

Ribozymes were initially discovered in 1981 by Cech et al. (Cech et al., 1981) with the observation that precursor rRNA could splice out an intron in the absence of proteins. The discovery of ribozymes lent the first solid support for the RNA world hypothesis. Ribozymes are actually quite rare leading to the suggestion that their roles have been taken over by protein counterparts. However, the few that do remain take part in critical biological processes which also hints at ancient origins. The ribosome itself is a large RNA machine and does not necessarily require the presence of amino acids to function (Nissen et al., 2000).

1.3.5. Artificial ncRNA

Recent discoveries of RNA function tie directly into the artificial synthesis of RNA specifically designed for binding small molecules or other functions. One such system is called Systematic Evolution of Ligands by Exponential Enrichment (SELEX). The SELEX process (Tuerk and Gold, 1990) begins with a very large library of randomly generated RNA sequences of fixed length. The sequences are exposed to the target (ligand), which may be a protein or small molecule. Sequences that do not bind are removed and the most tightly binding RNA sequence is identified. Some evolutionary methods exist that introduce random mutations and selection over several iterations to try and improve various aspects of sequence such as target specificity and binding strength.

Ribozymes mimics that are simple to synthesise and perform faster than their natural counterparts have been designed and tested (Bashkin et al., 1995). Although initially designed from RNA, such ribozymes are now created from DNA and incorporate a metal-based inorganic catalyst. Currently designed to target HIV mRNA with high specificity, they provide potential for cheap chemotherapy drugs without significant side effects.

1.4. Arabidopsis

The experimental data used within the experimental part of this thesis comes from the plant model organism *Arabidopsis thaliana*. *A. thaliana* itself has little economic value, but is used to research dicotyledon plants for a number of reasons including a short generation time, large seed production, convenient size, a relatively small and fully sequenced genome and the existence of well established transformation protocols. *A. thaliana* also has a wide variety of different accessions that were critical for the association mapping that is presented in Chapter 2.

1.4.1. ncRNA in Arabidopsis

Large scale annotation projects for mouse (FANTOM) (Claverie, 2005), human (ENCODE) (Weinstock, 2007) and fruit fly (Manak et al., 2006) have revealed that large, non-coding portions of these genomes are transcribed and these transcribed regions are potentially a rich source of undiscovered ncRNA. In plants, massively parallel signature sequencing (MPSS) of *A. thaliana* flowers and seedlings has discovered over 75,000 unique potential small RNAs (Lu et al., 2005). Aside from the usual mRNA, tRNA and rRNA, some of the newer families of ncRNA have also been predicted and observed in *A. thaliana* such as miRNA, spliceosomal RNA, snoRNA, riboswitches and signal recognition particle RNA (SRP).

Similar studies into other organisms reveal a surprisingly large number of new ncRNA genes (Eddy, 2001).

These ncRNA in *A. thaliana* have been discovered primarily by using a variety of bioinformatics means that rely on sequence homology, such as the snoRNAs that have been discovered in *A. thaliana* (Barneche et al., 2001; Qu et al., 2001). By searching the genome for complementarity to rRNA, snoRNAs were predicted that were eventually confirmed experimentally. miRNA were also discovered using a variety of homology means such as looking for the conserved segments in multiple alignments (Lu et al., 2006).

The use of sequence homology based means places two major restrictions on the discovery of ncRNA genes.

1. *Expanding known ncRNA families.* In the cases where ncRNA function relies on molecular structure, homology methods may not be able to provide complete coverage. For example, even within the same species, different tRNA shows a huge variety of different sequences and yet they all still perform the same function. If only homology were used in the identification of tRNA, then not all tRNA could be discovered. To address this, methods that rely on structural homology have been developed.
2. *Discovering new ncRNA families.* When using sequence homology methods, only known and documented ncRNA genes can be discovered and any novel ncRNA will be missed. Sequence homology methods may be able to discover some ncRNA families that do share high homology or rely on base pair complementarity for their function; however, they will miss ncRNA genes that do not share significant homology. Methods that address this issue will somehow need to find properties that are intrinsic to all functional ncRNA, if indeed they exist.

These issues have been partially addressed before with methods such as ALIDOT and PFRALI, which are used to detect homology on the structural level (Hofacker et al., 1998; Hofacker and Stadler, 1999). However, they still require multiple sequence alignments as a starting point and their computational requirements are prohibitive. For our approach we combine bioinformatics and experimental methods designed to potentially identify both novel and currently annotated ncRNA genes in different organisms and in particular *A. thaliana*. The experimental method is able to perform without relying on sequence homology. The bioinformatics method does not require sequence homology for the prediction of whether an

ncRNA is functional, but does use structural similarities to assign Rfam families. In tandem they are potentially able to identify novel functional ncRNA.

1.5. Key concepts

1.5.1. Folding RNA

The molecular shape of a folded single stranded RNA is believed to be crucial in determining the function. 3D RNA molecular structure is not yet able to be reliably predicted and most algorithms go only as far as secondary RNA structure. Although not a perfect representation of the 3D structure, secondary structure is likely to provide valuable information about ncRNA function. RNA folding algorithms are critical for the investigations presented in Chapter 3.

At a simplistic level of understanding, RNA folding is a problem of finding the structure for a single strand of RNA with the maximum possible number of base pairs, as this conformation can be assumed to correspond with the global free energy minimum, a state that functional folded proteins are thought to assume. The main complexity in RNA folding algorithms comes from the intricate scoring systems used, but the underlying algorithm in many tools is a fairly simple process and is implemented using top-down dynamic programming. Dynamic programming is a method of solving problems that exhibit the properties of overlapping sub-problems and optimal substructure. In the case of RNA, the entire RNA sequence to be folded can be broken down into sub-sequences wherein the optimal substructures can be found. By finding the optimal structure for all subsequences starting from a lengths of 0 and 1 (where there can be no structure), the algorithm can calculate the optimal sub-scores for increasingly large sequences until the optimal structure for the entire sequence, based on a particular scoring system, is found (Figure 1.3).

To find the structure with the maximum number of base-pairs, a simple scoring system of +1 per base pair and 0 for anything else is applied. For a sequence of length N, the optimal score for a subsequence ($S(i,j)$) is derived from the optimal scores of the smaller subsequences. For any subsequence in the range $i..j$, there are four possible ways that a structure of nested base-pairs can be defined.

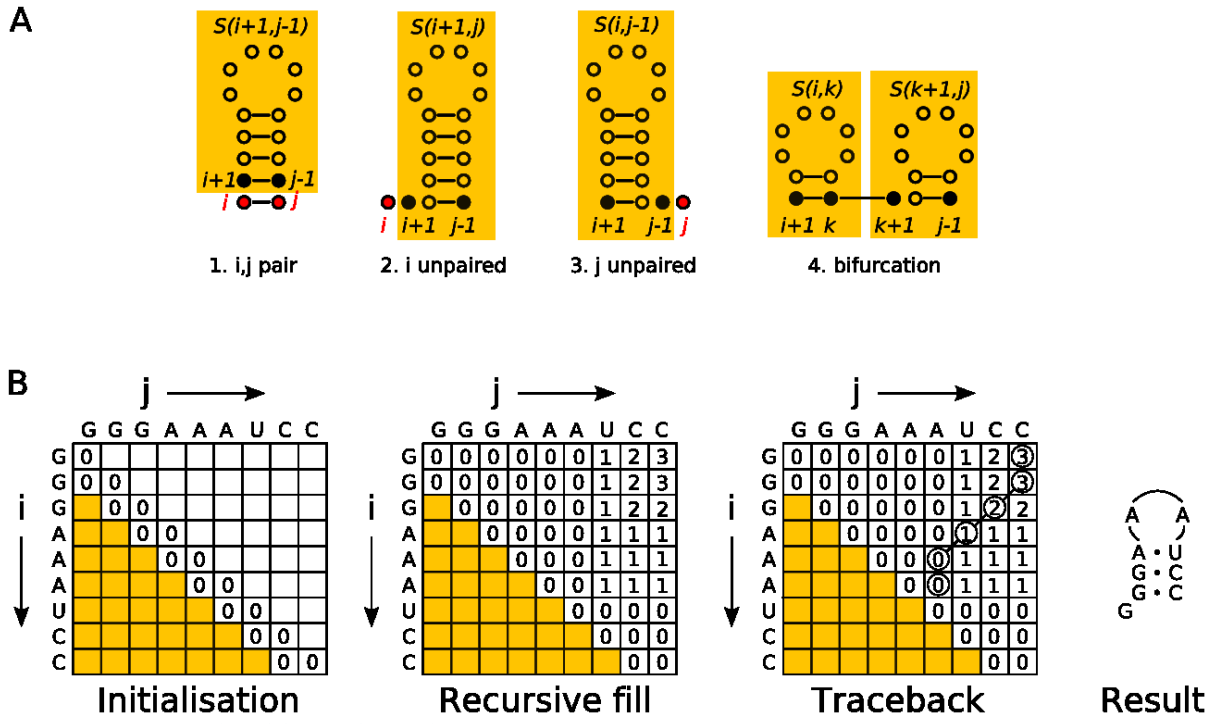

Figure 1.3: RNA folding algorithm Reproduced from (Eddy, 2004)

__A:__ The four possible cases considered by the dynamic programming algorithm. Red dots correspond with the nucleotides currently under consideration. Previously calculated optimal sub-sequences are highlighted in yellow. __B:__ The three major steps of the dynamic programming algorithm are demonstrated. The initial step initialises on the diagonal. The optimal score for all sequences that are one or two nucleotides long is 0. The algorithm then fills in the cells for longer subsequences from three till the length of the whole sequence in the upper right hand corner. The path taken by the algorithm is then traced backwards to the diagonal to produce the optimal structure for the sequence.

1. i, j are paired and added to the substructure $i+1..j-1$.
2. i is unpaired and added to the substructure $i+1..j$.
3. j is unpaired and added to the substructure $i..j+1$.
4. i,j are paired, but not to each other. Here the algorithm bifurcates and the optimal scores for the substructures $i..k$ and $k+1..j$ are added to each other.

For each of the first three cases, the nucleotides i and j are independent of the optimal substructure in the range $i+1..j-1$. Conversely, the optimal substructure in the range $i+1..j-1$ is independent from the bases i and j whether they are paired or not. Therefore, $S(i,j)$ in case 1 is $S(i+1,j-1)$ plus one, if i and j can base pair. In case 4, the two subsequences $i..k$ and $k+1..j$ are independent of each other and the score $S(i,j)$ is the sum of $S(i,k)$ and $S(k+1,j)$. Mathematically, these recursive cases can be calculated according to the formula:

$$S(i,j) = max \begin{cases} S(i+1, j-1) + 1 \\ S(i+1, j) \\ S(i, j-1) \\ max_{i<k<j} S(i,k) + S(k+1, j) \end{cases}$$

To implement this algorithm, an i by j matrix representing the scores for all possible subsequences is created. The matrix is initialised on the diagonal for subsequences of length 0 or 1 that have no structure *i.e.*, $S(i,i) = S(i,i-1) = 0$. The algorithm then work outwards on larger sequences, according to the rules defined earlier, until the upper-right hand corner is reached. This corner is the optimal score for the complete sequence $S(1,N)$. Finally, a path is traced back from the top right corner towards the diagonal following the choices that produced the optimal scores. Each vertical or horizontal step produces an unpaired base and each diagonal step a base pair. Using this scheme, only the optimal scores for the range $i..j$ need to be calculated and stored, which, for a sequence of length n, uses memory proportional to n^2 and a running time proportional to n^3.

This particular scoring system (the Nussinov scoring system) will predict the structure with the maximum number of base-pairs although it may not be the most thermodynamically stable structure. Due to stronger bonding between G-C base pairs than A-T base pairs, more thermodynamically stable structure could be expected to favour more G-C base-pairs and fewer A-T base-pairs. Thus, different scoring systems need to be used to determine the most thermodynamically stable structures. Typically, they seek to minimise the minimum free energy (MFE) produced through bond formation.

To better calculate ncRNA structures, synthetically constructed oligonucleotides are melted and the energy released when a base-pair breaks is calculated. By using these energies to form the foundation for a scoring system, more accurate predictions can be made. However, even using experimentally determined energies, folding tools are still too often wrong. The example presented in Figure 1.4 shows the true and predicted fold of a tRNA molecule using RNAFOLD.

Even with the perfect scoring system, these tools still lack the ability to generate more complex folds. The dynamic programming algorithm presented above is unable to predict pseudoknots, which have been experimentally determined. A pseudoknot is a folded structure with base-pairs that are not nested within other base pairs. Examples include “kissing hairpins”.

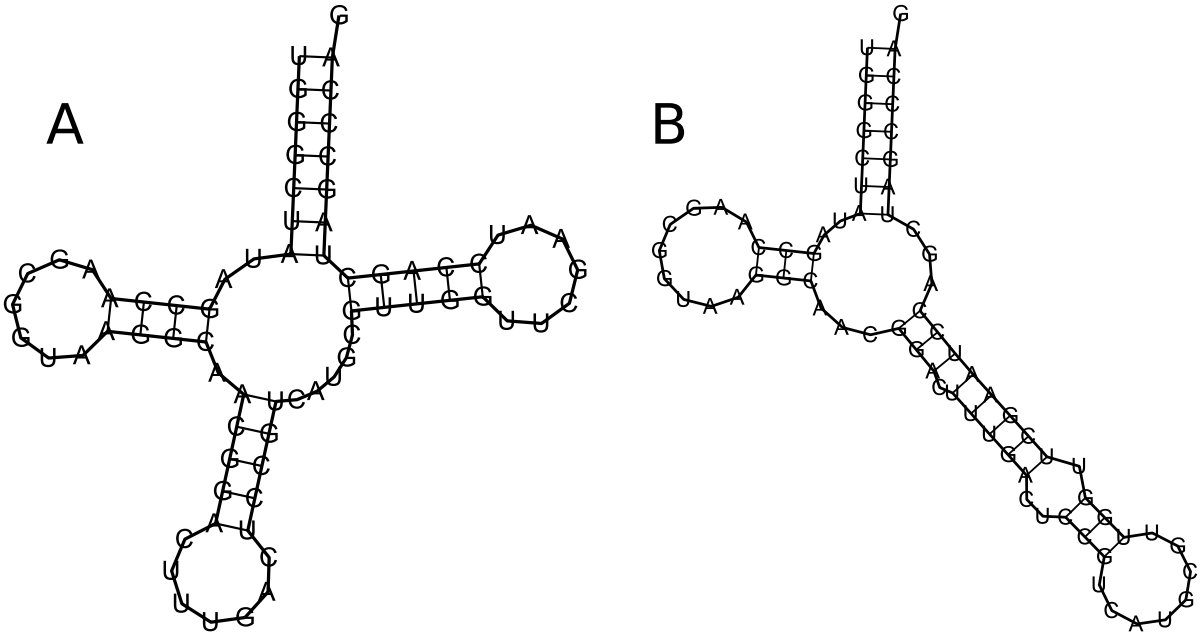

Figure 1.4: Imperfect folding

The structure labelled 'A' shows the correct fold that this particular tRNA sequence should assume. Folding programs still have problems predicting the correct structure as is shown by RNAFOLD's attempt labelled 'B'.

1.5.2. Graphs and graph theory

Graphs are a conceptual construct that show the relationships in a system. A graph is often depicted as a series of scattered dots (nodes/vertices) that are connected by lines (edges). The vertices can be connected to several edges and represent concepts or objects while the edges must be connected to only two vertices and represent the relationship between the vertices. Edges that represent a directional relationship between two vertices are represented as arrows. The vertex at the tail of the arrow is called the parent and the vertex at the head is called the child. Such graphs are called "directed graphs". A directed graph without cycles is called a "directed acyclic graph". A directed acyclic graph where each vertex is the child of only one parent is called a tree.

The structure of a graph can be employed to define and analyse different properties that could reflect the characteristics of the process or entity modelled by the graph. A property can be defined on the level of graph constituents (*i.e.*, vertices and edges) or on the level of the graph itself. Furthermore, computing a property may require limited or full knowledge of the graph. Based on these two criteria (level of detail and required knowledge of the graph), graph-theoretic properties may be classified into local (using limited knowledge of the graph and referring to a graph's constituent), local-global (using full knowledge of the graph and

referring to a graph's constituent), and global (using full knowledge of the graph and referring to the graph itself).

As a novel means for RNA function prediction, we implement an algorithm in Chapter 3 that uses graph theory as a means of producing the features required by support vector machines (described later) that reflect RNA structural and functional properties.

1.5.3. State machines

State machines are a very widely used tool used in biological analysis. State machines are directed graphs that depict the relationship between several states that a system can obtain and have beginning and end vertices. For example, a simple state machine could be used to represent a metabolic pathway by showing the possible series of chemical reactions that could transform glucose (a beginning state) into fatty acids or amino acids (end states) using various enzymes (directed edges). More specialised forms of state machines also exist that have been extensively used in the analysis of biological problems and RNA.

1.5.4. Hidden Markov Models

Hidden Markov Models (HMMs) were originally developed by Leonard E. Baum and speech recognition algorithms were one of the first early implementations (Rabiner, 1989). They are a type of state machine that assigns probabilities to the directed edges (transitions). The word "hidden" in the name refers to the end states, which are not directly observable and thus are predictions. A HMM is defined by specifying five components:

Q = the set of states = $\{q_1, q_2, \dots, q_n\}$

V = the output alphabet = $\{v_1, v_2, \dots, v_m\}$

$\pi(i)$ = the probability of being in state q_i at time $t = 0$ (*i.e.*, the initial states)

A = transition probabilities = $\{a_{ij}\}$,

where $a_{ij} = Pr[\text{entering state } q_j \text{ at time } t+1 \mid \text{in state } q_i \text{ at time } t]$. Note that the probability of going from state i to state j does not depend on the previous states at earlier times; this is the Markov property.

B = output probabilities = $\{b_j(k)\}$,

where $\{b_j(k)\} = Pr[\text{producing } v_k \text{ at time } t \mid \text{in state } q_j \text{ at time } t]$,

The three types of probability ($\pi(i)$, A, B) are calculated from experimental data.

MMN--LGCWMHVLF-
-MN--LGCFGLVL--
MMN--LPCPAVAAQN
MMNHRLGCWMLVLF-

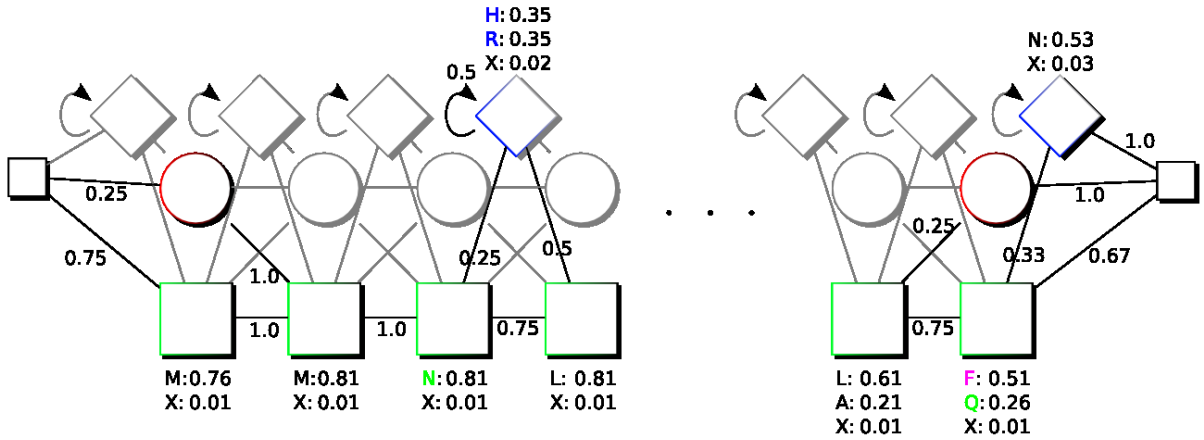

Figure 1.5: Hidden Markov Models

From the mock protein alignment given (above), the transitional and emission probabilities have been calculated and shown on the state machine (below). The beginning and end states are shown as small squares, the match states are shown as large green squares, the deletion states are shown as red circles and the insertion states are shown as blue diamonds. Any states or transitions that are not used have been greyed out. Emission probabilities are shown below match states and above insertion states and have been calculated with the "add-one" rule. This rule allows the possibility for other amino acids, which are not in the alignment, to be included.

As an example, a profile HMM, which are used to model multiple sequence alignments, can be build. A profile HMM has special beginning and end vertices that do not actually represent any position in the alignment, but are crucial for developing a proper model that can handle insertion and deletions. There are three types of vertices (match, insertion and deletion), allowing nine types of transition. Match-match, match-insertion, match-deletion, insertion-match etc… The standard profile HMM architecture is shown in Figure 1.5. Match and deletion nodes are calculated from the consensus sites in the alignment, while insertion nodes are calculated from the remaining alignment positions. When choosing consensus sites, a standard of >50% sequences represented is often used.

1.5.5. Covariance models

Covariance models (CVM) are a generalised form of HMM that have been developed specifically for investigating RNA structure and can be used to predict new members of

previously discovered RNA families. Instead of representing a position in a multiple alignment, the vertices of a CVM represent the consensus structure. There are two steps towards creating CVMs. The first step generates a guide tree that is a representation of the secondary structure. Based on the secondary structure, base pairs are assigned special vertices called “match pair” (MATP) and all other bases are assigned “match left”/”match right” (MATL/MATR) vertices. Branches are handled using a bifurcation (BIF) vertex, which allows the model to handle internal loops with several branches. The second step expands the guide tree to handle insertions and deletions. This step refers back to the original multiple alignment to retrieve the transitional and emission probabilities.

Aside from sequence homology, CVMs are probably the best current method for classifying ncRNA (Freyhult et al., 2007). In Chapter 3 we compare the performance of the tool developed to CVMs.

1.5.6. Microarrays

Microarrays are powerful high throughput tools used to analyse the expression of many genes simultaneously. The array itself is usually a simple glass slide with microscopic spots of DNA probes attached to the surface with each spot consisting of several identical sequences. The sequence of each spot can be random, targeted towards a certain gene, targeted towards artificial “spike-in” sequences or targeted towards some other sequence in a genome.

To use a microarray, the investigator extracts genomic or complementary DNA (cDNA) from the target organism(s). The DNA is amplified to levels detectable by the array and then labelled with a fluorescent dye. The array is washed with the labelled DNA and, under a fluorescent light, each spot glows with an intensity that corresponds to the amount of DNA that hybridised to the spot.

A tiling array is a microarray variant that is targeted to most, or all of, a genome. Depending on the array design, probes may represent either overlapping or spaced segments of a genome. Tiling arrays are normally used to detect transcribed regions of the genome under differing conditions, however the research presented in Chapter 2 uses tiling arrays to genotype *A. thaliana*.

1.5.7. Support vector machines

Support vector machines (SVMs) are supervised learning algorithms that classify quantitative data. They solve classification problems by finding the hyperplane that best divides two

classes of data in higher dimensional space. The data points lying closest to the hyperplane are called the support vectors. Each data point is projected into higher dimensional space using special equations called kernels. Kernels have been devised that are able to project data of many forms including sequences (Leslie et al., 2004) and graph topology (Borgwardt et al., 2005). SVMs are used in Chapter 3 for the prediction of ncRNA Rfam families. An overview of how they work is presented in Figure 1.6.

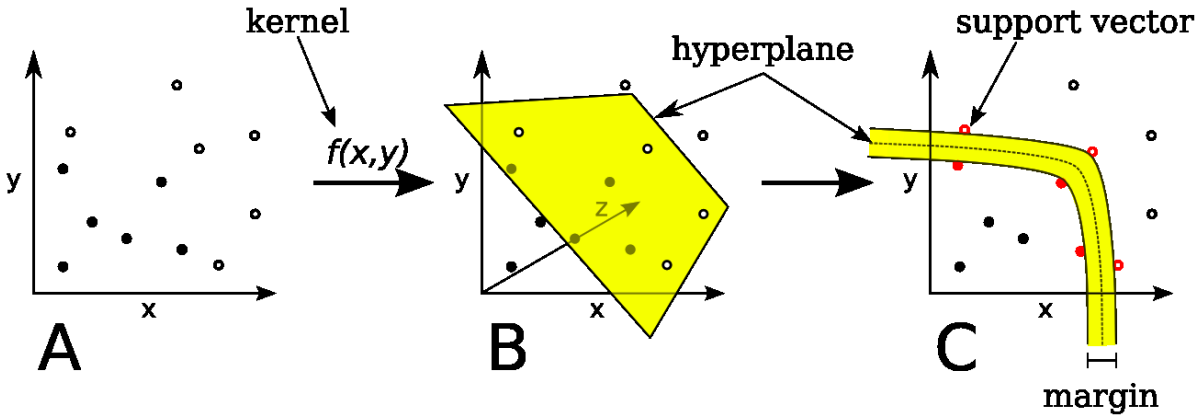

Figure 1.6: SVM algorithm

A: *A non-linear classification problem is computationally expensive to solve using non-linear methods.* ***B:*** *The SVM algorithm projects the data points into higher dimensional space using various mathematical functions called kernels. A linear solution in this higher dimension takes the form of a plane that divides the space by maximising the distance between the data points on the edge of their class space. This distance is called the margin.* ***C:*** *When the problem is looked at in the original number of dimensions, the plane becomes the hyperplane. The data points used to define the margin (red) are called the support vectors.*

1.5.8. Association studies

Association mapping is a method used to identify the possible underlying genetic causes for the quantitative variation of complex phenotypes. Genome-wide association analysis has become feasible due to recent advances in high-throughput technologies and computational power, and provides a clear advance over the traditional linkage analysis by increasing mapping resolution, reducing research time and providing greater numbers of alleles.

By conducting association mapping on several accessions belonging to the same species, the intrinsic genetic diversity of natural populations can be harnessed to resolve complex traits to single genes or even individual nucleotides. When considering association mapping within a species, the relatedness of the chosen accessions must be addressed. Ideally candidate accessions would be chosen that exhibit no population structure as this may lead to spurious associations between genotype and phenotype.

Linkage Disequilibrium (LD) is a measure of the degree of non-random association between alleles at different loci. In the absence of other forces, random mating will break down LD. Generally, LD extends to much longer ranges in self pollinated crops, such as wheat, than in cross-pollinated species, such as maize. Genome-wide LD determines the mapping resolution and marker density for a genome scan. If LD decays within a short distance, mapping resolution is expected to be high, but a large number of genomic markers is required. On the other hand, if LD extends a long distance, then mapping resolution will be low, but require a relatively small number of genomic markers.

Association mapping can be used to find associations between many different types of phenotypic and genomic data. Genomic data typically comprises of polymorphism data such as single nucleotide polymorphisms, single feature polymorphisms and short tandem repeats. Phenotypic data is usually a quantitative measurement such as fresh weight, metabolite measurements and enzyme activity measurements. Once this data has been obtained, statistical methods are used to find patterns of association between the genomic and the phenotypic data.

The underlying statistical methods used for association mapping are linear regression, analysis of variance (ANOVA), t-test or chi-square test. However, if allelic variation is too strongly associated with population structure, spurious genotype-phenotype associations can occur. These methods are usually augmented to deal with this complication through the use of genotypic information from random molecular markers across the genome, which accounts for genetic relatedness in association tests either explicitly (structured association) or through *ad hoc* adjustment (genetic control). With genetic control, a set of random markers is used to estimate the degree that test statistics are inflated by population structure, assuming such structure has a similar effect on all loci. In contrast, structural association first uses a set of random markers to estimate the population structure (Q) and then incorporates this estimate into further statistical analysis.

1.6. Outline

In this thesis, new experimental methods and bioinformatics tools were developed and applied to the study of ncRNA in the hopes of obtaining new information about ncRNA and their functions. In Chapter 2, tiling arrays were used to analyse genomic DNA, instead of the usual transcriptome, to identify genomic markers. These markers were then associated with phenotypic and metabolic traits to determine the potential underlying genetic cause for the

variation observed in the traits. This then paves the way for identifying any ncRNA genes that may be responsible. In Chapter 3, graph theoretic properties were calculated from graphs representing folded RNA secondary structure. These properties were then used to train and test support vector machines that predict whether an RNA molecule is functional or not and, if so, the potential function. The context of the results are then discussed in the final chapter along with proposed future projects and the criteria that need to be considered when designing new bioinformatics tools.

Chapter 2.

Arabidopsis genotyping[1]

2.1. Abstract

A collection of 54 Arabidopsis natural accession-derived lines were subjected to deep genotyping through single feature polymorphism (SFP) detection via genomic DNA hybridisation to Arabidopsis Tiling 1.0 Arrays for the detection of selective sweeps, and identification of associations between the genome, sweep regions and growth-related metabolic traits. A total of 1,072,557 high-quality polymorphic sites were detected and indications for 3,943 deletions and 1,007 duplications (all larger than 350 nucleotides) were obtained. Different genomic fractions and various functional gene classes varied significantly in polymorphism frequency. Significantly lower than expected SFP frequency was observed in protein-, rRNA-, and tRNA-coding regions and in non-repetitive intergenic regions, while pseudogenes, transposons, and non-coding RNA genes are enriched with SFPs. Gene families involved in plant defence or in signalling were identified as highly polymorphic, while several other families including transcription factors are depleted of SFPs. Five selective sweeps regions were detected with an overrepresentation of transporter genes. The sweep regions show no significant increase in associations with 9 measured traits (fresh weight, protein, amino acids, sucrose, starch, myo-inositol, beta-alanine, threonic acid and erythritol)

[1] This chapter is done in collaboration with Prof. Dr. Thomas Altmann, Prof. Dr. Karl Schmid, Dr Hanna Witucka-Wall and Torsten Günther. Although the goals and structure of the larger project are presented here, parts of this chapter have been altered to provide focus on SFP detection and genome wide trait-marker association.

than random sampling from the genome and a whole genome association produces 192 significant associations.

2.2. Introduction

Many plant species show a wide geographical range across contrasting ecological environments. Reciprocal transplantation experiments of ecotypes from different geographic origins in a common garden, provided evidence for local adaptation within the species range that resulted in fitness differences (Clausen et al., 1940; Korves et al., 2007; Schemske, 2007). Local adaptation by means of natural selection affects both the interaction with the abiotic environment, such as soil, temperature or photoperiod, as well as biotic interactions with competitors, pathogens and pollinators (Bradshaw Jr and Schemske, 2003). Frequently, plant species show a high level of phenotypic variation among ecotypes in morphology, phenology and biochemistry that is geographically structured (Holub, 2007; Koornneef et al., 2004). The high levels of phenotypic diversity lead to a key question: Which traits are polymorphic due to natural selection and which ones result from the differential fixation of random mutations by genetic drift? Since plants are sessile organisms showing highly structured populations and a propensity for self-fertilisation, random drift is expected to play a significant role in local populations. Therefore, one of the central goals of plant evolutionary biology is to disentangle the effects of both processes and to identify traits under selection along with the genes controlling these traits.

Natural accessions of *Arabidopsis thaliana* are characterised by a high level of phenotypic variation in morphological, developmental, metabolic, resistance and reproductive traits (Koornneef et al., 2004). Since *A. thaliana* shows a large distribution range throughout the Northern Hemisphere (Hoffmann, 2002) and grows in highly differentiated local habitats, it is ideally suited to investigate the roles of drift and selection in phenotypic variation. For example, flowering time and the vernalisation response show a latitudinal gradient across Europe (Dijk et al., 1997; Stinchcombe et al., 2004), suggesting local adaptation to different day lengths. Early flowering plants with a summer annual habit are predominately observed in Southern Europe whereas late flowering plants with a strong vernalisation response tend to occur in Northern Europe. Major candidate genes controlling flowering time and vernalisation response have been identified and allelic variation segregating at these loci shows evidence of positive Darwinian selection (Balasubramanian et al., 2006; Le Corre, 2005; Toomajian et al., 2006). A similar variation in flowering time was also observed in other species, indicating that latitudinal pattern does not result from past historical processes

like the re-colonisation of glacial habitats and the fixation of new mutations by genetic drift, but resembles local adaptation. Genes that mediate resistance to pathogens (R genes) are also well known to be highly polymorphic in *A. thaliana* (Bakker et al., 2006) and indications of various modes of selection acting on these genes have been obtained (Allen et al., 2004; Bakker et al., 2006; Bergelson et al., 2001; Mauricio et al., 2003; Mondragon-Palomino et al., 2002; Stahl et al., 1999; Thilmony, 2006; Tian et al., 2002).

In addition to flowering time and pathogen resistance, the utilisation of available resources is of central importance for plant fitness. Nutrient availability and the environmental conditions influencing carbon fixation show a high level of spatial and temporal variation. Substantial variation of biomass accumulation and metabolite composition has recently been observed across *A. thaliana* accessions (Cross et al., 2006; Keurentjes et al., In Press) (Sulpice et al., submitted) suggesting that local adaptation involves changes in metabolic traits that may result from natural selection. In *A. thaliana*, a positive correlation between above-ground biomass and fecundity was found (Aarssen and Clauss, 1992) as well as a strong correlation between biomass and the metabolic profile in a recombinant inbred line population of two genetically divergent accessions (Meyer et al., 2007). A subsequent analysis of 94 *A. thaliana* accessions uncovered much phenotypic variation in more than 90 metabolites and significant negative correlations between biomass and several metabolites (Sulpice et al., submitted). This work also revealed starch as a major integrator of the metabolic response. Allelic variation at two candidate genes, whose expression is affected by carbon regulation are associated with biomass, thereby providing a target for natural selection. While positive selection has been shown to operate on genes involved in the synthesis of compounds involved in plant defence and to drive diversification in plant secondary metabolism (Benderoth et al., 2006), and despite the documentation of significant genetic variation in enzyme activities of primary (and secondary) metabolism in *A. thaliana* (Mitchell-Olds and Pedersen, 1998), very little information is available on the extent and mode of selection on primary metabolic traits in plants.

Hybridisation-based arrays have been used to address a broad range of questions related to genomic variation (Gresham et al., 2008), such as the identification of natural variation between closely related organisms (Gilad and Borevitz, 2006). Tiling arrays hold sets of microarray hybridisation probes that contain both coding and intergenic portions of the genome and can be used to discover polymorphisms throughout the entire genome. Such polymorphisms are called single feature polymorphisms (SFPs) because they differ from a

common reference sequence. Although the exact nature of polymorphism in SFPs is difficult to identify they have been established to be a useful alternative to the genotyping of single nucleotide polymorphisms (SNPs) (Kim et al.).

The study of non-coding RNA (ncRNA) in *A. thaliana* shows that very little is understood about the biological role that RNA plays. Recently massively parallel signature sequencing (MPSS) of *A. thaliana* flowers and seedlings has discovered over 75,000 unique potential small RNAs (Lu et al., 2005) the majority of which are unannotated. Among the families of ncRNA discovered are miRNA (Bartel, 2004), snoRNA (Barneche et al., 2001; Qu et al., 2001) and riboswitches (Sudarsan et al., 2003). A brief look at Rfam, a comprehensive database for ncRNA (Griffiths-Jones et al., 2003), reveals that there are 1,255 ncRNA sequences with a predicted function. The disparity between the numbers of experimentally supported and computationally predicted small RNAs, identifies a key area where ncRNA investigation has much to contribute.

In the present study we seek to widen the scope of previous genotyping experiments by increasing the number of studied accessions and the density of SFP calling and by broadening the analysis to include representative primary metabolic traits. Two previous studies both investigated ~20 accessions with two different types of arrays: the Affymetrix ATH1 microarray (Borevitz, 2006) and the Perlegen high density re-sequencing microarray (Clark et al., 2007). While the high density re-sequencing arrays provide the best possible resolution available, high cost limits the number of accessions that can be investigated. The ATH1 microarray on the other hand is limited to ~11 probes per annotated gene. By using the Affymetrix Arabidopsis Tiling 1.0 Array and analysing 54 accessions, we substantially increased the coverage of the *A. thaliana* population analysed by deep genotyping and phenotyping. We also aim to add to the type of traits studied in *A. thaliana* by including phenotypic and metabolic data. Metabolic traits have been chosen primarily based on their correlation with biomass in a set of 94 accessions (Sulpice et al., submitted). Seven of them significantly correlate with biomass (proteins, total amino acids, beta-alanine, sucrose, starch and threonic acid) while two metabolites do not show any significant correlation (erythritol and myo-inositol). Using both types of data, we demonstrate how genome-wide SFP sites can be used to detect genomic regions showing signs of recent selection that are associated with phenotypic variation in metabolic traits. Our goal was to identify genomic regions, coding genes and non-coding genes which harbour genetic variation that played a role during the recent history of *A. thaliana* and may have contributed to local adaptation. We achieved this

by comparing regions with unusual haplotype structure that may have resulted from the rapid fixation of advantageous mutations with phenotypic variation of growth-related and metabolic traits such as fresh weight, protein content and metabolite levels to investigate whether genetic variation in these regions played a role in local adaptation during the recent species history. We also seek to provide further experimental and computational annotation of ncRNA in *A. thaliana*.

2.3. Methods

2.3.1. Selection of accessions

For this study, we chose a total of 54 accessions representing native ranges in Central Europe, Southern Europe, Asia and Eastern Europe from a larger set that had been subjected to phenotypic characterisation (Sulpice et al., submitted). The *A. thaliana* accessions used in this study were obtained from various sources: Col-0 from G. Rédei (Univ. of Missouri-Columbia, USA); Ler-1 from M. Koornneef (Wageningen University, Netherlands); Te-0 from S. Misera (Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben, Germany); Bor-4, Est-1, Lov-5, NFA-8, from D. Weigel (Max Planck Institute Tübingen), Ak-1, Bur-0, Enkheim-D, Jea, Mh-1, Oy-0, Petergof, Pyl-1, Shakdara, Stw-0, Ta-0, and Te-0 from the Versailles stock centre. All others were obtained from the Nottingham Stock Centre (NASC), through which all accessions are now available. Accessions were homogenised by single-seed propagation and were bulk-amplified prior to the analyses conducted (Toerjek et al., 2003).

2.3.2. Plant material preparation for phenotypic analysis

Plant growth conditions. Accessions were grown in short-day conditions (8/16h light/dark) in moderate light and well-fertilised soil to apply a moderate C deprivation, and harvested during the last hour of the day, 5 weeks after germination when they were still in the vegetative growth phase. Rosette fresh weight at the harvest time was measured as an indicator of biomass. To assess potential links between biomass and primary metabolism, six metabolic traits significantly correlating with biomass and two that that did not were selected (Sulpice et. al. submitted). Protein, total amino acids, beta-alanine, sucrose, starch, threonic acid, erythritol and myo-inositol were measured as representatives of structural and storage components (see below). Phenotype data were calculated as least squared means values over the experiments. Phenotyping for the traits used in the association mapping was carried in 6 to 8 independent experiments, in particular for fresh weight, protein, amino acids, sucrose

and starch in 8 experiments and for myo-inositol, beta-alanine, threonic acid and erythritol in 6 experiments.

Metabolic trait analyses. Metabolic traits were analysed as described by Sulpice et al. (submitted). In brief: Chemicals were purchased as described by Gibon et al. (Gibon et al., 2004) and total protein, starch, sucrose, total amino acids were assayed as described by Cross et al. (Cross et al., 2006). Alanine, threonic acid, erythritol and myo-inositol were determined by gas chromatography (GC-MS) coupled to mass spectrometry. Metabolite extraction for GC-MS was carried out on the exact same samples as used for enzymes and metabolites determined by spectrophotometric methods as described previously (Schauer et al., 2006). 50 mg of Arabidopsis shoots were homogenised using a ball mill pre-cooled with liquid nitrogen. Derivatisation and GC-MS analysis were carried out as described previously (Lisec et al., 2006). The GC-MS system was comprised of a CTC CombiPAL autosampler, an Agilent 6890N gas chromatograph and a LECO Pegasus III TOF-MS running in EI+ mode. Metabolites were identified in comparison to database entries of authentic standards (Schauer et al., 2005).

Linkage disequilibrium and population structure. The accessions were genotyped with 460 SNP markers: 149 framework SNPs assembled in the frame of the *A. thaliana* "HapMap" project, (J. Borevitz, pers. communication; see http://naturalvariation.org/hapmap and links given therein) and 311 SNPs with intermediate allele frequency selected by N. Warthmann et al. (Warthmann et al., 2007). SNP genotyping was carried out at Sequenom Inc (San Diego, CA, USA). Microsatellite typing was perform using Li-Cor 4300 (LI-COR Biosciences GmbH, Germany) and allele scoring was done using the SAGA GT genotyping software (LI-COR Biosciences GmbH, Germany).

The program STRUCTURE 2.2 (Pritchard et al., 2000) was used to determine population structure and assign accessions to subpopulations. For analysis of population structure and kinship among the accessions selected for this study, 417 of the 460 SNP markers were selected for <25% missing data. We used an ancestry model that allows population admixture, and allele frequencies among population were assumed to be correlated (*i.e.*, allele frequencies were likely to be similar due to shared ancestry or migration). The optimal number of subpopulations was simulated by setting k (number of clusters) from 2 to 10. The length of burn-in period as well as Markov Chain Monte Carlo iterations (MCMC) after burn-in were set to 100,000 for each run and each run was iterated 10 times. An accession was assigned to the subpopulation or group to which it showed the highest probability of

membership. Two criteria were used to determine the k value which best fits the data. First, the $Pr(X/k)$ value should be less than or equal to zero, and second the value of alpha as a measure of population admixture should remain constant (<0.2) and the cumulative value of Δ for alpha-factor for next sub-cluster should not decrease. The resulting Q matrix was used for controlling population structure in subsequent association analysis.

2.3.3. Plant material preparation for genotypic analysis

Plant growth conditions. Growth conditions for phenotypic analyses were exactly as previously described by Cross et. al. (Cross et al., 2006). Briefly: Seeds were germinated and grown for the first 7 days with a day length of 16 h, temperature of 6°C at night and 20°C during daytime, humidity 75%, and luminosity 145 µmol m-2 s-1. After 7 days, seedlings were transferred to a phytotron. Growth was continued in an 8-h-light/16-h-dark regime at temperatures and humidities of 16°C and 75% at night and of 20°C and 60% during the day. Illumination was 145 µmol m-2 s-1. At the age of 2 weeks, plants of average sizes were transferred to pots of 6 cm in diameter (5 plants per pot). Plants were switched to a controlled small growth chamber after 1 further week in the short-day conditions outlined above for 2 weeks more. Day length was then 8 h, temperature a constant 20°C, and illumination an average 125 µmol m-2 s-1. Plants were watered daily. Harvests of 15 plant rosettes were performed at the end of the light period. Each sample typically contained three rosettes, equivalent to 500 mg fresh weight, depending on the accessions. The entire sample was powdered under liquid nitrogen and stored at –80°C until its use.

Extraction of genomic DNA. 25 ml of preheated CTAB buffer were added to approximately 4g of plant tissue, ground in a Retch-mill in liquid nitrogen and incubated for 20 min at 65˚C. 10 ml of chloroform:isoamyl alcohol (24:1) were added to the samples, which were then incubated for 20 min at room temperature followed by centrifugation for 5 min at 3000 rpm. The upper phase was transferred to a new Falcon tube and 17 ml of isopropanol were added, followed by incubation for 10 min on ice and centrifugation for 5 min at 3000 rpm. The upper phase was discarded and 4 ml of H_2O were added to the pellet and mixed. 4 ml of 4M $LiCH_3COO$ were added followed by an incubation step for 20 min on ice and then cleared by centrifugation for 10 min at 3000 rpm. The upper phase was transferred to a new Falcon tube and 16 ml of 96% EtOH were added and incubated at 4˚C for 1 hour. The sample was pelleted by centrifugation for 20 min at 4000rpm and dried. Afterwards the pellet was re-suspended in 900µl H_2O, transferred to two new Eppendorf tubes (2ml) and 140 µl of RNAse

(10 mg/ml) were added to the each sample and incubated for 30 min at 56˚C. 130µl of 5M NaCl were added to the samples followed by centrifugation for 5 min at 12000 rpm. The upper phase was again transferred to a new Eppendorf tube and 360 µl of phenol/chloroform were added gently mixed and spun-down by centrifugation step for 5 min at 14000 rpm. In the next step, the upper phase was transferred to a new tube mixed with 1 ml of chloroform:isoamylalkohol (24:1) and cleared by centrifugation step for 5 min at 14000 rpm. The whole step was repeated and the upper phase was transferred to a new Eppendorf tube and mixed with double volume of 96% EtOH following by incubation for 5 min on ice. The mix was centrifuged for 5 min at 14000 rpm and the upper phase was discarded. The pellet was washed with 70% EtOH, shortly centrifuged and the upper phase was discarded. The last step was repeated. Samples were placed on a thermo block at 37˚C for drying. The pellet was dissolved in 50 µl of H_2O. The DNA quality was tested using a NanoDrop photometer (Thermo Scientific). Only samples with 230/280 factor > 1.85 were used for further analysis.

DNA digestion and labelling. 3.5 µl 10x One-Phor-All Buffer (GE Healthcare), 2.1 µl 25 mµ CoCl and 1 µl 1:3 1U/µl DNase (Promega) were quickly added to 25 µg genomic DNA and was filled up to a total volume of 35µl with H_2O. Digestion was carried out for 4 min at 37˚C. To stop the reaction, samples were incubated for 15 min at 95˚C and then cooled down to 4˚C.

The quality of digested products was checked by 2% agarose TBE gel electrophoresis. Samples were classified as well digested if the smear of DNA fragments was about 50 base pairs.

2µl of 20 U/µl Terminal deoxynucleotidyl transferase and 2 µl of Biotin-N6-ddATP (Enzo Life Sciences Inc.) were added to 20 µg of digested DNA and incubated for 1 h at 37˚C, followed by a cooling step to 4˚C. The reaction was carried out in a dark room.

Array hybridisation. Labelled DNA samples were sent to the RZPD (Berlin) service unit for hybridisation. Affymetrix array hybridisation was carried out according to the manufacturer's standard instructions for expression array hybridisations. For each accession, a total of three independent DNA samples were used. Three reverse arrays were hybridised (one for each of the biological replicates) and a fourth forward array was hybridised using a DNA mixture of the three biological replicates.

2.3.4. SFP identification

Signal intensities obtained from the array hybridisation were analysed using customised

software (available upon request) to enable exploration of different normalisation strategies and SFP calling methods. The main functionality of the software is based upon the TILING-ARRAY ANALYSIS SOFTWARE (TAS) from Affymetrix. To calculate the log fold change, we used the Hodges-Lehmann estimator method. For each probe, the log fold change for all pairs of Col and non-Col chips was calculated and the median was taken.

We utilised the sequence dataset, referred to as the "2010" dataset, from Nordborg et al. (Nordborg et al., 2005) to assess the accuracy of our SFP prediction method by estimating the False Discovery Rate (FDR). The 2010 dataset comprises 1,213 fragments of Columbia genomic regions, each approximately 500 nucleotides long, produced through dideoxy-sequencing of 97 accessions. 12 of our accessions were found represented in the 2010 dataset. To allow direct comparison of our results, we converted the SNPs identified in the 2010 dataset to SFPs by mapping them onto the Affymetrix probes. The accuracy was then assessed using Receiver Operator Characteristic (ROC) plots. A False Negative Rate (FNR) was also estimated, using the 2010 data, as the fraction of the remaining sequence-verified polymorphic probes that were not predicted to be polymorphic using our methodology.

We tested several normalisation strategies including invariant normalisation, mode and range normalisation, and quantile normalisation. For invariant normalisation, we identified invariant probes; *i.e.*, probes not found to be polymorphic based on the 2010 dataset, and determined the normalisation shift that aligns the invariant probe signal intensity distributions to those of the Col-0 control hybridisations. Mode and range normalisation adjusts the mode of the signal histogram while ensuring that the lowest and highest values of the distribution are equal to the lowest and highest values of the control distribution. For quantile normalisation, the quantiles of the signal intensity distributions are rendered identical to the control distributions.

We also investigated several methods to improve the FDR of SFP calling. A typical Affymetrix microarray analysis would use the difference between perfect match and mismatch probes as the signal. Here, we also analysed the perfect match probe alone. We also studied the effect of removing outliers reported in the CEL files and substituting an average of the local array area. However, this approach masked the SFPs we seek resulting in poor sensitivity. We also tested the number of forward and reverse arrays required to produce the best result using the initial test cases (Ler-1, Ak-1 and Bch-1). A combinatorial approach using three forward and three reverse arrays, shows that a balance of forward and reverse arrays produces the best result. Naturally, the more arrays used, the better the results.

However, the improvement provided by using more arrays peaks at around four. To maximise the number of accessions possible to analyse, we chose to use 3 reverse arrays and 1 forward array per accession.

A Linear Discriminant Analysis (LDA) using STATISTICA 7.1 (http://www.statsoft.com) was performed to compare different SFP calling techniques and identify the best possible combination of predictive variables. It determined that log base 2 (log_2)-fold change of the perfect match probe alone reflects 98% of the discriminatory information. Among the other variables used in the LDA were the Wilcoxon signed rank test P value, t-test P value, GC content (%G+C), RNA-folding free energy (MFE) of the probe as determined by RNAFOLD (Hofacker et al., 1994), normalised absolute intensity signal, difference between observed and expected values as calculated by significance analysis for microarrays (SAM) (Tusher et al., 2001), and the position relative to the centromeres. As a result, we chose to use the log_2-fold change as the only variable to call SFPs. Based on the ratio of true positives to false positives as determined in the ROC plots, we defined a threshold which discriminated SFP from non-SFP calls. The final combination of parameters used for SFP calling were quantile normalisation, perfect match probes only, no outlier removal, and calling SFPs on the log_2-fold change. The threshold for calling SFPs was set at a log_2-fold change of -1.5 which is the equivalent of a 2.8-fold decrease in signal intensity.

2.3.5. Analysis of functional gene categories

We obtained genome structural, gene family as well as functional category annotation information from TAIR and NCBI. Forty gene families were chosen based on families analysed in previous studies (Clark et al., 2007) and the gene families available in TAIR. To identify statistically significant enrichments or depletions in particular genomic regions (coding regions, intergenic regions and others), the observed number of SFPs in a particular region is compared to the expected number assuming that SFPs are distributed purely proportionally to the total size of the region. Statistical significance was evaluated using a paired Students t-test.

2.3.6. Detection of duplications and deletions

From the resulting SFP dataset, consecutive stretches of 10 consecutive probes or longer were considered deletions. The gene content for each deletion was extracted and assigned to the functional categories to test for different deletion frequencies. Putative gene duplications were detected by searching for stretches of 10 consecutive probes with a greater than two-

fold increase in signal. The functional categories in these regions were also analysed.

2.3.7. Whole genome haplotyping

To reduce the impact of multiple testing and the time required for association testing, the genome was represented as a series of haplotypes based on the linkage disequilibrium and population structure calculated in the phenotypic analysis (section 2.3.2). The memory requirements of the SFP data proved prohibitive for analysis with tag SNP selection tools so an alternative method was developed.

First, the SFP data was divided into tiled regions defined as half the length of the average LD. The length was chosen as a trade-off between capturing an accurate representation of the local haplotype and reducing the number of tests to be made. Second, each region is then clustered using a k-means clustering algorithm. The parameter k was chosen based on the population structure determined by STRUCTURE 2.2. The resulting clusters were considered the haplotype for the particular region of the genome being analysed.

2.3.8. Identification of selective sweeps

The detection of selective sweeps followed the approach of Toomajian et al. (Toomajian et al., 2006). The PHS statistic (Toomajian et al., 2006) was computed for all SFPs with a minor allele frequency of at least 5%. This test for selection corrects for the population structure by taking the pairwise similarity of accessions into account, therefore it should be suitable for our structured sample from natural populations. Additionally, it accounts for the local recombination rate by using genetic distances between the markers. For this conversion, we fitted a polynomial curve to 253 markers for which physical and genetic position are known, as previously done (Schmid et al., 2005). After calculation of PHS only the highest scoring SFP among highly linked neighbouring SFPs ($r^2 > 0.5$ and r positive) was included in the following analysis (standardisation identification of sweep candidates). The scores were then standardised for their allele frequency f like follows:

$$standardized\ PHS\ =\ (PHS - median_f[PHS])/SD_f[PHS].$$

SD_f and $median_f$ refer to standard deviation and median, respectively, computed for all alleles with a frequency of f; the median is used instead of the mean because it is less sensitive to extreme values.

To identify sweep candidate regions, the density of top 0.5 % standardised scores were

determined by calculating the proportion of such outliers in a sliding window along the genome (window size of 1000 SFPs and offset of 50 SFPs between adjacent windows). Windows with an outlier proportion of more than 45 % were denoted as sweep candidates and within windows the highest scoring SFP was taken for the associations. The identified sweeps are represented as a focal SFP site that best resembles all the SFP sites in the sweep.

2.3.9. SFP-trait associations

Marker-trait associations were calculated using general linear models (GLM) (Yu et al., 2006) as implemented in the TASSEL program (Bradbury et al.). The GLM requires a population matrix (Q), which was obtained from the program STRUCTURE 2.2. Association mapping was performed with the whole genome haplotypes earlier identified in section 2.3.7 and the focal SFPs of the five sweeps identified in section 2.3.8.

The GLM (general linear model) function of the TASSEL program was applied with the STRUCTURE results for $k = 4$ to determine the significance of marker trait associations for following traits: fresh weight, protein, amino acids, sucrose, starch, myo-inositol, beta-alanine, threonic acid and erythritol. The p-values obtained from TASSEL are corrected for multiple testing by generating an empirical distribution through 10,000 permutations of the input data for each marker-trait pair (Churchill and Doerge, 1994). Phenotype data were calculated as least squares means (LSM) values over the experiments. The GLM that was used for the calculations was

$$v = M + Env + Q_{k=4} + E,$$

where v is the dependent variable, M is a marker effect, Env is an effect of environment of different experiments, Q is population structure cofactor, E is experimental error.

For the sweep-trait associations, the focal SFP site from each sweep was associated with each of the traits. The SFP sites in the sweep bear a resemblance to the focal SFP site. Whole genome associations were carried out in a two-step process. In the first step, the whole genome haplotypes were associated with each of the traits. The haplotypes with significant associations to the traits were then expanded into the regions that they represent and the associations were re-run using the SFP sites to provide a finer resolution association.

2.4. Results

2.4.1. Patterns of SFP diversity

Using Affymetrix Arabidopsis Tiling 1.0 Arrays, we created high density SFP maps for 54 genetically diverse *A. thaliana* accessions (Figure 2.1). An analysis of various possible SFP prediction criteria using a Linear Discriminant Analysis (LDA) reveals that the $\log_2$-fold change between Col-0 and non-Col-0 arrays accounts for greater than 98% of the predictive power. Other methods such as significance analysis for microarrays and t-tests perform just as good or worse despite significantly increased calculation costs. Based on these results, we chose to apply a threshold to the $\log_2$-fold change in order to call SFPs based on hybridisation signals. The quality of the predicted SFPs for 12 accessions was tested against the 2010 dataset of SNPs determined by sequencing (Nordborg et al., 2005) using Receiver Operator Characteristic (ROC) plots. Based on the plots prediction performance and weighing true positive against false positive rates, the threshold we chose to apply to the log2-fold was -1.5 (2.8-fold decrease in signal) to produce the best False Discovery Rate (FDR – the proportion of false positives among the called SFPs) across all 12 accessions. At this threshold, we were able to predict SFPs at FDRs between 0% and 32.1%. False Negative Rates (FNRs – defined as the proportion of false negatives to the polymorphic probes) were estimated to be between 65.5% and 77.7%. We also tested the effect of various properties on the probe signal intensity; the position of an SNP in a probe, the number of SNPs in a probe and the %G+C of probe (Figure 2.2). By using every unique probe on the array to call SFPs, we obtained 1,072,557 polymorphic sites in 54 accessions. On average, 85,968 SFPs were called in each accessions; numbers range from 42,251 in Co-3 to 122,079 in H-O-G. Our control Col-0, which is the reference genome of the array, produced only 199 SFPs, thus confirming a low FDR of the applied SFP calling procedure.

SFPs occur in high densities across all 5 chromosomes sometimes occurring in densities of up to one-fifth of the 50kb window. When compared to gene density, we observed a negative linear correlation of -0.37 (p-value = 2.05E-78) indicating that SFPs were depleted in genic regions.

Using the TAIR8 genomic annotation information, we calculated the proportions of SFPs in each feature and compared them to an expected value under the assumption of a random SFP distribution. For each genomic feature, we calculated the significance of the difference from random expectation using a paired Students t-test. As expected, SFP frequency per nucleotide

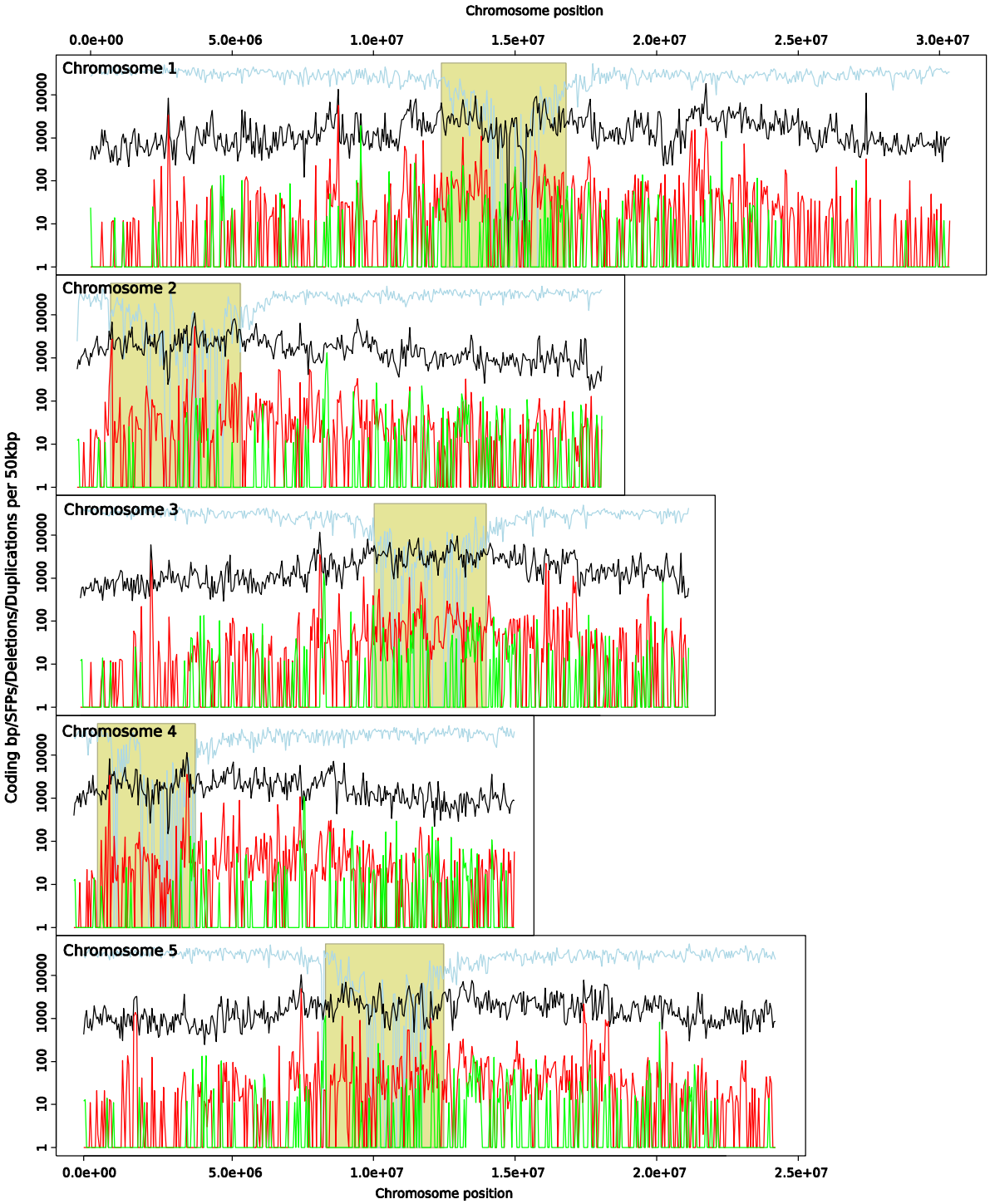

Figure 2.1: The density of SFPs and genes across each chromosome of Arabidopsis.

SFP density is calculated for all accessions per base pair using a sliding window of 50k nucleotides and the density at each position is plotted on a log scale (shown in black). The gene density, calculated as coding nucleotides per every 50k genomic nucleotides is also calculated (shown in light blue). The olive regions show the position of the centromeres. Deletions are shown in red and are defined as 10 probes in a row that were called SFPs (~350 nucleotides). The number of deleted nucleotides calculated in a 50k nucleotide window is shown in red. We have defined a duplication to be at least 10 probes in a row with a log2-fold change of greater than 1 (2-fold increase in signal). In this figure, the number of duplicated nucleotides is calculated in 50kb windows across the chromosome (green).

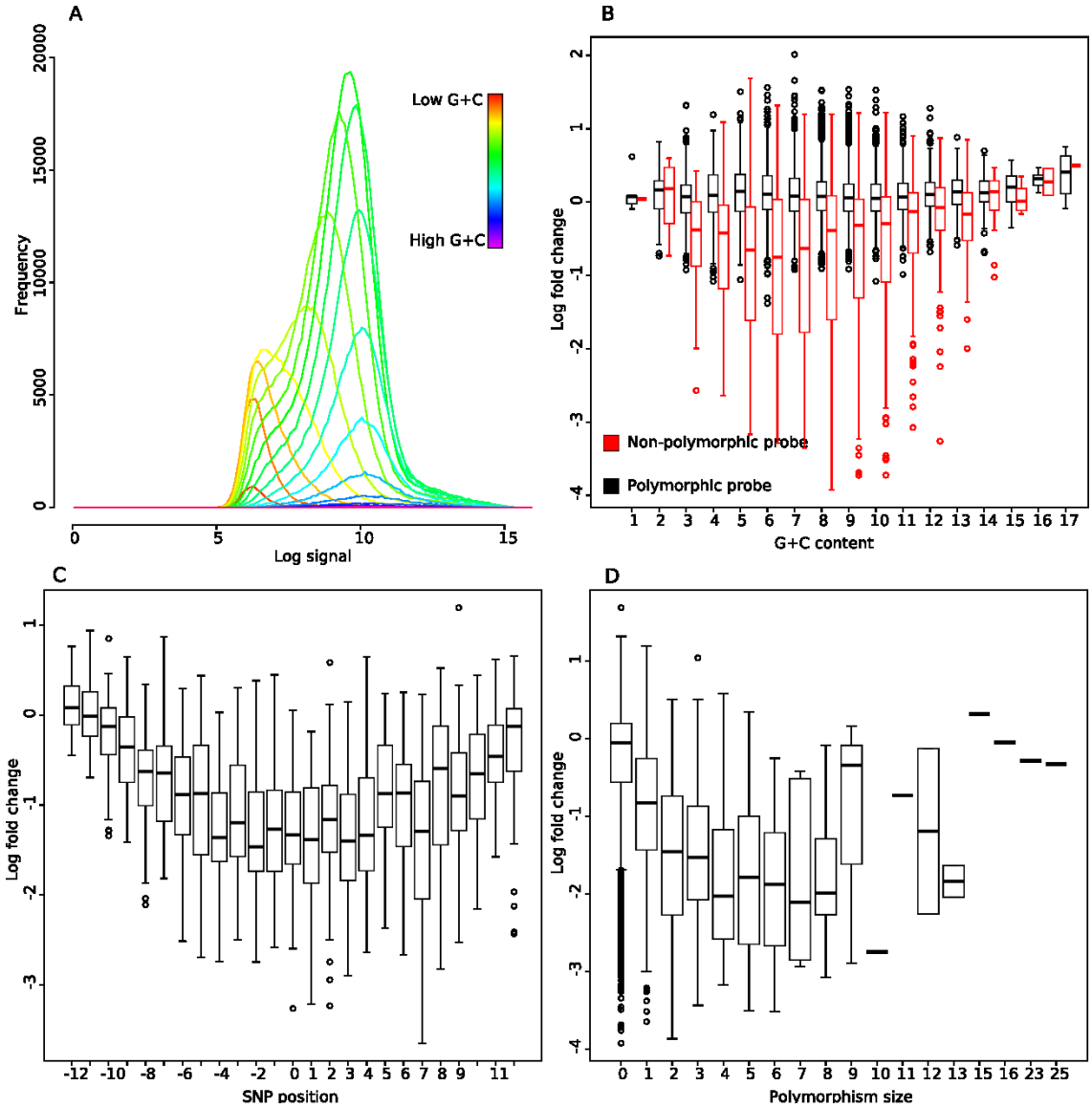

Figure 2.2: Tiling array signal properties

__A:__ The signal for a probe is highly dependent on the %G+C content. The lines show the signal distribution for probes containing a spectrum of %G+C content starting from low %G+C content (red) going to high (purple). In general, the higher the %G+C content, the higher the signal. __B:__ %G+C content had a different effect on the signal depending on whether the probe was polymorphic. Polymorphic probes show a much larger change in signal and variability for middle range %G+C contents. __C:__ The signal of a probe targeting a polymorphism is dependent on the position of the polymorphism. If the polymorphism occurs towards the centre of the probe, the change in signal will be much stronger. __D:__ The number of SFPs or the size of the polymorphism in a probe will also change the fold change. The more SFPs in a probe, the stronger the fold change.

is significantly reduced in functional regions of the chromosomes. By contrast, there is a significantly higher than expected frequency of SFPs per nucleotide in non-coding RNA, pseudogenes, and transposons. Introns, miRNA, snRNA and snoRNA show no significant difference from a random distribution (Figure 2.3).

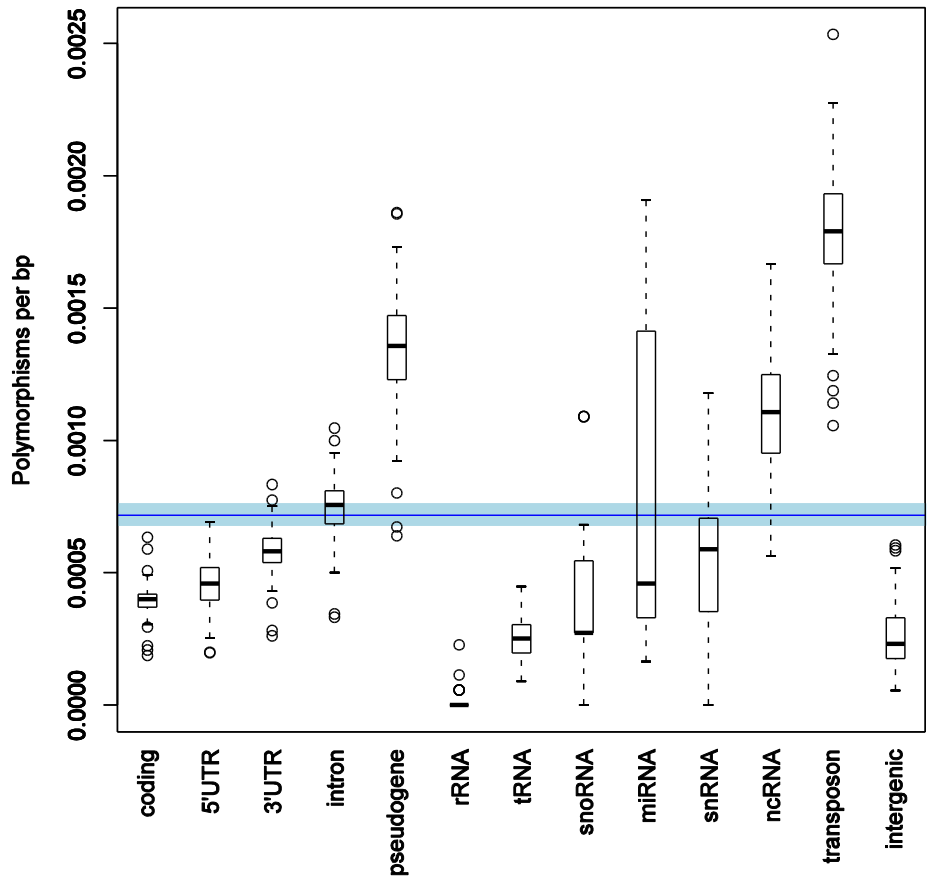

Figure 2.3: Observed and expected SFP frequency.

The presence of SFPs in various genomic structural features is shown as the number of SFPs per base pair for each entity. The spread of observed values across all accessions is shown in black. The blue band shows the range between the 1st and 3rd quartile while the blue line shows the median value. A chi-squared test is used to determine whether the observed amount of SFPs per class was close to the expected number due to random chance. All classes are significantly greater than or less that random ($p < 0.05$) except for the classes 'intron', 'snRNA' and 'snoRNA'.

For protein coding regions, we collated functional categories from TAIR and NCBI. Of these 40 chosen gene families, 20 contained a significantly greater than random amount of SFPs and 18 contained significantly fewer SFPs than expected from a random distribution. Disease resistance, plant defence and signalling genes comprised the majority of the gene families which were enriched in SFPs. Various transcriptions factors comprised the majority of the top 20 most significant gene families which were observed depleted in SFPs (Figure 2.4).

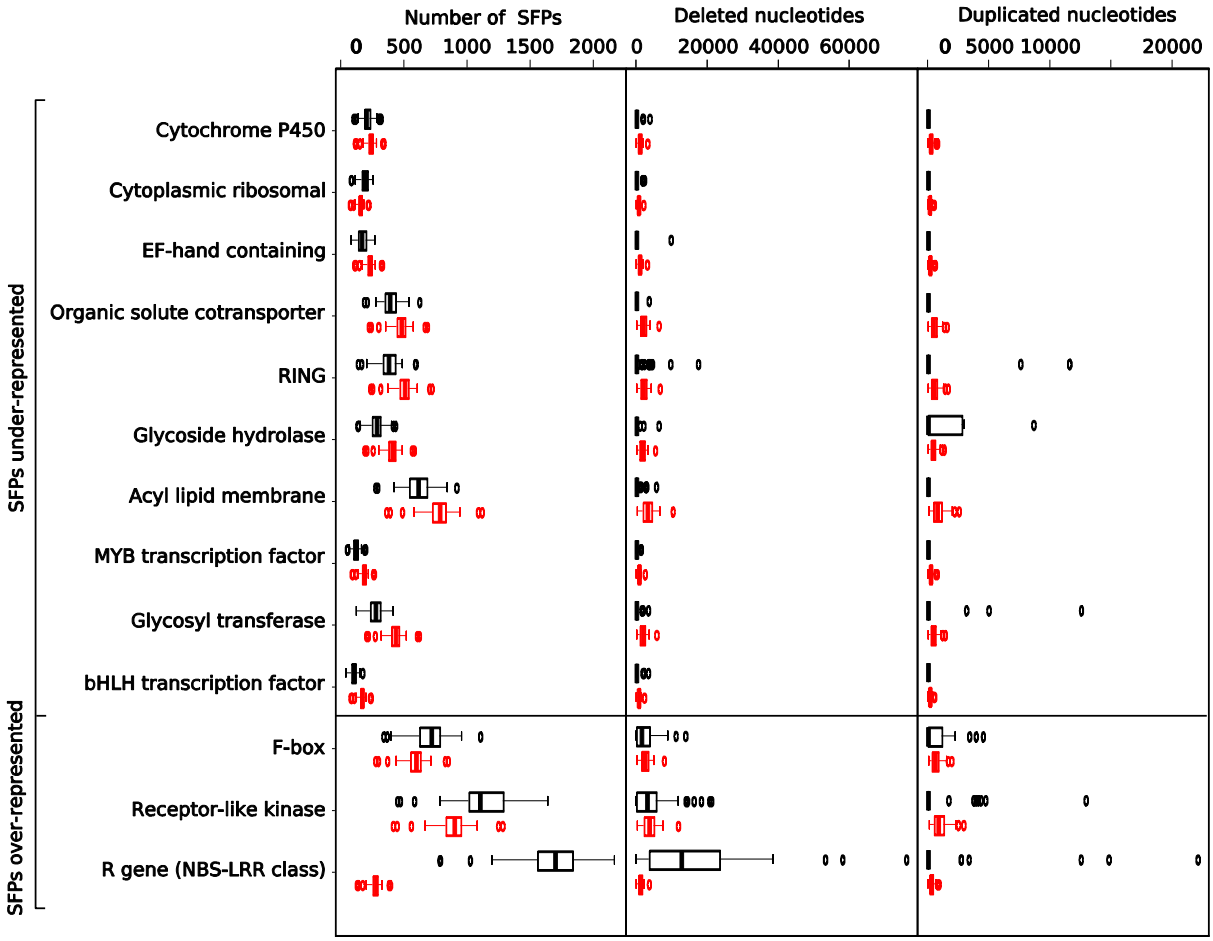

Figure 2.4: Number of SFPs in selected gene families.

Of the available gene families, 13 are shown here that have been investigated in previous studies for comparison purposes. The observed number (black) of SFPs in a gene family is compared to the expected number (red) due to random chance. Gene families are sorted by the significance of the difference between observed and expected SFP numbers. All gene families show significantly different observed SFPs than expected except for RING genes. Deletions follow a similar pattern to SFPs by removing the most nucleotides from resistance, F-box and receptor-like kinase genes but differ by removing fewer nucleotides from cytoplasmic ribosomal genes. Duplications rarely occur in any of the chosen gene families. Although they do occur in resistance genes, the difference between observed and expected is not significant.

2.4.2. Deletions and duplications

We observed 3,943 deleted regions with a mean length of 1,086 nucleotides (on average 28 probes in a row). A minority of the deleted regions occur across many accessions; however, the majority occur in only one. The accessions show a broad range in the number of deleted regions ranging from 25 in Co-3 to 415 in Dal-12, with a mean of 135 deleted regions per accession. Many of the areas with a high SFP density are explained by an abundance of deletions. Of the chosen gene families, the combined length of deletions is largest in NBS-

LRR genes and smallest in bHLH transcription factors.

We detected 1,007 duplicated regions, with a mean length of 1,088 nucleotides. Similar to deleted regions, the majority of duplications occur only in a single accession, while a small minority occur in many accessions. The accessions also show a range in the number of duplicated regions from 8 in Blh-1 to 83 in Shakdara, with a mean of 34 duplicated regions per accession. Like deletions and SFPs, NBS-LRR and F-box genes have the highest total length of duplicated regions whereas many other categories have no duplicated regions at all (Figure 2.4). However, in this case, the enrichment in duplication is not significant for these gene families but it is significant for MADS-box transcription factors and glycoside hydrolases. It is of note that the same gene classes show a high frequency of insertions and deletions.

2.4.3. Population structure analysis

The analysis of population structure with both the SFP data and the independently genotyped SNPs confirms the existence of population structure, even among accessions from Central Europe. It was previously thought that they represent admixture zones since past glaciation events (Schmid et al., 2006; Sharbel et al., 2000). A STRUCTURE analysis (Pritchard et al., 2000) indentified four major clusters of accessions. These clusters separate the Central Asian, Eastern European, Central/Western European and Iberian accessions. A high proportion of the Central European and Eastern European accessions are admixed individuals, probably reflecting the effect of postglacial population admixture and subsequent genomic recombination in suture zones. The STRUCTURE analysis is supported by clustering based on pairwise distance of both the SNP and SFP data. Distance matrices from both marker systems are highly correlated. The SNP markers also show a significant correlation between geographic and genetic distance although it is weak, possibly because of the large proportion of individuals from Central Europe and the disturbance of the geographic distribution by human agriculture. Taken together, the population structure analysis indicates the need for correction of population structure for this sample in the subsequent association study.

2.4.4. Choice of phenotypic traits

Nine phenotypic traits were selected from a larger data set, including rosette fresh weight and eight metabolic traits; six of which showed high correlations with rosette fresh weight (Sulpice et al., submitted). All of the measured traits used in the association study were shown to be normally distributed.

In a regression analysis across a larger set of 120 accessions, significant Spearman's R values were found for sucrose vs. protein, protein vs. alanine and alanine vs. sucrose (0.43, 0.32 and 0.28, respectively) as well as for protein vs. threonic acid (0.26) (Sulpice et al., submitted). Hence, the associations for these traits in a given sweep may not be independent.

2.4.5. Analysis of selective sweeps

Putative selective sweep regions were identified by detection of unusual patterns of shared haplotypes among accessions, based on the notion that genomic regions with selected haplotypes will be longer than regions without selected haplotypes due to hitchhiking of linked variation with the selected mutation (Sabeti et al., 2002). Hence, young and selected alleles will be surrounded by extensive linkage disequilibrium (LD) which produces longer haplotypes, while older alleles would show lower LD due to recombination. Plots of pairwise SFP haplotype sharing indicate regions with high levels of haplotype sharing. We used the pairwise haplotype sharing score (PHS) which was previously applied to *A. thaliana* data (Toomajian et al., 2006). Five significant sweep regions were identified and designated as SR1 to SR5 (Figure 2.5). These regions show a high proportion of significantly scoring SFPs with unusual haplotype patterns.

The candidate regions range from a length of 145 kb (SR2) to 353 kb (SR4) and the average length of haplotype sharing around the focal SFP ranges from 1.2 kb (SR3) to 6.4 kb (SR1). Two candidate regions (SR2, SR4) are further characterised by positive Tajima's D values indicating some form of balancing selection. The genome-wide highest scoring SFP is located in SR3 and, thus, was chosen as focal SFP for this region. Candidate region SR5 includes the gene FRIGIDA which was previously identified to evolve under selection (Korves et al., 2007; Toomajian et al., 2006) and which reflects the power of the PHS test to detect true positive signals of selective sweeps.

The PHS test was not able to show significant signatures of selection in the candidate regions previously suggested by Clark et al. (Clark et al., 2007). Although we found extensive haplotype sharing in one of their candidates on chromosome 1, this region showed no clear peak in PHS density even if we relaxed the threshold for significant scores. This discrepancy might be caused by the different composition of our sample, but the most important difference is our method is able to identify sweep candidates. Clark et al. simply identify regions with a low diversity over a long range. The PHS test compares the length of the different haplotypes around a core allele and, additionally, accounts for the population

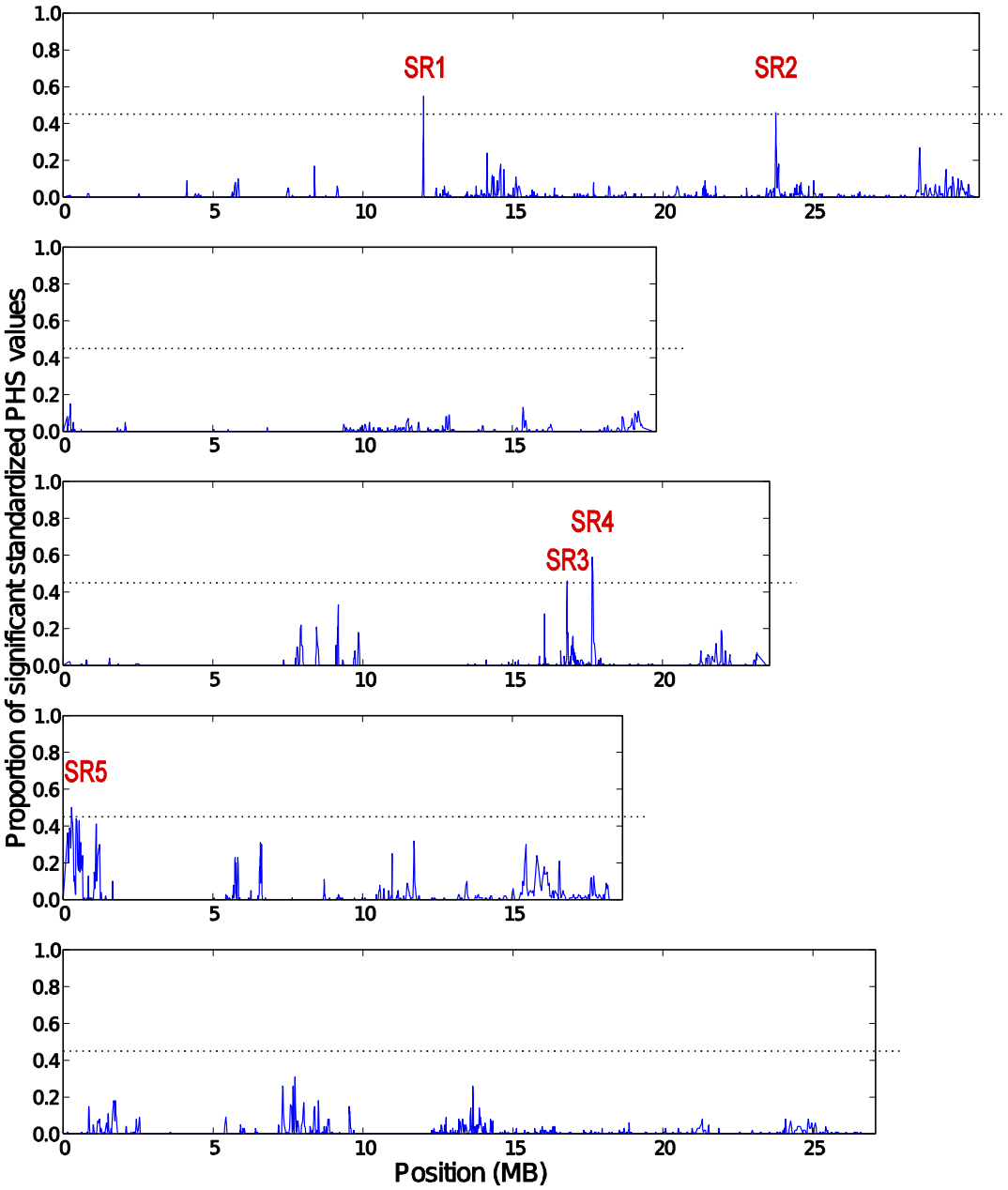

Figure 2.5: Location of selective sweeps

Map over all sweep candidates - The graph shows the density of significantly scoring SFPs measured by the proportion of significant SFPs in a sliding window (window size of 1000 SFPs, offset between windows of 50 SFPs) across the genome. The dashed line represents the cut-off for the definition of sweep candidates. Sweep candidate regions (SR) are denoted in red.

structure and local recombination rate. Hence, the PHS test represents a more rigid measurement to identify selective sweeps then the approach by Clark et al. (Clark et al., 2007).

In addition to being highly polymorphic, NBS-LRR protein-encoding genes also have a

relative overrepresentation of SFPs with extreme iHS values (multiple testing adjusted p=1.6E-10), suggesting that they are preferred targets of selection in *A. thaliana*. The Receptor-like kinases genes-category was the other functional class significantly enriched in extreme iHS values (p=0.04). Interestingly, although cytoplasmic ribosomal protein genes and F-Box genes are among the most highly polymorphic genes (Figure 2.4), none of the chosen gene families show significant enrichment or depletion in the sweep.

2.4.6. Association mapping

The variation in the phenotypic measurements is likely to be caused by mutations in the genome. Those that are beneficial are potentially selective sweep generators and thus many more sweep SFP sites are expected to be associated with the nine traits than non-sweep SFP sites. We tested the hypothesis that an SFP in a selective sweep is more likely to be associated with a trait than a random SFP by randomly sampling 5000 SFPs from our dataset and performing an association test on the random SFPs along with the focal SFPs from each sweep. Nine traits were associated with 5 focal SFPs and 5000 random SFPs using a general linear model for a total of 45,045 associations. To control for population structure in the GLM, associations for all traits were calculated using a Q-matrix as cofactor based on 417 SNPs estimated in STRUCTURE using $k = 4$ populations. Multiple testing was controlled with 10,000 permutations per marker-trait comparison. The ranks, based on F-score, for the focal SFPs were compared with the ranks of the random SFPs using Wilcoxon's Rank Sum Test, which showed no significant shift in sweep SFP p-values towards greater significance than randomly sampled SFPs. This result is likely to be due to the small number of traits measured and tested.

The whole genome haplotyping produced 4,762 haplotypes leading to a large reduction in the computational resources required while allowing a whole genome association test to be performed. 42,858 first round associations were performed with 10,000 permutations for each marker-trait pair to correct for multiple testing. Of these, 192 associations were significant with a p-value less than or equal to 0.005. A lower threshold was chosen to limit the number of second round associations to a manageable amount. With 50 significant associations, fresh weight had the most, followed by amino acids (36), sucrose (27), beta-alanine (21), myo-inositol (20), threonic acid (15), erythritol (12), protein (7) and starch (4).

After the haplotypes were expanded into individual SFP sites, 24,148 second round associations were performed (also with 10,000 permutations) allowing a more exact location

within each 25,000 nucleotide region to be pinpointed as the causal polymorphism. Of the second round associations, 1082 were significant with a p-value less than or equal to 0.05. Of these, 886 were in non-coding regions or introns and 419 lay in either the introns or exons of coding regions. The SFP sites in non-coding regions are dealt with in Chapter 4. The 419 SFPs in genes covered 163 genes. An over-representation analysis of Gene Ontology (Eilbeck et al., 2005) (GO) terms for each trait reveals many significantly over-represented terms (Appendix Section I) including "generative cell mitosis" and tRNA wobble uridine modification" for amino acid, "proton-transporting ATP synthase complex assembly" and "response to nematode" for beta-alanine, "DNA topoisomerase complex (ATP-hydrolyzing)" and "regulation of embryonic development" for fresh weight, "response to osmotic stress" and "glucosylceramidase activity" for myo-inositol, and "specification of organ position" and "cell fate commitment" for sucrose. The remaining traits had too few significantly associated SFP sites to perform statistically significant tests.

2.5. Discussion

Genotyping through hybridisation of genomic DNA to genome tiling arrays as applied in this study is characterised by an extraordinarily high degree of multiplexation surpassing all other genotyping methods except the use of whole-genome re-sequencing microarrays (Clark et al., 2007) and genome re-sequencing using second generation sequencing methods (Mardis, 2008). The Arabidopsis Tiling 1.0 array represents 3,039,991, 25 nucleotide sites of the *A. thaliana* nuclear genome, of which 3,035,275 occur uniquely, and can thus be simultaneously used for sequence variation detection defined as a substantial deviation of the obtained hybridisation signal from that of the reference (DNA of the genotype whose sequence has been used to design the array, Col-0 in the case of *A. thaliana*). Through the use of Tiling 1.0 arrays, we were able to monitor in a collection of 54 different accessions over 1,000,000 high quality SFPs. This substantially expands previous data sets (Borevitz et al., 2007; Clark et al., 2007). Furthermore, we could detect haplotypes and selective sweeps, and associate metabolic and phenotypic traits with the responsible genomic regions. We confirm that SFP monitoring is a viable and cost efficient alternative to SNP typing as proposed by Kim et al. (Kim et al., 2006) and also demonstrated by Borevitz et al. (Borevitz et al., 2007), which (in contrast to the latter approach) does not depend on prior knowledge of sequence variation.

2.5.1. Genomic data

Using known sequence information (the `2010 dataset´) (Nordborg et al., 2005) of 12 *A.*

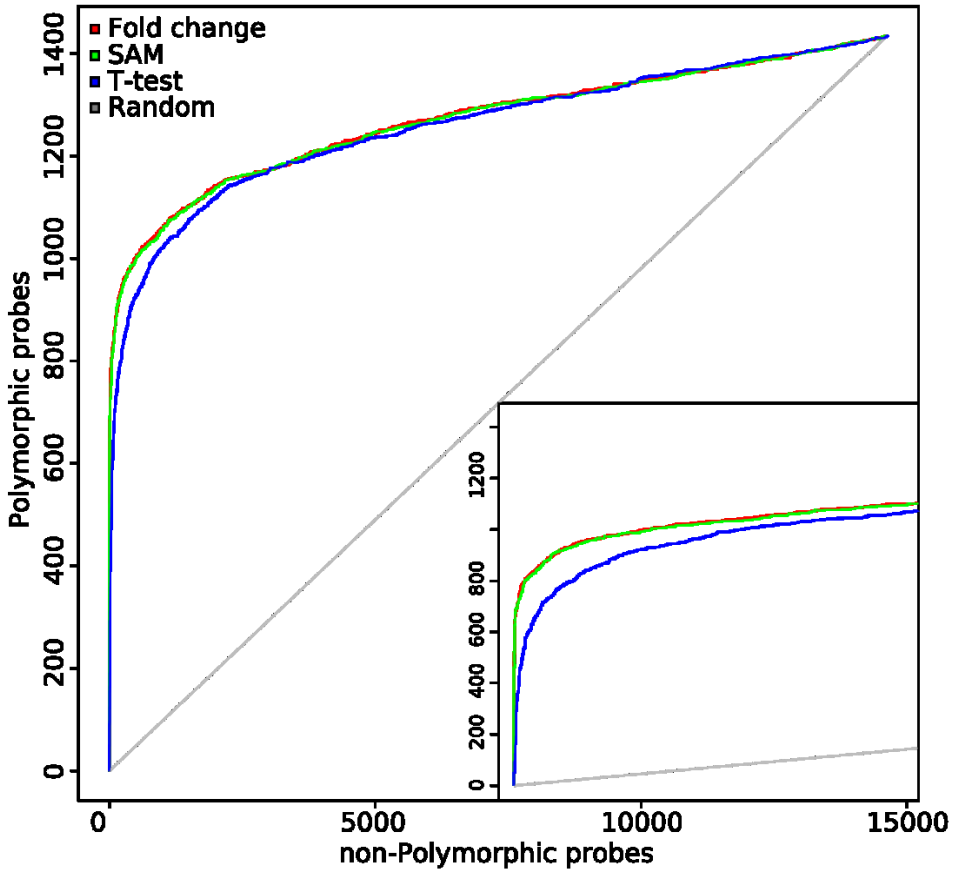

Figure 2.6 Method performance comparison

The log fold change method's performance is comparable to other commonly used methods. In general, it performs very well. For a typical ROC curve, the area under the curve (AUC) is 0.89, which is generally considered an indicator for a good predictor. However, because there are almost ten times more non-polymorphic probes than polymorphic probes, there are always a large number of false positives called. The inset shows the performance if the axes have the same scale. After the first 800 or so polymorphic probes are called, there is little gain in the number of true positives versus the number of false positives. For this reason, the threshold for calling SFPs is set very low (-1.5 log fold change).

thaliana accessions also used in this study, we confirmed that the method used here is suitable for detection of high quality SFPs. The FDRs observed here are similar or lower than those reported in previous studies (Borevitz et al., 2007; Borevitz et al., 2003), which may be due to improvements of the data normalisation and/or the metric used for SNP calling or due to higher stringency of SFP calling. The latter is reflected in a FNR of between 65.5% and 77.7%. In this method, a moderately high FDR is, to some degree, unavoidable. Due to the large numbers of probes that weren't polymorphic when compared to Col-0, and the inexact nature of microarrays, a large proportion of non-polymorphic probes will show a significant fold change decrease (Figure 2.6). We expect, however, that any false positives will be randomly distributed throughout the genome and their effects diluted. The high FNR caused

by the conservative threshold, may also cause problems for the associations performed. We expect a reduction in the number of significant associations due to missing polymorphisms.

Despite the limitations of the approach, which lie in coverage (~70% of the genome), obscurity (the exact mutation is not identified) and novel/duplicate sequence (unable to analyse novel sequence or the position of duplications), we were able to create high density SFP maps. On average we predicted approximately 86,000 SFPs per accession, which is equivalent to about one SFP per 1.6 kb of genomic sequence in each pair-wise comparison of an accession with the reference. About 25% to 30% of the SFPs were previously identified in each of the 10 accessions that were also investigated by Clark et al. (Clark et al., 2007) using resequencing microarrays. The limited overlap between the two data sets may at least in part be explained by the different SNP / SFP detection procedures used in the two approaches: The resequencing microarray hybridisation data were analysed based on differential hybridisation of 8 oligonucleotides for each position in the genome in order to infer base substitutions (Clark et al., 2007). This approach is sensitive to further adjacent polymorphisms that interfere with the hybridisation to all 8 oligonucleotides (Zeller et al., 2008). In contrast, SFP calling based on signal intensity comparisons of individual probes between hybridisations of DNA from a particular accession and from the reference is even enhanced at sites with two or more mismatches. During the course of the analysis, we observed that in addition to the number of SNPs in a probe, the position of an SNP in a probe and the %G+C of a probe all have significant effects on the probe signal. When we consider each of these, they may limit the number of the SFPs that can be called but they have little effect on the quality. Mid-range %G+C probes show a stronger effect change in probe signal than %G+C poor or rich probes (Figure 2.2). SNP position limits the SFPs we can predict to probes where the position of the SNP(s) is not on the edges (Figure 2.2). A greater number of SNPs in a probe will cause a decrease in the signal but have little effect on SFP quality (Figure 2.2).

The frequency of SFPs among the genomic features shows greater selective pressure on functional regions of the genome especially in coding areas, rRNA and tRNA. SFPs occurring in pseudogenes, transposons and ncRNA appear to be under much reduced selective pressure and, in some accessions, can accumulate to rates as high as one SFP every 400 nucleotides. The polymorphisms present in ncRNA could be explained by the pressure to conserve structure over sequence. Of the remaining genomic regions, it is remarkable that non-repetitive intergenic regions also experience a relatively low mutation rate, even lower

than that of coding regions. This could be indicative of undiscovered functional regions in the Arabidopsis genome.

In addition to individual SFP sites, regions of multiple, consecutive, SFPs were detected. Consecutive stretches of 10 or more polymorphic probes (equivalent to at least 350 nucleotides) were classified as deletions. Across the 54 accessions almost 4000 deleted regions of a mean length of approx. 1.1 kb were detected, most of which occur in only a single accession and on average 135 deletions were detected per accession with a total length of 974 nucleotides. According to the threshold of 10 consecutive SFPs, smaller deletions were not considered as they could not be readily distinguished from instances of multiple SNPs or small insertions/deletions (INDELS) in close vicinity. Furthermore, as only probes with a unique occurrence in the genome sequence were considered, this analysis was restricted to non-repetitive sequences. The fraction of deleted genome sequences here detected has thus to be regarded as a conservative minimum number. Conversely, regions were called duplicated if 10 or more adjacent probes displayed greater than a two-fold increase in hybridisation signal. Again, the total of about 1000 duplicated regions of an average length of approx. 1.1 kb (across the 54 accessions with a mean of 34 duplicated regions per accession has to be regarded as a minimum number. The notion of substantial variation in presence or copy number of genomic sequences in *A. thaliana* accessions is supported by the previous observation of 750 INDELS larger than 100 nucleotides detected in a 263 million nucleotide shotgun sequence of the Ler-1 accession compared to the Col-0 sequence, where 93% of the INDELS identified in total were shorter than 100 nucleotides (Jander et al., 2002). A remarkable variation in nuclear DNA content has been observed through very accurate flow cytometry of 21 *A. thaliana* accessions (Schmuths et al., 2004). This analysis revealed approx. 10 % higher DNA content in the accessions with the largest genomes as compared to those with the lowest values. Interestingly, the Col-0 accession, whose full genome sequence has been used to design the Arabidopsis Tiling 1.0 array, had the lowest genome content of the 21 accessions analysed. Throughout the 54 accession analysed in the present study we observed fewer duplications than deletions. While it remains to be analysed how much of the variation in genomic DNA across accessions is accounted for by repetitive DNA, our results indicate that the additional DNA in the accessions with larger genomes is not merely caused by increased copy numbers of sequences present in the Col-0 accession (which should have been detected as a much higher incidence of duplications). Thus, a substantial fraction of the genome complement of the *A. thaliana* species represented

by the additional nuclear DNA in many accessions is apparently yet uncharacterised.

The obtained polymorphism information was used to investigate the degrees of diversity among a broad range of functional gene classes. Whether they are duplications, deletions or individual SFPs, there are much greater numbers of polymorphisms than expected at random in disease resistance and defence related gene families (Figure 2.4). Across a broad geographic range, an encounter a large spectrum of pathogens can be expected and the high occurrence of polymorphisms in the resistance genes reflects the need for defence genes to adapt to the local pathogens and environmentally specific challenges. Among the genes that have the most significantly greater number of SFPs than expected, are the NBS-LRR, anthranilate synthase, F-box, terpene and monooxygenase gene families. All these gene families are either directly involved or implicated in plant defence. Transmembrane receptor, thioredoxin, leucine-rich repeat kinase, receptor-like kinase, and wall-associated protein kinase gene families and a subset of the proteins encoded by the other genes families identified to be highly polymorphic are involved in inter- and intracellular signalling potentially reflecting variation in responses to external and internal signals that may be relevant for local adaptation. F-box, NBS-LRR and transmembrane receptor genes have been previously identified as having a high number of polymorphisms (Borevitz et al., 2007; Clark et al., 2007).

On the other hand, many transcription factors, calmodulin binding genes, helicases, ABC transporters, glycosyl transferases, and acyl lipid metabolism genes contain significantly fewer SFPs than expected. Proteins encoded by these genes thus appear to be involved in more conserved processes.

While the detected duplications are so rare that they do not affect any of the chosen functional categories in most accessions, their frequency in some functional categories is particularly striking in a few accessions: For example, 22,000 nucleotides of the NBS-LRR disease resistance genes in accession El-0 are duplicated. This is significantly more than the expected 307 nucleotides. In a similar case, 37,000 nucleotides of the MADS-box transcription factors are duplicated in the Wei-1 accession.

2.5.2. Associations

The 54 accessions investigated here have been thoroughly characterised with respect to their biomass accumulated during growth under controlled conditions and they have been investigated for more than 90 metabolic parameters including protein and chlorophyll

contents, starch, a large number of low molecular weight metabolites and several enzyme activities (Sulpice et al., submitted). Substantial variation in biomass accumulation and many of the metabolic traits as well as complex relations between them were observed, probably due to underlying mutations. Of these mutations, those that are sufficiently beneficial are likely to be spread throughout the population resulting in a selective sweep leading to the hypothesis that we will observe more significant marker-trait associations within the selective sweeps than a random selection of other SFPs. As a first step towards addressing this question, associations were tested between focal and random SFPs of the sweeps and plant fresh weight as well as eight metabolic traits chosen as representatives of different groups of cellular components including structural and storage compounds as well as central metabolic intermediates. No significant associations were detected between the focal SFPs and the nine phenotypic traits (fresh weight, protein, amino acids, sucrose, starch, myo-inositol, beta-alanine, threonic acid and erythritol contents). However, this does not lead us to completely reject the hypothesis as the number of traits included in this analysis was limited. A more comprehensive selection of traits is required to execute a full investigation.

We observed that traits with many associations to genomic regions are likely to be controlled by a large number of genes, such as fresh weight and amino acid content. However, protein content produced an unusually small number of significant associations. Although the high FNR may reduce the number of significant associations found, if it is assumed that the reduction is uniform, than we can conclude that the number of protein associations is genuinely relatively low. This could be an indicator that protein production is regulated by a small number of genes or possibly that proteins are regulated by different genes in different accessions. Such a regulatory system would remain undetected by association mapping.

Some of the significant marker-trait associations found by the whole genome association mapping are obviously related to the over-represented Gene Ontology annotation of the genes that their markers lie within; such as amino acids (various amino acid synthesis and catabolism terms), beta-alanine (various H+ transport terms), fresh weight (embryonic development) and myo-inositol (cell-signalling). Considering the relatedness between the genes identified by the significantly related SFP sites and the traits they are suggested to influence, it is probable that the association is not only statistically significant but also biologically significant as well.

2.5.3. Natural selection mapping versus molecular genetics.

We have identified genomic regions with genetic variation that is associated significantly with growth-related metabolic traits in greenhouse conditions. Even though the phenotypic analysis was not conducted under natural conditions or in nature itself, the associations strongly suggest the location of variants that have a robust effect on complex traits. It can not be excluded, though, that in nature/ evolutionary history, a different phenotypic effect of the sweep region may have been selected. Furthermore, if the region is large, then the phenotypic effect could hitchhike with a selected mutation that is actually responsible for another type of phenotypic variation.

Like every other method, natural selection mapping may have some biases. By focusing on genes with a strong signal of selection that readily sweep through the population, genes with little epistatic interaction and with low pleiotropic effect are being detected. For example, major functionally different alleles for two flowering time genes FLC and FRI are in significant LD across chromosomes in a central European population, which is maintained by epistatic selection and has significant fitness consequences (Korves et al., 2007). However, even though we identify significant LD across the loci, they do not show up as significantly selected or extraordinarily polymorphic regions in our analysis. In the near future, this limitation will be circumvented by analysing hierarchically structured samples from the complete species range and the inclusion of additional methods for detecting non-neutral evolution such as LD between distant haplotypes. In *A. thaliana*, efforts are underway to re-sequence a large number of genomes (www.1001genomes.org) that will lead to much improved detection of sweeps.

A critical issue in natural selection mapping is the age of the sweeps which may occur. Even though phenotypic variation is associated with the sweep, it may not necessarily be the trait that evolved under selection but may even be a (slightly) disadvantageous trait. If the sweep occurred recently, there may not have been much time for compensatory mutations to neutralise this phenotypic effect (Smith and Macnair, 1998).

Nevertheless, this analysis represents a precedence of a viable procedure to identify genomic regions carrying functionally relevant variation in (probably regulatory) genes, which is potentially involved in adaptation. On the one hand these results open the exciting opportunity to identify and characterise the genes responsible for the phenotypic variation of growth-related metabolic traits through the investigation of the sequence variation of the

identified regions across the accession population and the functional analysis of the encoded genes and their alleles through phenotypic characterisation of genetic substitution lines, mutants, and transgenic plants grown in a variety of conditions. On the other hand they indicate how powerful this approach will become in the near future upon availability of deep genotype/re-sequencing information and phenotypic data of hundreds of *A. thaliana* accessions.

Chapter 3.

ncRNA functional prediction

3.1. Abstract

The study of non-coding RNA genes has received increased attention in recent years fuelled by accumulating evidence that larger portions of genomes than previously acknowledged are transcribed into RNA molecules of mostly unknown function, as well as the discovery of novel non-coding RNA types and functional RNA elements.

Here, we demonstrate that the structural abstraction of RNA molecules offers a computational means for a reliable identification of their function. Our results show that specific properties of graphs associated with the predicted RNA secondary structure reflect functional information. We introduce a computational algorithm and an associated web-service (GRAPPLE) for classifying non-coding RNA molecules into Rfam families based on their graph properties. Unlike sequence-similarity-based methods and covariance models, GRAPPLE is demonstrated to be more robust with regard to increasing sequence divergence, and when combined with existing methods, leads to a significant improvement of prediction accuracy. Furthermore, graph properties identified as most informative are shown to provide an understanding as to what particular features render RNA molecules functional. Thus, GRAPPLE may offer a valuable computational filtering tool to identify potentially interesting RNA molecules among large candidate datasets.

3.2. Introduction

Non-coding RNA genes (ncRNA) are critical components of many biological processes including translation (tRNA, rRNA), RNA splicing (ribozymes), gene regulation through mRNA hybridisation (miRNA, piRNA), gene regulation through metabolite binding (riboswitches), and RNA methylation and pseudouridylation (snoRNA) (Meyers et al., 2008). Functions such as translation and RNA splicing have been long considered to be the sole role of ncRNA. However, new and unexpected functions have been recently discovered, revealing that RNA molecules assume highly diverse functions and are more actively involved in biological processes than previously thought (Mattick, 2007). The intensified study of ncRNA and search for new functional roles of RNA is further propelled by the realisation that a larger portion of intergenic space than previously acknowledged is actually transcribed. For instance, 85% of the fruit fly genome (Manak et al., 2006), 62% of the mouse genome (Claverie, 2005), and a staggering 93% of the human genome (Birney et al., 2007; Weinstock, 2007) have been reported as transcribed. Understanding the functional role of this otherwise seemingly wasteful transcription requires the analysis of large amounts of genomic data. Thus, computational methods have a great potential to contribute significantly toward this work by predicting potentially functional non-coding regions and their respective function.

The structure of ncRNA is thought to provide insight into the biological function (Mathews and Turner, 2006). In the folding process, characteristic nucleotide base-pairing and stacking interactions play significant roles and are governed by molecular forces acting on and within any molecule in aqueous solutions (*e.g.*, electrostatic interactions) (Tinoco et al., 1971). The adopted shapes or folds can be highly complex and are capable of carrying out a variety of molecular functions, such as binding metabolites and proteins with high specificity (Lee et al., 1993; Mironov et al., 2002; Nahvi et al., 2002; Schilling et al., 2004; Winkler et al., 2002). RNA is particularly suited for hybridising with nucleotide sequences allowing for highly specific targeting of genes and genomic regions (Brouns et al., 2008; Kurihara et al., 2008; Nakashima et al., 2007). Furthermore, it is conceivable that two ncRNA molecules with completely different nucleotide compositions would still fold to form the same structure and have the same function. For example, the secondary structure of tRNA has a characteristic cloverleaf shape; however, the nucleotide composition of tRNA can vary to the degree that two tRNAs can have completely different sequences. Thus, methods that incorporate ncRNA structural, and not just sequence information, are required for an accurate

prediction of function.

Due to the importance of RNA structure, several computational RNA folding tools have been developed, such as: MFOLD (Zuker, 2003), RNAFOLD (Hofacker, 2003), VSFOLD (Dawson et al., 2006), EVOFOLD (Pedersen et al., 2006) and SFOLD (Ding and Lawrence, 1999). The majority of these algorithms work on an input sequence to determine the folded secondary structure that minimises the free energy by optimising the intramolecular basepairing. The input sequence may come from publicly available repositories, *e.g.*, Rfam (Griffiths-Jones et al., 2003) which currently contains 636,138 sequences, grouped in 603 ncRNA families, that are largely computationally annotated (Griffiths-Jones et al., 2005).

The listed structure-prediction tools are fast and accurate when operating on sequences of less than 200 base-pairs; however, they are not suitable for longer sequences (Freyhult et al., 2005). The accuracy of the predicted structure has been improved by algorithms that use multiple sequence alignments to produce a consensus structure. Another, relatively recent class of algorithms are designed to fold pseudoknots—structures where each bonded base pair is not required to be bounded by another bonded pair of bases that are closer to the ends of the molecule. An example of a pseudoknot is the "kissing hairpin" where the loops of two hairpins are bound to each other. The problem of folding pseudoknots was shown to be NP-complete and many tools do not attempt to fold them (Lyngso and Pedersen, 2000). There is also an algorithm that predicts the three-dimensional structure (as opposed to secondary and pseudoknotted structures) of RNA molecules (Das and Baker, 2007). This method involves using fragments of RNA whose three-dimensional structure has been experimentally determined as building blocks to assemble the shape of an investigated molecule. Any methods developed to predict ncRNA function are fully reliant on the ability for these tools to predict the structure accurately.

There are very few tools that deal with the classification of functional versus non-functional RNA sequences. One attempt to develop such a tool investigated the idea that the minimum free energy (MFE) of functional RNA sequences should be lower than that of random, shuffled and non-functional genomic sequences (Rivas and Eddy, 2000). In the study, MFE was identified to be largely unhelpful except in a later study which discovered that MFE can be used to identify miRNA (Bonnet et al., 2004). Other studies calculate the thermodynamic stability of multiply aligned structures as a means of identifying functional RNA (Washietl et al., 2005) and have been applied to the genomes of *Saccharomyces cerevisiae* (Steigele et al., 2007) and *Plasmodium falciparum* (Mourier et al., 2008).

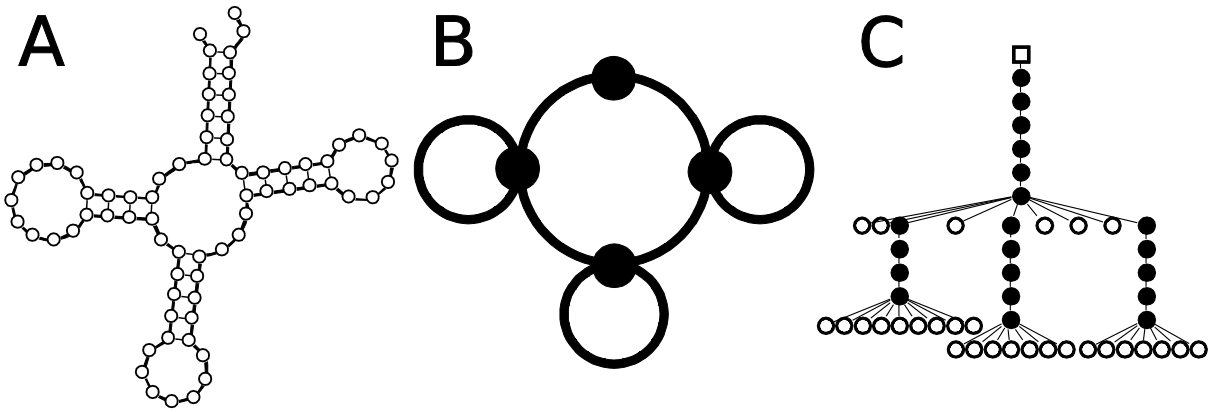

Figure 3.1 Representations of RNA structure using graphs.

A: *A typical tRNA structure represented using the bracketed graph representation. Nucleotides are represented as nodes (open circles) and bonds (both base-base hydrogen bonds and backbone ester bonds) as edges. The secondary structure of the tRNA is reminiscent in the shape of the graph. This is the chosen graph representation in the current paper.* ***B:*** *The dual-edge graph for the same tRNA is shown. Stems are converted to nodes and loops to edges. Information about dangling ends is lost in this representation.* ***C:*** *Planar tree representation uses a special node for the root (5'/3' end of the structure) depicted here as an open square. Base-pairs are converted to "stem" nodes (closed circles) and loop nucleotides are converted to "loop" nodes (open circles). The tree is built by following the strand from 5' to 3' and the order of children is important. Information about dangling ends is also lost in this representation.*

Assigning unannotated RNA sequences to an Rfam family is better investigated and there are a wide variety of ncRNA family specific predictors, most of which focus on miRNA (Cao and Chen, 2006; Lim et al., 2003; Myslyuk et al., 2008; Zhang, 2005). Covariance models, as general predictors, are used to identify nucleotide pairs which vary together across multiple alignments and are thus likely to be bonded in secondary structure (Eddy, 2002). Such models require multiple alignments and are computationally time consuming, limiting the number and type of sequence that may be processed.

A large portion of recent RNA-related research applies concepts developed in graph theory to the analysis of RNA structure (Fera et al., 2004; Janssen et al., 2008; Kim et al., 2004). A graph is an abstraction of the relationship among objects, which uses nodes to represent the objects and edges to represent the relationship between two objects. There are many ways to represent RNA structure with graphs, including the bracketed (where nucleotides a converted to nodes and bonds to edges), planar tree (where base-pairs are converted to "stem" nodes and loop nucleotides are converted to "loop" nodes while following the molecule from 5’ to 3’) and dual graph representations (where stems are converted to nodes, while loops to edges) (Figure 3.1), each with different advantages and disadvantages including information loss and

complexity of calculation (Fera et al., 2004). Graph topology derived from RNA structure has also been used to assign Rfam family (Karklin et al., 2005). Although the ability to discriminate between functional and non-functional genes was not demonstrated, this approach appears quite successful in terms of classification.

The structure of a graph can be employed to define and analyse different properties that could reflect the characteristics of the process or entity modelled by the graph. A property can be defined on the level of graph constituents (*i.e.*, nodes and edges) or on the level of the graph itself. Furthermore, computing a property may require limited or full knowledge of the graph. Based on these two criteria (level of detail and required knowledge of the graph), graph-theoretic properties may be classified into local (using limited knowledge of the graph and referring to a graph's constituent), local-global (using full knowledge of the graph and referring to a graph's constituent), and global (using full knowledge of the graph and referring to the graph itself). Thus, graph representations of RNA molecules offer a means to capture both local-global and global structural properties that can be used to deduce the large- and small-scale structural, and therefore functional, differences between the molecules.

Here, we go another level of abstraction higher than previous methods and address the question of how a set of selected graph-theoretic properties derived from a graph representation for predicted RNA secondary structures can be used as characteristic features for the classification of RNA molecules. Among the immense number of existing graph-theoretic properties, we select several representatives based on the following three criteria: (1) polynomial-time computation, (2) relevance to local and global levels of the graph, and (3) usage in complex network research. As a means of exploring the relationship between graph properties and Rfam families, we attempt to recall the Rfam families of ncRNA sequences using support vector machines (SVMs) trained on the selected graph properties. Furthermore, we show that graph properties can be employed to differentiate between functional and non-functional sequences as well as predict a likely function. In this study, a small number of graph properties are identified as most relevant for the correct classification of ncRNAs and their interpretation is demonstrated to shed light on structural properties that may render RNA molecules functional compared to their non-functional counterparts.

3.3. Methods

3.3.1. The data set

The seed-RNA sequence alignment dataset was obtained from Rfam release 9.0 (Griffiths-

Jones et al., 2005) and all redundant identical sequences were removed using CD-HIT (Li and Godzik, 2006), yielding 18,974 unique sequences for analysis. These sequences were split into 210 Rfam and 8 compound families, which were formed out of several smaller related Rfam families (CD-box, HACA-box, internal ribosome entry sites, leader sequences, miRNA, riboswitches, ribozymes, and scaRNA). All RNA sequences were folded into their predicted secondary structures using RNAFOLD (Hofacker, 2003).

3.3.2. Calculating graph properties

The bracketed graph representation was used to represent the predicted structure (Figure 3.1). It was calculated by converting all nucleotides to nodes and all bonds between nucleotides (both ester and hydrogen) to edges.

From the three different ways in which a property can be defined and calculated, here we used the summary statistics for the local-global properties, since they provide insight not only on the global level of the graph itself, but also on the level of its nodes and edges. The employed statistics (mean and variance) allow for a uniform way of summarising the distribution of values an investigated local property may assume. For instance, the node-betweenness used in our analysis is given by the mean and variance of the distribution of node-betweenness values over all nodes of a graph. Similarly, we used bibliographic coupling as given by the mean and variance of bibliographic couplings over all pairs of nodes.

All properties were calculated using the igraph R package (Csárdi and Nepusz, 2006) for complex networks with our own extensions to the presently implemented algorithms that facilitate the extraction of the graph representation and calculation of the necessary summary statistics. We focused on the following global properties: number of articulation points, diameter, girth, density, and transitivity, together with the local-global properties (given by the mean and variance): Burt's constraint, path length, node betweenness, edge betweenness, degree, co-citation coupling, bibliographic coupling, coreness, and closeness (A brief definition of all graph properties used in this study is provided in the Appendix Section II).

3.3.3. SVM training and testing

We used the following procedure for training and testing all support vector machines (SVMs): First, we produced matched training/testing sets with randomly selected, but non-overlapping sequences and matching graph property sets. SVMs were then created from the training sets using LIBSVM software (Chang and Lin, 2001). All graph properties for the

training sets were initially scaled between -1 and 1 to prevent graph properties with larger numerical ranges from dominating those with smaller ranges. A 10-fold cross validated grid search, based on the training set, was used to optimise the initial parameters C (the cost parameter) and γ (the kernel width). In addition, the SVM was trained on the full training set using the optimised values. The radial basis function (RBF) kernel was employed as it is able to identify non-linear relationships between class-labels and features (graph properties), requires fewer hyper-parameters, and presents fewer numerical difficulties than other kernels. The testing sequences were in turn submitted to test the SVMs, and results are reported in the Results section. Each SVM was trained 100 times with different sets of random sequences.

The importance of the graph properties was calculated using the F-score (Chen and Lin, 2006). The F-score is a simple measure that discriminates between two sets of real numbers. Given training vectors $x_k, k = 1..m$, if the number of positive and negative instances is n_+ and n_-, respectively, then the F-score of the i^{th} feature is defined as:

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}, \text{ (Eq. 1)}$$

where x_i, $x_i^{(+)}$, and $x_i^{(-)}$ are the average of the i^{th} feature of the whole, positive, and negative data sets, respectively; is the i^{th} feature of the k^{th} positive instance, and is the i^{th} feature of the k^{th} negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. The larger the F-score, the more likely it is that this feature is more discriminative. This algorithm is available using FSELECT which is available on the LIBSVM internet site.

3.3.4. Functional vs. non-functional RNA sequence prediction

SVMs were trained to differentiate functional from non-functional RNA using graph properties. Sequences available from Rfam are considered functional and comprise the set of all functional sequences (here, mRNA is considered non-functional). A non-functional set was created by shuffling each Rfam sequence once while preserving dinucleotide content using USHUFFLE (Jiang et al., 2008). 200 functional and 200 non-functional sequences were randomly chosen for the training and testing sets with each family having an equal chance of being chosen, yielding 400 sequences for each set. After classification, the important graph properties were determined by calculating the F-score.

3.3.5. Predictive power of graph properties

We attempted to determine whether graph properties alone can be used to recall Rfam families. To remove the influence of sequence similarity and length, we filtered the training and testing sequences in two ways.

First, to account for sequence similarity, we created diverging testing and training sets. A distance matrix for each family was created by an all-against-all comparison of sequences within a family using the similarity score provided by CLUSTALW pairwise alignments (Larkin et al., 2007) (http://www.clustal.org). Each family was then divided into diverging training and testing sets, where the greatest similarity between a member picked from a training set and a member chosen from the paired testing set would be less than or equal to a given threshold. We set the initial threshold to give a maximum similarity between the two sets of 90 percent identity (%id) to allow the training and testing sets to become highly similar but not identical. We then decreased this threshold in steps of 10%id. As any two completely random RNA sequences are expected to have 25%id due to random chance, we set the lower bound to 20%id, thus creating a total of eight sets (20, 30, 40, 50, 60, 70, 80, 90%id). The divergence algorithm used is presented as Algorithm 3.1.

Second, graph properties were calibrated for the potential bias introduced by length and GC content (%G+C). A set of random sequences was generated, in which all combinations of the lengths 50-1000 nucleotides (in steps of 50) and the %G+C 10-100% (in steps of 10) were represented 100 times, producing a matrix for each graph property with 10,000 entries. The graph properties of each sequence were then calibrated by dividing by the entry with the closest length and %G+C in the corresponding calibration matrix. The F-score was also used here to calculate the predictive power of each graph property.

As the maximum similarity between the training and testing set decreases, the number of available sequences also decreased and many families became too small to be used leaving, finally, 18 families for analysis. Training sets were restricted to 50 random members, while testing sets were restricted to 20 from each family.

The sensitivity (Q^D) and specificity (Q^M) of SVM-based predictions for each individual family were calculated using the following equations:

$$Q_i^D = \frac{z_{ii}}{\sum_j z_{ij}} \text{ (Eq. 2) and } Q_j^M = \frac{z_{jj}}{\sum_i z_{ij}}, \text{ (Eq. 3),}$$

where z is an entry in a confusion matrix, i is an index for the actual family and j is an

Algorithm 3.1: Diverging sequences

Given: Distance Matrix: `D`, Distance threshold: `thr`

```
1:  Initialise the training and testing sets
2:  Seed the sets with the furthest random sequence pair
3:  Initialise empty training_candidate and testing_candidate sets
4:  for all sequence do
5:      if distance between sequence and training_seed <= thr do
6:          add sequence to training_candidates
7:      if distance between sequence and training_seed <= thr do
8:          add sequence to testing_candidates
9:  while both candidate sets have members do
10: add farthest training_candidate sequence from testing to training
11: adjust all testing_candidate distances to distance from new training
```

index for the predicted family.

We also investigated the possibility of combining the results of the SVM with the sequence-based assignment of Rfam family using WUBLAST in order to improve accuracy. For each sequence, the SVM produces probabilities (p-score, note that this score is different to the p-value) that the sequence belongs to each ncRNA family. The sum of the p-scores totals to 1. Similarly, for each sequence, we produced an E-value for each family using WUBLAST. This E-value was adjusted to the same scale as the SVM p-score by calculating the inverse E-value as a fraction of the total inverse E-values:

$$pscore_{BLAST} = \frac{\frac{1}{e_i}}{\sum_{k=0}^{n}\frac{1}{e_k}}, \text{(Eq. 4)}$$

where e_i is the E-value obtained for an individual family and e_k is the sum of E-values over all families. The two values were then combined linearly using a weighting factor, α, as follows:

$$pscore_{MERGE} = (1-\alpha) \times pscore_{BLAST} + \alpha \times pscore_{SVM}. \text{(Eq. 5)}$$

As a result, we obtained for each sequence a merged p-score for each family that, although not considered a probability, indicates how likely the sequence belongs to that family. The

family with the highest p-score was assigned to the sequence.

Standalone WUBLAST (Gish, W. (1996-2004); http://blast.wustl.edu), the INFERNAL package (Eddy, 2002) (http://infernal.janelia.org) and HMMER (Eddy, 1998) (http://hmmer.janelia.org) package were used to provide references to methods which are expected to perform either poorly and well on the diverging training sets. As the sets diverge, the performance of sequence comparison based methods, such as WUBLAST, should degrade whereas structure based methods would ideally remain stable. The comparison was performed using the same training and testing sets. A description of how these methods were applied can be found in the next section

3.3.6. Comparison to other methods

To compare our method to existing tools, we a chose representative method from each class of classifier presented in a previous study used to benchmark a number of other tools (Freyhult et al., 2007). We chose WUBLAST from the homology-based methods, HMMER from the Hidden Markov Model-based methods and INFERNAL from the covariance model-based methods. Training sets of 50, 100 and 200 sequences per Rfam family were generated, which resulted in 25, 8 and 3 Rfam families of sufficient size for each training set respectively. All tools were used with the default settings following the same procedure described in the previous section.

For comparison with WUBLAST, each training set was split into the constituent Rfam families and converted into blastable databases. The testing set was then blasted against each database, using an E-value threshold of 100, resulting in a set of E-values for each sequence that measures how well it matched each Rfam family. Sequences were then classified according to the family with the lowest E-value.

A similar procedure was followed using INFERNAL and HMMER. Training sets were split into constituent Rfam families and aligned using MUSCLE (Edgar, 2004) (http://www.drive5.com/muscle). From each family alignment, covariance and Hidden Markov models were built. The testing set was then searched using each of the models and each sequence was scored on how well it matched a given family. Sequences were classified according to the best identified matching family.

3.3.7. Generalised classifier

Our general classifier is named GRAPPLE, an acronym for Graph Property-based Predictor and Likelihood Estimator. 400 random members were taken from each family in the non-redundant Rfam dataset and split evenly into testing and training sets, creating sets with equal family representation. Families with less than 400 members were ignored leaving 23 families for analysis. To measure the confidence of classification, we applied Shannon entropy to the p-scores that a sequence produces by using the equation:

$$entropy = -\sum_{i=0}^{n} pscore_i \times \log\left(pscore_i\right), \text{(Eq. 6)}$$

where $pscore_i$ is the p-score for a family and n is the number of families.

3.3.8. Availability

A tool based on this method called the GRAph Property based Predictor and Likelihood Estimator (GRAPPLE) is available as a web-based tool at: http://grapple.mpimp-golm.mpg.de. Programmatic access using the Simple Object Access Protocol (SOAP) is available at: http://grapple.mpimp-golm.de/cgi-bin/grapple.soap.py.

3.4. Results

In the current paper, we developed three approaches to investigate graph properties and their ability to reflect the functional information of RNA molecules. In the first approach, we tested the ability of graph properties to discriminate between functional and non-functional RNA molecules. In the second, we removed any bias that may be introduced through sequence similarity, length and GC content (%G+C) by using calibrated and diverging training and testing sets to test the predictive power of the graph properties alone when predicting the Rfam family of an ncRNA sequence. In the third, we removed the limitations imposed in the second approach and compared the ability for the developed method to predict Rfam family to other established tools.

In the first approach, an SVM-based classifier was able to classify RNA sequences into functional and non-functional classes with Matthew's Correlation Coefficients (MCC) ranging between 0.61 and 0.98 with an average MCC of 0.87, and sensitivity and specificity of 0.73 respectively. This indicates that graph properties can be used to identify functional RNA sequences and performs better than random assignment (MCC = 0). The discriminatory power of each graph property was then calculated using a measure called the F-score (see Methods) (Figure 3.2). This score revealed that the "number of articulation points" possessed

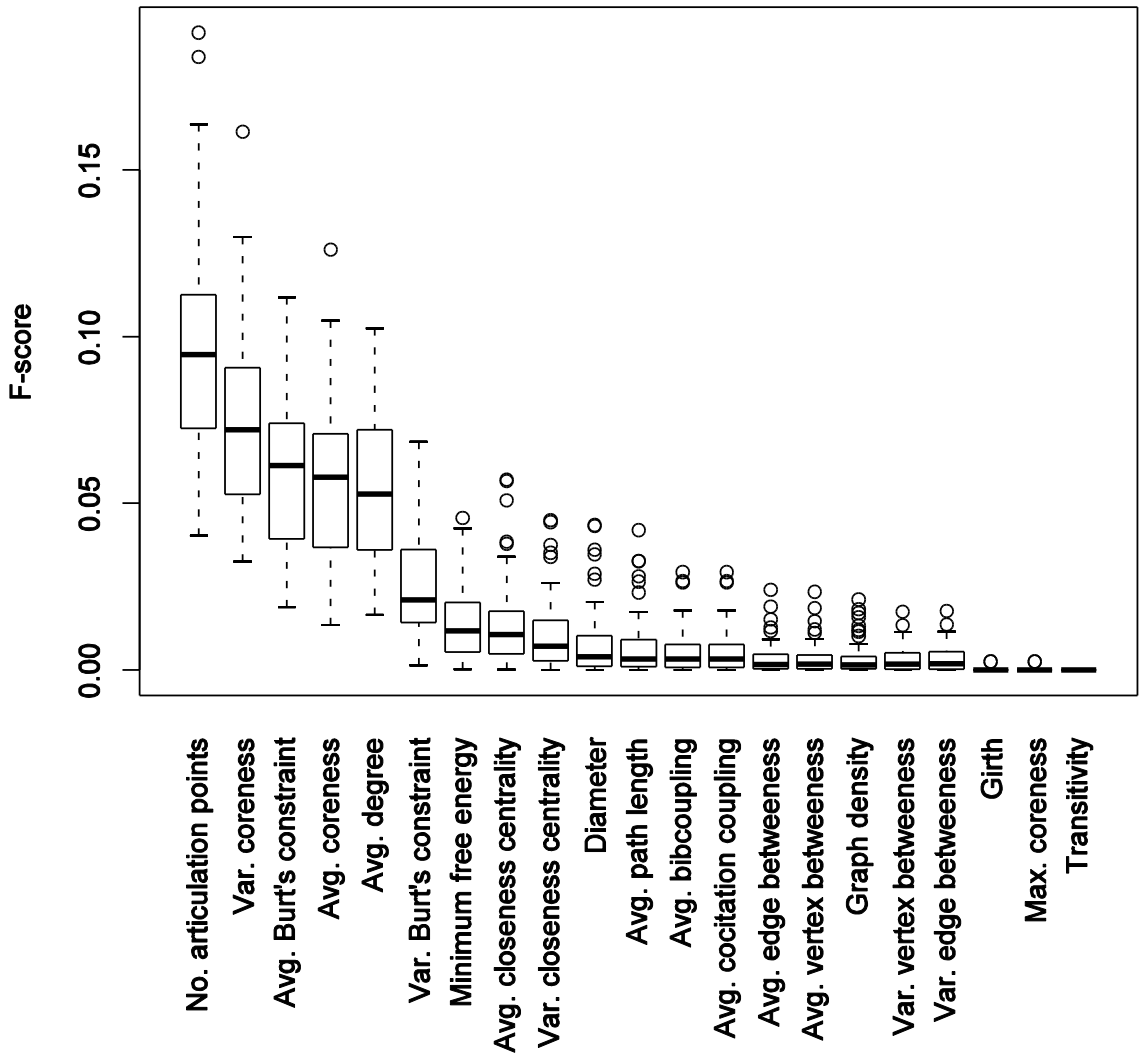

Figure 3.2: Graph property discriminatory power for functional vs. non-functional classification.

The discriminatory power of each graph property was determined by calculating the F-score (Equation 1) with larger F-scores indicating more relevant properties. The distribution of F-scores is shown for each graph property as a box plot where the middle bar is the median, the outer edges are the 10 and 90 percentiles and the edges of the box are the 25 and 75 percentiles. Outliers are shown as circles. When classifying functional versus non-function RNA, we find that the "number of articulation points", "variance of coreness", "average coreness", "average Burt's constraint" and "average degree" consistently have significantly higher F-scores than the other graph properties

the most discriminatory power with an average F-score of 0.094 followed by the "variance of coreness" (0.080), "average coreness" (0.062), "average Burt's constraint" (0.062) and "average degree" (0.056). The F-score dropped off significantly for the remainder of the graph properties along with the MFE. "Girth", "maximum coreness" and "transitivity" had little or no discriminatory power and were included to provide baseline support for high-scoring graph properties.

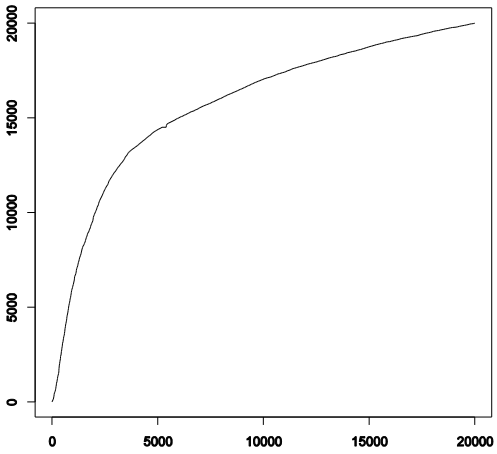

Figure 3.3: ROC curve for functional vs. non-functional classification.

By changing the threshold at which we label a sequence as functional or non-functional, we obtain a plot that enables us to find the ideal threshold. Such a plot reveals that sequences are labelled significantly better than random chance. The average Matthew's Correlation Coefficient was calculated to be 0.87 and the area under the curve to be 0.79.

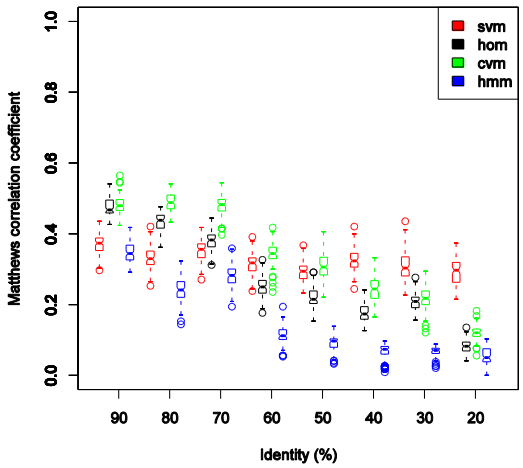

Figure 3.4: Rfam family recall for length-calibrated, diverging sets.

The performance of the SVM method (svm) is significantly decreased when classifying sequences calibrated for length and %G+C indicating that RNA families have distinct lengths and %G+C that affect the graph properties. However, the performance remains stable as the sequences diverge whereas the performance of homology methods (hom), covariance models (cvm) and Hidden Markov Models (hmm) degrades sufficiently that SVMs still outperforms them at maximum similarities of 50% and below.

The method used in this approach produces two p-scores that indicate the probabilities that the query sequence is functional or non-functional whose sum equals one. These p-scores shouldn't be confused with the p-value produced by statistical tests. By changing the p-score at which we accept a sequence as functional, we can decrease the number of false positives that we predict (Figure 3.3).

The second approach explored the idea that graph properties are able to reflect RNA structure and function in greater detail by attempting to recall the correct Rfam family without the influence of sequence similarity, length and %G+C. To control for sequence similarity we created diverging testing and training sets. To control for sequence length and %G+C, we performed a calibration using generated random sequences of various lengths and %G+C (see Methods). SVMs trained on calibrated graph properties classify RNA sequences with an

average MCC of 0.32 (Figure 3.4); *i.e.*, substantially above the expected rate when guessing. This value is relatively stable at all 8 thresholds of sequence divergence as it varies between 0.29 and 0.37, and shows that the method is robust at all levels of sequence divergence.

To illustrate that the training and testing sets really do diverge, we tested the dataset with WUBLAST. When using WUBLAST to classify the sequences based on sequence similarity, there is a significant drop in the average MCC from 0.48 (90% maximum similarity between training and testing sets) to 0.08 (20% maximum similarity) that demonstrates this divergence (Figure 3.4). To demonstrate that structure should remain the same as the sets diverge, we used INFERNAL, a covariance model-based method. Unexpectedly, we observe a drop in the average MCC as the sets diverge from 0.49 to 0.12 similar to WUBLAST (Figure 3.4) implying that within an Rfam family, the structures do not folded similarly. HMMER was also used on the diverging training and testing sets and performs significantly worse at all levels of divergence than the other methods.

When other methods, such as WUBLAST, are augmented with graph properties, the resulting predictions show a significant increase in accuracy (Figure 3.5). This suggests that the graph properties are able to access information independent of the sequence similarity used by WUBLAST to make predictions. At all levels of divergence, a 10-20% increase in prediction accuracy is observed.

In the cases, where the training and testing sets have at most 20% similarity, the F-scores of the graph properties, fall into four broad categories (Figure 3.6). The “average Burt's constraint”, “average degree” and “average coreness” have the highest F-scores while “girth”, “maximum coreness” and “transitivity” do not contribute at all to the SVM. The remaining graph properties fall into two roughly equal groups that have average F-scores around 0.6 and 0.4. Thus, the important graph properties which determine functional versus non-functional RNA sequences and those that determine the Rfam family differ slightly. While the “average Burt's constraint”, “average degree”, and “average coreness” remain among the most important, the “number of articulation points” and the “variance of coreness”, which were important for functional versus non-functional classification, are ranked among the least important for assigning Rfam families.

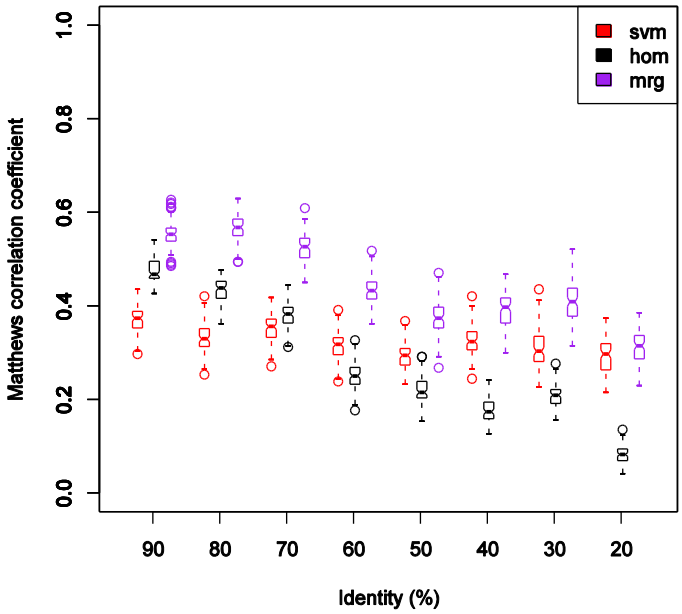

Figure 3.5: Linear combination of graph property based SVM and BLAST for calibrated, diverging sets.

By merging the results of the WUBLAST (hom) and SVM (svm) methods, we obtained improved classification at all thresholds of divergence (mrg). This indicates that both methods capture independent information that allows more accurate classification when combined.

The SVM method does not work evenly across all Rfam families. When the sets are maximally divergent, the families SECIS (0.96 sensitivity, 0.58 specificity), Intron gp II (0.72, 0.73), 5S rRNA (0.63, 0.70), tRNA (0.42, 0.83) and MIRNA (0.79, 0.46) all perform well with high specificity, high sensitivity or both. IRES, LEADER and SRP are associated with the worst sensitivity and perform only slightly better than random assignment of Rfam families; 0.07 versus 0.05.

As the graph-property-based prediction approach may capture relevant aspects of RNA molecules that are not properly reflected by sequence similarity searches alone, (as demonstrated by the more robust behaviour of our graph-based method when tested on diverging sequence sets; Figure 3.3), combining both methods may result in increased prediction performance compared to each individual approach. By combining a p-value calculated from the WUBLAST E-value and the SVM p-score in a linear fashion with a properly chosen weighting factor (0.5), MCC values higher than those produced by each method individually (Figure 3.5) were obtained. The average MCC for the combined methods

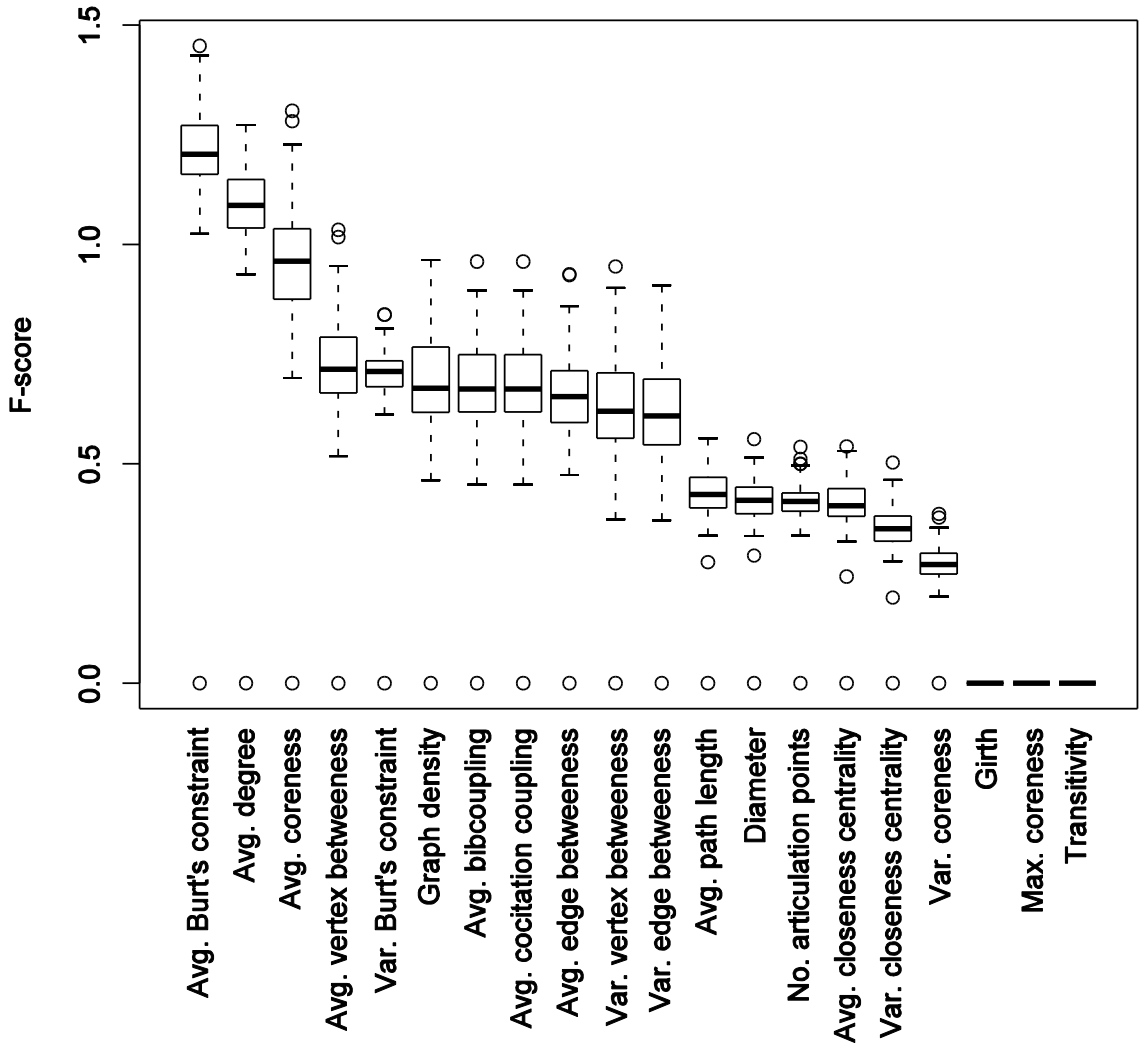

Figure 3.6: Graph property discriminatory power for Rfam family assignment.

The discriminatory power of each graph property is calculated to identify the important graph properties (Equation 1). The most important properties are the "average Burt's constraint", "average degree" and "average coreness".

is 0.446 and ranges from 0.313 in sets that are 20% similar to 0.567 in sets that are 90% similar.

Although we applied rigorous calibration to the sequences to identify whether the graph properties themselves were responsible for prediction or the influences from sequence similarities and length, it would be imprudent not to use this information when constructing an SVM intended for actual classification of RNA sequences outside the testing protocol. Thus, for comparison with other methods, the third approach used non-calibrated graph properties and imposed no similarity restrictions between the training and testing sets.

Having established that graph properties are sufficient to distinguish between Rfam families, we took a third approach that compared the performance of the SVM method to WUBLAST, INFERNAL and HMMER using training sets with 50, 100 and 200 sequences per Rfam family. The size of the training set and the number of families to be classified has a significant impact on the performance of the tested method (Figure 3.7). The SVM-based method shows performance increase from a median MCC of 0.88 for training sets of size 50 to 0.96 for training sets of size 100 and 0.98 for training sets of size 200. INFERNAL also show an increase (0.96, 0.99, 0.99) whereas WUBLAST remains stable (0.97, 0.95, 0.96) and HMMER shows a decrease (0.97, 0.94; 0.88). The results indicate that SVMs trained on graph properties are able to perform slightly better than homology-based methods and slightly worse than covariance model-based methods.

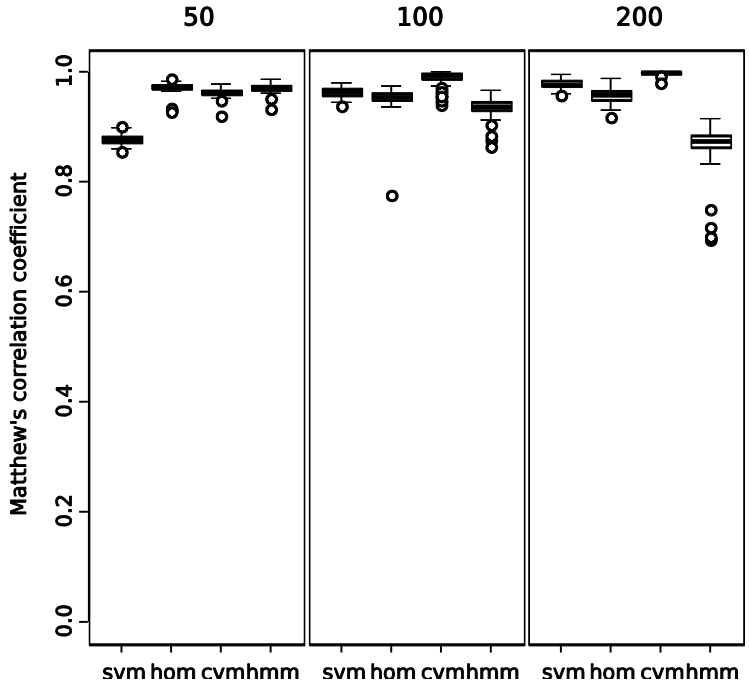

Figure 3.7: Comparison to other methods.

Four difference classes of method were compared; SVM (svm), homology (hom), covariance models (cvm) and Hidden Markov Models (hmm). Each method performs differently depending on the size of the training set. The SVM based method performs with a median MCC of 0.88 when trained with sets of 50 sequences. The median MCC increases considerably for training sets with 100 and 200 sequences to 0.96 and 0.98 respectively. With larger training sets, SVMs compare favourably with the other methods such as homology methods, which had median MCC values of 0.97, 0.95 and 0.96 for training sets of size 50, 100 and 200 respectively, covariance models (0.96, 0.99, 0.99) and Hidden Markov Models (0.97, 0.94, 0.88). Hidden Markov Models show a reverse trend that may be explained by the impact that the increased variability has on the transitional probabilities.

The developed tool, GRAPPLE, predicted Rfam families with an average MCC of 0.96

ranging between 0.83 and 0.98. Sequences which were correctly classified show much lower entropy (see Methods) than those that are incorrectly classified (Figure 3.8). Entropy was used as a means of measuring the confidence that a classification is correct. Sequences that the methods are able to confidently classify have a much higher probability of belonging to a certain family and thus show lower entropy across the p-scores (Figure 3.8). Those that are more ambiguous tend to have no family with a particularly high p-score and thus exhibit greater entropy. By using entropy as a measure of confidence, GRAPPLE predicts whether a novel sequence belongs to a family that the SVM was trained upon, or fits a none-of-the-above category better.

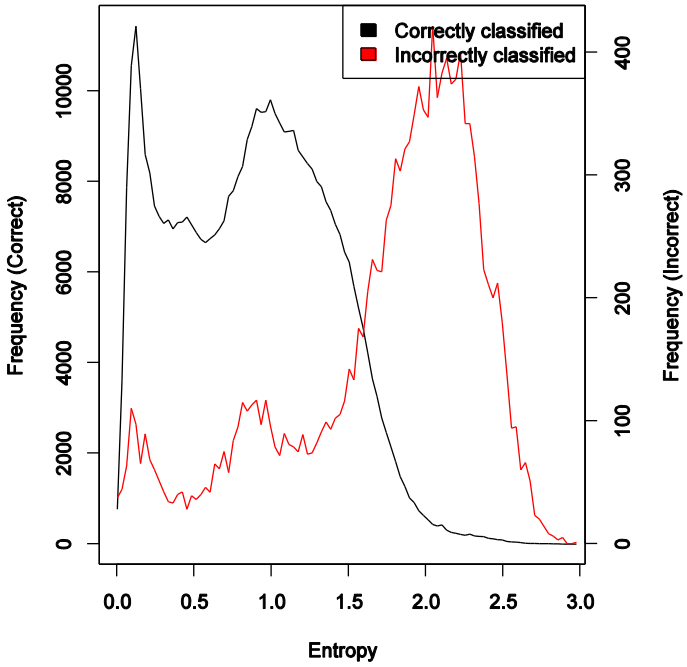

Figure 3.8: Entropy histogram for correctly and incorrectly classified sequences.

Applying Shannon entropy (Equation 6) to the merged p-scores of a sequence was used to estimate the confidence associated with the prediction. The lower the entropy for a sequence, the more certain the method is that the sequence belongs to the family with the highest p-score. The bimodal nature of the curve is explained by the combination of the BLAST and GraPPLE. The BLAST method tends to produce very high E-values for a single family and very low for the rest while the SVM method provides a more balanced distribution of SVM p-scores.

3.5. Discussion

Currently, there exist very few predictors capable of assigning function to uncharacterised ncRNA molecules and even fewer that can predict whether or not an ncRNA molecule is functional. Available methods are based on sequence comparison (Altschul et al., 1990),

covariance models (Yao et al., 2006), graph topology (Karklin et al., 2005) and structural alignments (Washietl et al., 2005). However, the disadvantages of existing methods limit their application. Here, we presented a novel method for de-novo ncRNA and Rfam family prediction that addresses some of the problems of the existing methods, expands the repertoire of available methods and, hopefully, helps enrich the understanding of the RNA world.

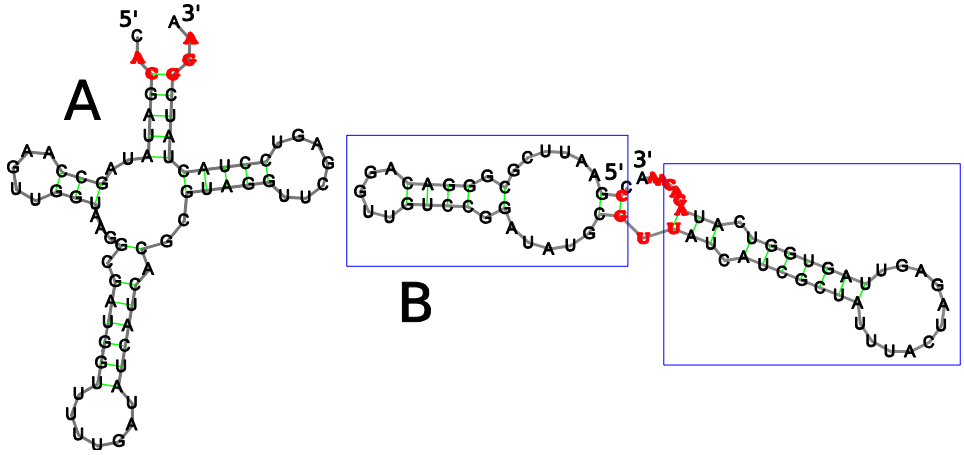

Figure 3.9: Structural relevance of articulation points.

Articulation points are nodes that, if removed, will disconnect the graph. In a structural sense, these are either "dangling" nucleotides at the 5' and 3' ends of a molecule or "bridge" nucleotides that connect potentially separable structures. A tRNA molecule (A) and a shuffled counterpart (B) are shown. The red nucleotides are articulation points in the graph representation and the structures within each blue box are potentially separable secondary structures. Ester bonds are shown in black and hydrogen bonds are coloured in green. If graphs with fewer articulation points tend to represent functional ncRNA, then functional RNA molecules are more likely to have fewer dangling/bridge nodes and potentially separable structures than non-function RNA molecules. The tRNA graph contains 5 articulation points and not separable structures whereas the shuffled counterpart contains 10 articulation points and two separable structures.

By analysing the manner in which the graph properties are calculated, we may gain insight into how functional RNA is formed and what topological features render functional RNA molecules unique compared to their non-functional counter parts. Analysing the “number of articulation points” may serve as an example as it is relatively easy to interpret. In graphs representing RNA secondary structure, there are only two situations where removing a node in the graph can disconnect the graph: when a node is removed from the dangling 5' or 3' end, or when a node is removed from a bridge connecting potentially separable structures (Figure 3.9). From this study, we find that graphs with more articulation points are more likely to represent non-functional structures, indicating that functional structures minimise the length of the dangling ends and, in addition, either minimise the length of the bridges between

separable structures or the number of separable structures. The remaining graph properties that were determined as important are calculated using more complex algorithms and a structural interpretation is more complex (Gross and Yellen, 2004). Further biological interpretations of some of the other graph properties can be found in the Appendix Section II.

When using graph properties, it is important to remove factors that may obscure the attempt to determine whether they reflect sufficient information about Rfam family to make a prediction; otherwise, simpler methods such as sequence alignment or predictions based on just the sequence length would suffice. After removing these confounding factors, the graph properties themselves are shown to maintain the predictive power. As a result, we gain insight into structurally important properties for functional RNA by interpreting the way graph properties are calculated in a biological context (demonstrated here with the number of articulation points). The graph properties also provide sufficient information for Rfam family prediction on highly divergent sequences. Combined with the aforementioned sequence properties, the accuracy of the method improves significantly.

ncRNA exhibits greater conservation on the secondary structure level than the primary structure (Yao et al., 2006), as is demonstrated by the large variety of tRNA molecules, and thus sequence similarity is a potentially suboptimal choice of a classification criterion. In many cases, the sequences are sufficiently dissimilar at the sequence- and even the secondary-structure (inferred from covariation) level that neither sequence alignment nor covariance models are able to recall the function, and yet a classifier built on the graph properties; *i.e.*, based on a higher-level of abstraction, still manages to perform accurately. This finding indicates that, to a certain degree, our method is sequence independent and that there are properties inherent in the structure of ncRNA indicative of function.

Of the many graph representations available, such as dual graph and planar tree representation, we chose bracketed graph representation. This representation was chosen over the others as it is more sensitive to small changes in the underlying RNA structure due to the greater number of nodes and edges. The graph space of the other representations is far smaller, potentially reducing the predictive power of the derived graph properties. This representation also minimises information loss (*e.g.*, dangling ends and, in dual graph representation, the length of the stems and loops) and is simpler to calculate.

On the other hand, there are several potential disadvantages of our proposed method: (1) usage of sequences biased toward certain species, (2) dependence on one folding algorithm of

choice, and (3) usage of secondary structure prediction, thus neglecting pseudoknots and tertiary structures. Note that all of the identified issues are exogenous to our method, and one can account for them by conducting comparison tests—a task beyond the scope of this paper.

The sequences available in Rfam are largely bacterial and viral, and thus the method we have developed will be biased towards the prediction of purely bacterial and viral Rfam families or, where the family occurs in several kingdoms, higher accuracy prediction in bacterial and viral sequences. When more ncRNA sequences become available, our method will benefit from being trained upon a more specific choice of sequences, *e.g.*, purely plant or animal sequences.

Often the predicted structure of an ncRNA sequence is quite different from the experimentally determined structure. As we obtained all secondary structure assignments using RNAFOLD (Hofacker, 2003), our method is reliant on RNAFOLD producing consistent predictions. As the tools for RNA folding prediction improve, we plan to upgrade the folding algorithms, hopefully yielding higher accuracies and better understanding of the biological and structural relevance of graph properties. An immediate possible improvement is the use of multiple alignments which improves the accuracy of the predicted secondary structure (Hofacker et al., 2002).

We chose to calculate graph properties from secondary structure, rather than pseudoknots or predicted 3D RNA structure, which potentially limits the predictive power of the graph properties. By using secondary structure, the maximum degree (number of connected edges) for any node is limited to three *i.e.*, two ester bonds and a hydrogen bond. With pseudoknots and 3D structure, the possibility for much more complex graphs emerges, which would have far reaching consequences on the graph properties. Such graphs are likely to better reflect functional information, which is a point for further study.

We expect further improvement through careful selection of the graph properties as potential discriminatory features. Of the current graph properties, we would ideally use only those that are most informative and perhaps more biologically relevant. There are also many more properties that can be calculated than the 20 chosen and experimentation with new graph properties may lead to improved accuracy and greater insight into ncRNA functionality. Including properties such as minimal folding energy (which has been shown to be informative for miRNA), %G+C and perhaps dinucleotide frequencies in SVM training should provide a significant boost to the accuracy of the described method.

Many of the existing methods have shortcomings that limit their applications. As many ncRNA families show little sequence homology, homology-based methods are unable to completely cover each family. Covariance models, although highly accurate, require long computation times and neither method are able to discriminate between functional and non-functional ncRNA sequences. The method developed in the current paper, can cover more sequences than homology-based methods at quick speeds typical of SVM-based methods, which provides a good compromise between the two methods. It also exhibits the ability to identify functional RNA sequences and may also help to determine the properties that render an RNA molecule functional. Finally, combining graph properties with other methods provides a significant boost to performance.

In conclusion, ncRNA is not simply a primitive form of molecule as it is active in a wide variety of roles not typical for proteins. We developed a computational method that represents a necessary first step for future ncRNA investigation tools. With a plethora of potential functions still undiscovered and many more molecules whose functional role is still unassigned, we believe that higher-level structural abstraction and their respective properties will play a key role in discovering new ncRNAs and their plausible biological role.

Chapter 4.

General discussion

In this thesis, experimental and bioinformatics methods were presented as means for the investigation of ncRNA. The experimental method produced a high resolution genomic map for 54 *A. thaliana* accessions from single feature polymorphism (SFP) predictions that were then used to identify a large set of candidate genomic regions that significantly associated with nine measured phenotypic traits. The distribution of SFPs in the *A. thaliana* genome and the significantly associated coding regions were annotated with publicly available data. Four-fifths of the significantly associated regions were annotated to be non-coding, suggesting a strong influence of non-coding regions upon trait variability.

A bioinformatics tool developed to predict functional ncRNA functional and assign Rfam families was presented in Chapter 3. In this method, the secondary structures of ncRNA, which were obtained from Rfam, were used to construct graphs from which the graph properties were calculated. The graph properties were used to train two types of support vector machine (SVM). One type was able to predict if a query RNA sequence was functional and the second type was able to assign an Rfam family to the sequence, if it was determined to be functional. Both predictors proved to be highly accurate; the functional predictor had a Matthews Correlation Coefficient (MCC) of 0.87 and the Rfam predictor had a MCC of 0.98 when trained on sets of 200 sequences. In this chapter, the results of the previous two chapters are presented in a broader context, the requirements and enabling technology for the developed software are discussed and future considerations are covered.

4.1. Broader context

4.1.1. Paradigm shift

Proteins were, and still are, the prime targets for research into the underlying molecular causes for the wide variety of phenotypes and traits observable in an organism, which is probably a natural result of being a highly visible functional component in biological processes. They are used as the primary agonist in almost every biological system imaginable and their functions cover catalytic, structural and regulatory roles. Perhaps it is their structural and functional versatility and stability that makes them a prime candidate for these roles. However, it would appear that sometimes proteins are not the ideal molecule for a particular set of regulatory roles and ncRNA genes provide an alternative, more optimal solution.

With the discovery of new ncRNA roles, research efforts have been focussed on exploring the number and types of ncRNA families mostly through homology based techniques. They have revealed a significant number of new ncRNA families and the manner in which they perform their functions, although it is possible that more families exist that remain undetectable through the implemented algorithms as suggested by the findings presented in Chapter 2. Furthermore, the involvement of ncRNA in metabolic pathways and their regulatory effects are also a major focus, especially for the ncRNA families of riboswitches and miRNA. By understanding both the protein and ncRNA sides of biological networks, we stand a better chance of fully elucidating the inner workings of all organisms.

4.1.2. High-density, genome-wide association mapping

The data for genome wide association mapping has been available for a while, usually in the form of SNP markers, however the resolution has been often quite poor. Only recently has high resolution data capable of finely mapping the genetic causes of phenotypic trait variation onto the genome become available with the advent of several large scale genotyping efforts (Borevitz, 2006; Clark et al., 2007; Nordborg et al., 2005). The work presented in Chapter 2, although preliminary, represents a first step towards finding the exact combination of underlying genetic elements responsible for phenotypic variation on a genome-wide scale.

After adjusting for problems introduced by population structure and multiple testing, several probable genomic causes for phenotypic variation were identified, the large majority of which lay in non-coding regions. This suggests a large influence of non-coding genomic regions on trait variability, possibly raising many more questions than the case where mainly

coding regions were identified. In further investigations we could ask: What exactly are these regions? Are they transcribed or could the DNA itself have some influence? If they are transcribed are they miss-annotated as non-coding? If they are transcribed and non-coding, do they belong to an entirely new class of ncRNA? Do they act through antisense complementarity, bind molecules or even have catalytic functions? Mutations in ncRNA also raise the possibility of very subtle changes. A mutation in a protein tends to produce highly visible changes in conformation and amino acid composition; however, a mutation in ncRNA could slightly disrupt a stem structure, cause a stem/loop to extend a little bit farther, or indeed disrupt the entire structure completely. Such small changes could change a target, slightly alter binding specificity and strength, or change translational efficiency among a host of other possible effects both large and small. Should further research show that non-coding regions truly do have a large influence on trait variability, than these are all issues that will need to be considered.

4.1.3. ncRNA functions

Classifying an RNA molecule as functional or non-functional remains a difficult task and identifying the properties that make an ncRNA functional should be considered a priority. Tools that tackle this problem initially seek conserved sequences across several genomes using a variety of methods, include a second step for further confirmation such as a significantly lower minimum free energy (MFE) or a shared secondary structure (Mourier et al., 2008; Pedersen et al., 2006; Song et al., 2009), and have predicted varying numbers of previously unannotated ncRNA. However, such approaches are likely to miss those ncRNA families that show little sequence conservation among the members. Experimental methods uncover an abundance of potential ncRNA molecules (Lu et al., 2005) and the conflict in reported numbers of potential ncRNA raises the question of exactly how many ncRNAs are expected to be found. Through the use of GRAPPLE the properties that functional ncRNA share may be unveiled and some of the answers to such questions could be found.

Even after identifying functional ncRNA and predicting a function, the exact role of the molecule in a biological system also needs to be elucidated. There are a handful of high-throughput reverse genetics techniques that provide the necessary bandwidth required to provide experimental confirmation for the large number of predictions being made by the bioinformatics techniques (Bargmann, 2001; Clark and Ding, 2006; Sessions et al., 2002; Stuitje et al., 2003).

4.2. Future Considerations

4.2.1. Grappling with SFPs

The methods developed and the results from the previous two chapters are able to be combined with the aim to identify potential ncRNA for experimental confirmation. Several lines of research indicate that there are several unannotated ncRNA genes in *Arabidopsis thaliana*. Multiple alignments of the *A. thaliana*, poplar (*Populus trichocarpa*), grape (*Vitis vinifera*), papaya (*Carica papaya*) and rice (*Oryza sativa*) genomes (Song et al., 2009), although limited to using mostly draft genomes, identify 21 ncRNA in 16 novel ncRNA families. Massively parallel signature sequencing (MPSS) of *A. thaliana* flowers and seedlings has discovered over 75,000 unique potential small RNAs (Lu et al., 2005). Finally, from Chapter 2, the 886 out of the 1,082 markers that significantly associated with variability in nine phenotypic traits lay in non-coding regions indicating a significant influence of non-coding regions on trait variability. Thus, exploring *A. thaliana* for more ncRNA families is not only likely to produce more ncRNA in both existing and new families but, more importantly, there are indications that ncRNA have a heavily involved in influencing biological and molecular phenotypes. If this is so, then only by fully characterising and describing the ncRNA that constitute an organism are the biological processes that take place within cells able to be fully understood and defined. Many more ncRNA families are likely to be discovered especially if they retain little sequence similarity and thus evade current techniques for ncRNA identification. GRAPPLE is an ideal tool to be used for the identification of ncRNA in these regions as it is able to both identify regions that are possibly functional and provide clues about the family to which the potentially functional ncRNA belong. A bioinformatics integration of the two chapters is presented here and for further progress, experimental confirmation of the results is required.

From Chapter 2, all SFP sites with significant associations to the variability of one of the nine traits and which fell in non-coding regions were further analysed. A 100 nucleotide region around each SFP site (50 upstream, 50 downstream) was extracted from the *A. thaliana* genome and submitted to the GRAPPLE ncRNA identification and Rfam family prediction tool. GRAPPLE produces two results for each sequence. The first indicates the probability that the sequence has some sort of function, and the second is an entropy score indicating the confidence of the assigned Rfam family. The probability score used by Support Vector Machines (SVMs) is based on the distance of the submitted sequence's graph properties from the support vector and ranges from 0 to 1 where higher scores indicate a greater distance from

the support vector and thus a better probability of belonging to a class. Normally a threshold of 0.5 is set for functional classification, but for this analysis we use set a threshold to 0.7 as a balance between high confidence results and the number of significant functional sequences found. From the sequences predicted to be functional, we consider Rfam families assigned with entropy of 1.5 or less to be assigned with high confidence. Any higher entropy is considered low confidence.

Of the 886 significant marker-trait associations that occur in intergenic regions and introns, 55 are considered to be within functional ncRNA with a p-score of 0.7 or greater (Appendix Section III). Two potential ncRNAs have Rfam families assigned with entropy indicating high confidence; a T-box ncRNA associated with sucrose and a signal recognition particle (SRP) also associated with sucrose. The T-box ncRNA belongs in promoters and accordingly, the one identified is found 532 nucleotides upstream of an ERF/A2 transcription factor, still potentially within the promoter region. The region around the SRP ncRNA is annotated as a transposable element, throwing doubt on the accuracy of the prediction. The remaining ncRNA have possibly novel functions.

Currently, the results presented so far are tenuous at best and await experimental confirmation. The correlation between the phenotypic traits and the associated gene functions shown in Chapter 2 (*e.g.*, amino acid and amino acid synthesis/catabolism genes) provides support that the significantly associated regions are likely to have some sort of biologically relevant connection to the trait. If the associated region is truly intergenic or an intron, then we must look further than protein coding genes for a causal agent.

Although mostly found in bacteria, T-box ncRNA are not entirely new to *A. thaliana* and have been previously described (Chan et al., 2001). The potential T-box:sucrose association revealed by the combined methods hints at a previously undiscovered T-box. Furthermore, the T-box lies 532 nucleotides upstream of a transcription factor whose expression is significantly differentially expressed under sucrose starvation (Contento et al., 2004). Any investigation seeking experimental confirmation of the results presented in this chapter should start with this association. Such investigations could involve checking whether the proposed region is transcribed and if it really is a promoter for the gene. SALK lines with T-DNA insertions in the area are available and could be used to quickly identify the behavior of the disrupted ncRNA under different sucrose levels in the growth medium.

To further investigate the remaining ncRNA predictions, further bioinformatics and

experimental analyses could be performed. Bioinformatics analyses tend to be limited to comparing the sequence or structure of the candidate ncRNA with known families. However, as these predictions are low confidence, they could either belong to other ncRNA families than indicated or could be altogether novel families of ncRNA. The latter case limits the effectiveness of bioinformatics tools thus necessitating an experimental approach. As a first step in experimental confirmation, tiling arrays or RT-PCR would be fairly simple and effective methods for checking whether the predicted region is expressed. Later, reverse genetics approaches, such as knock-outs or insertions, could define the function of the predicted region more narrowly and the interaction it has in gene networks.

Sequence similarity is the most commonly used method for both (1) identifying genes that are homologous to a query gene and (2) identifying unannotated regions of homology across several genomes. This approach is perhaps the most intuitive, as the conservation of function is most easily achieved through the conservation of sequence. Thus the current methods used to search *A. thaliana* for novel ncRNA families rely upon sequence comparison to discover similar regions and thus potentially conserved functional genes (Song et al., 2009). While it seems unlikely that any ncRNA family would have little sequence-based homology among the members, we only need to look as far as tRNA to find a family that has a conserved function, poorly conserved sequence homology and highly conserved structural homology. It is also worth considering that any ancient ncRNA genes would have sufficient time for homologous sequences to diverge to the point where they no longer resemble each other. Therefore, it may be insufficient to rely completely on tools that seek sequence similarity and, instead, apply tools and methods that are able to identify novel functional ncRNA based on other criteria such as structural homology and implication through association. There are very few tools that are able to do this. The programs ALIDOT and PFRALI are examples of tools developed to rectify this problem, but still require an initial multiple sequence alignment and are unable to process large amounts of data (Hofacker et al., 1998; Hofacker and Stadler, 1999), thus the need to develop a tool that is able to able to predict whether or not a genomic region of interest may be functional.

Although GRAPPLE was designed to process complete ncRNA sequences, preliminary tests in developing a genome scanner indicate that it is also able to perform considerably well when provided the fragmentary ncRNA that occurs when windowing genomic DNA. Two cases are possible: (1) An ncRNA fragment, (2) The whole ncRNA and surrounding genomic sequence. This is possibly due, in part, to the inclusion of minimum free energy (MFE) and

the support provided by other graph properties. The performance of GRAPPLE is still likely to be impaired and thus we chose more stringent thresholds when determining whether the submitted sequence is functional.

In this section, an example of how two techniques can work in tandem to further investigate the RNA phenomenon is presented. The association mapping produced both coding and non-coding SFP sites that significantly associated with the variability of different biological traits. There were many more significantly associated non-coding SFP sites than coding and thus, if only the coding SFPs were characterised only one third of all significantly associated regions would be addressed. By applying GRAPPLE to the analysis of the non-coding SFPs, we sought to provide insight into the families of potential ncRNA and the role they play in the regulation of the associated trait. One likely ncRNA candidate is proposed for sucrose regulation along with 53 candidates for novel ncRNA families. With further experimentation, the nature of these non-coding SFP sites will be revealed.

4.2.2. A genome scanner

An obvious first step in further developing the detection of ncRNA family and function using graph properties is the implementation of a genome scanner. Such a scanner should be able to receive long (>1000 nucleotides) sequences as input and report the possible locations and lengths of ncRNA that lie upon it.

There are a few considerations that need to be addressed when adjusting the method described in Chapter 3 from a static window of variable length to a sliding window of fixed length. Four scenarios can be envisaged when scanning a genome for ncRNA.

1. There is no ncRNA in the window.
2. The window is the same size as the ncRNA.
3. The window is smaller than the ncRNA.
4. The window is larger than the ncNA.

For case 1, the result is straightforward. The predictor is expected to report no functional ncRNA. For case 2, little changes from the predictor developed in Chapter 3. However, for the remaining two cases the predictor is expected to report a functional ncRNA and two obstacles must first be overcome before the method is applied.

The first obstacle, when the window is smaller than the ncRNA, is to first ascertain whether the fragment of ncRNA that is inside the window can be detected. The important graph

properties that were previously determined are likely to be very different from those that are important in this case. Because the ncRNA doesn't entirely fit inside the window, "dangling ends" are likely to be generated, which leads to an increase in the number of articulation points. From Chapter 3 it is known that the more articulation points, the more likely a functional ncRNA will remain undetected. When the window is larger than the ncRNA, a similar problem arises. Thus we need to re-train the SVMs to handle these particular scenarios and to re-determine which graph properties provide the most discriminatory power.

The size of the window itself may also be important. Too small and we may not pick up on long range interactions that provide vital structural clues. Too large and the signal provided by the ncRNA will get lost in the surrounding noise. As a step towards solving this problem, we need to try a variety of different window sizes. This approach may also provide clues about the size of the ncRNA predicted as the ncRNA is expected to behave best when folded in a window of roughly the right size.

4.2.3. RNA structure in translation initiation

As shown in the case of riboswitches, RNA can regulate gene expression through structures encoded in mRNA. This ability may provide an alternative to the most well supported hypothesis for translation initiation in prokaryote genomes. The current hypothesis is that the Shine-Dalgarno (S-D) sequence (consensus AGGAGG), is required for the translation initiation of a gene (Shine and Dalgarno, 1975). This sequence is often found in the 5' untranslated region (UTR) ranging from 22 to 2 nucleotides upstream of the start codon. The S-D sequence is suspected to bind to the anti-S-D sequence, which is found on the tail of the 16S small ribosomal subunit; an ncRNA gene. More importantly, a significant number of genes do not have an S-D sequence (54% in fully sequenced plastids) implying that alternative translation initiation mechanisms may exist. Established methods, such as searches for alternative conserved sequence motifs, fail to turn up any promising results and thus the search for alternative signals, which could reveal a more global translation initiation mechanism, may prove helpful.

Such an alternative signal could lie in the folded structure of the gene's 5'UTR as proposed by the cumulative specificity mechanism (Nakamoto, 2009). The cumulative specificity mechanism hypothesises that secondary structure selectively interdicts access to most of the non-initiator methionine codons while leaving open the true initiation site and that the final recognition of the initiation site occurs by the cooperativity and cumulative specificity of the

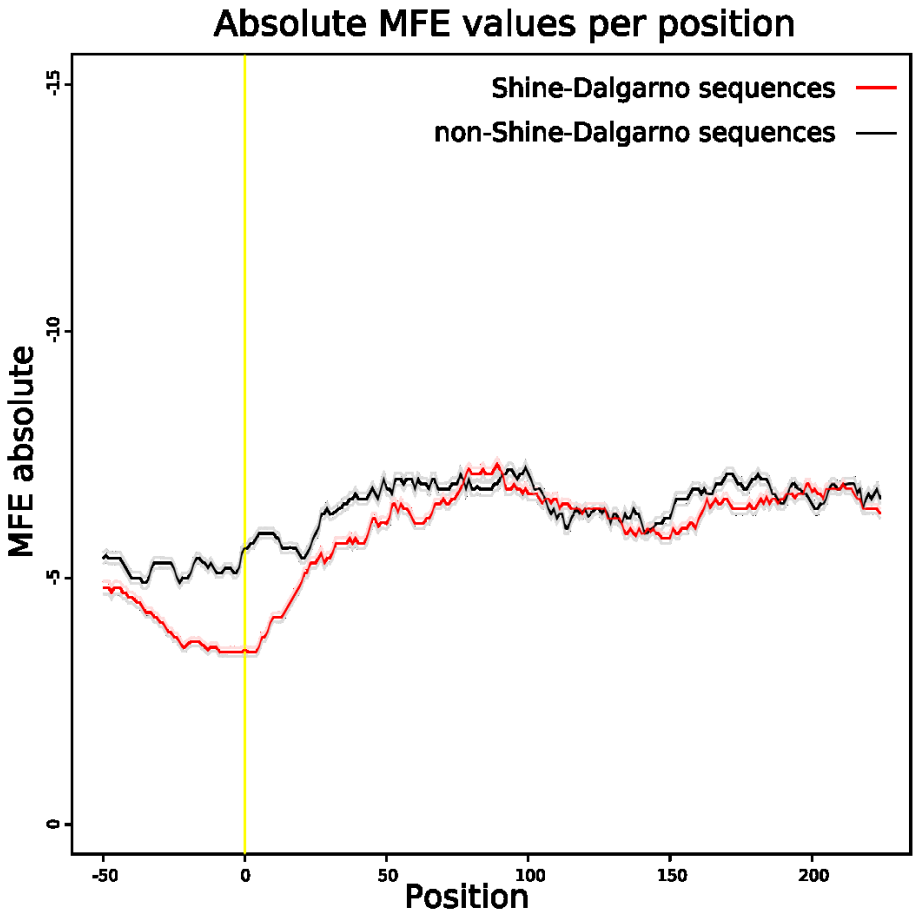

Figure 4.1: Average RNA structure around plastid start codon

A sliding window of 50 nucleotides has been applied to the region 75 nucleotides upstream to 250 nucleotides downstream of the start codon for all protein coding genes shared among at least 60% of fully sequenced and annotated plastids. The windows are folded using RNAFOLD *to obtain the running minimum free energy. The y-axis has been flipped from (-15, 0) to (0, -15) to more intuitively represent the amount of structure at each point* i.e., *A trough on the graph represents an area with low structure and* vice versa.

several ligands recognition site of the ribosomes, which confer broad substrate specificity to the system. Previous research indicates that RNA structure around the ribosomal binding site on mRNA influences the translation efficiency of *E. coli* (Kudla et al., 2009). By folding the area around each methionine codon in mRNA, and comparing all non-initiator codons with initiator codons, we tested this hypothesis. In plastids we observed that the start codon, on average, has much less structure than the coding region of the mRNA (Figure 4.1) and we observe decreased structure around the start codon also in *Escherichia coli, Arabidopsis thaliana* and metazoan mitochondria. There are hints that human genes also share this property.

The relationship between the genes that have the S-D sequence and those without has also been investigated. S-D positive sequences were defined as genes whose upstream region,

Algorithm 4.1: Simulated annealing algorithm

Given: Initial heat: h_0, Heat decay: dh, Stable state: h_n, Sequence sequence

1: Initialise h_i to h_0
2: Calculate a fitness score for the sequence as old_score
3: **while** $h_i > h_n$ **do**
4: Mutate a random position in the sequence using probabilities generated from codon usage table.
5: Calculate a fitness score for the new sequence as new_score
6: **if** new_score < old_score **then**
7: Accept new sequence
8: **else**
9: Accept new sequence using Boltzmann probability calculated from new_score − old_score and h_i

from -22 to -2, hybridised with the 16S rRNA tail (containing the anti S-D sequence) with a minimum free energy of -4.4kJ. These criteria were chosen according to earlier methods developed for detecting the S-D sequence (Starmer et al., 2006). This reveals that there is a large difference in energy signatures between the genes with S-D sequences in the promoter region and those without. On average sequences with an S-D sequence show a drop of roughly 3 kcal mol^{-1}, whereas those without show a drop of about 1 kcal mol^{-1}. Semi-random sequences, which have been designed to resemble plastid mRNA, show a significantly smaller difference, between those that have an S-D sequence detected in the 5'UTR and those that do not, across a significantly shorter range than non-random sequences. This is highly suggestive that the S-D sequence itself is not responsible for the drop in structure observed and that selective pressure maintains lower structure in S-D positive mRNA than S-D negative mRNA around the start codon.

To test whether the amount of structure around the start codon will have an effect on translation initiation, several reporter gene constructs can be generated that show both high and low structure around the start codon. Once inserted into an expression platform such as *E. coli*, the level of expression can be measured and statistical test will show whether constructs with lower structure around the start codon have higher expression or not.

To generate the constructs, a bioinformatics tool is under development to mutate a given sequence. The mutations will preserve the coding sequence and other regions that the user

specifies and simultaneously attempt to optimise the amount of structure to an amount also provided by the user. To avoid rare codons, a species specific codon usage table can be provided. This table will be used to calculate the probabilities at which a random codon will be chosen. The sequence will undergo a similar process to simulated annealing (Algorithm 4.1). This process allows large changes to the sequence in the initial stages and slowly "cools" the system until the sequence does not undergo any more mutations.

4.3. Software development

4.3.1. Software accessibility and user friendliness

When developing new bioinformatics tools, it is important that the finished product is easily accessible and user friendly. To make a tool such as GRAPPLE accessible, there are a few methods that are possible and it is highly desirable that two in particular are implemented.

The first desirable method is to design a web page that provides visually guided access through an internet browser (such as Firefox or Microsoft Internet Explorer) to the underlying tool. Such a method makes the tool accessible to a wide audience and is particularly advantageous for those with little programming skill. The presentation of a web page is particularly important. The interface must be intuitive and the results easily understandable. An intuitive interface can be achieved through the tolerance of many input styles, hiding more advanced options along with an uncluttered and minimal presentation. GRAPPLE was designed with these principles in mind. The only input that GRAPPLE requires is a single sequence either as direct input into the page, or uploaded as a file. This sequence can be either a simple string of nucleotides or a proper FASTA file. The structure of the sequence can also be provided allowing the user to submit a custom fold. The home page allows immediate access to the tool and other information such as the description of the algorithm lies in other easily accessible pages. The results have also been presented to be easily interpretable. The most relevant results are explained in full English sentences with important information emphasised in bold text. More detailed results are also available and are presented further down the results page.

The second desirable method is to make the tool available as a "web service". A web service allows programmers programmatic access to the tool over the internet. Protocols such as the Simple Object Access Protocol (SOAP) allow programmers to tightly integrate the remote tool with the local tools and scripts that they develop.

A third, more complex, method to make a tool accessible is to make it downloadable. This method comes with a number of advantages and disadvantages. Cross-platform compatibility is always an issue with downloadable tools. Developing the tool in cross-platform languages can circumvent this problem, however, if the tools is written in a non-cross-platform tool such as C++, then compilation will be needed, which requires programming knowledge on the user's side. If the tool is resource intensive, then allowing the user to download alleviates the load on the provider's computers. However, this also requires that the user has the computing power to run the tool. The programmatic libraries used by the tool also need to be installed causing further complications for the user. Although this method is ultimately the most powerful option, the uncertainties associated with it make it less desirable to implement for small to medium sized tools. Ultimately, one must consider the advantages and disadvantages of all the possibilities to arrive at the best decision.

4.3.2. High performance computing

The bioinformatics investigations presented in this thesis would not have been possible without the availability of high performance computing. Use of a computer cluster enabled a large number of independent analyses to be conducted in parallel, significantly reducing the time required for completion. The large amount of data generated by the tiling arrays, the many repetitions required by the SVMs for statistical significance and especially the >1 billion association mapping comparisons performed emphasises the importance of high performance computing in modern and future bioinformatics.

4.4. Conclusion

The search for alternative methods for the investigation of biological phenomena can provide new perspectives and discoveries. By using the relatively novel approach of applying genomic DNA, instead of the usual cDNA, to tiling arrays, we were able to create a genomic map of 54 *A. thaliana* accessions and by associating the predicted markers with a variety of phenotypic traits, we discovered that roughly four-fifths of the genome, believed to influence the traits, is non-coding. To analyse the non-coding regions, we needed a tool that was able to determine the nature of the markers contained within. The available tools relied on sequence and structural homology which limited classification to classes that already exist and did not address to question of whether the proposed ncRNA sequences were functional or not. By using graph properties derived from graphs representing RNA structure, we developed an ncRNA predictor that fulfilled these objectives while potentially revealing clues about which

structural properties define functional ncRNA. We applied the predictor to the non-coding markers to begin investigations into the involvement of ncRNA in the regulation of phenotypic traits. Combined with high performance computing, we were able to process large amounts of experimental data and perform the many permutations required for significance calculations in statistical methods.

The research in this thesis emphasises the shift from studying proteins as the primary biological molecule to include ncRNA, which appear to be more suited to executing certain roles than proteins are. These molecules have the ability to perform many regulatory and catalytic roles. The discovery that large portions of the genome that are non-coding and with no known function are transcribed and that the majority of genetic markers that associate with phenotypic trait variability are also in non-coding regions, seems to suggest that a significant portion of the metabolic and regulatory networks remains missing when only proteins are considered. Through the study of ncRNA, a journey is made into the past as we reveal the nature of the ancient RNA mechanisms that comprised early life on Earth while simultaneously exploring the frontiers of the modern RNA world.

I. Significant GO term over-representation

GO terms were tested with a Fisher Exact test for over-representation and p-values <= 0.01 were labelled as significant.

Amino acid (181 SFP sites in 31 genes)

GO term	p-value
generative cell mitosis	1.38E-03
tRNA wobble uridine modification	1.38E-03
isopentenyl-diphosphate delta-isomerase activity	2.76E-03
protein localization	2.76E-03
actin binding	2.88E-03
oxidoreductase activity	8.25E-03
intracellular	8.41E-03
protein amino acid phosphorylation	8.64E-03
protein serine/threonine kinase activity	9.13E-03
cytidine metabolic process	9.62E-03
lipoxygenase activity	9.62E-03

Beta-alanine (141 SFP sites in 20 genes)

proton-transporting ATP synthase complex assembly	9.13E-04
response to nematode	1.19E-03
proline catabolic process	1.83E-03
proline dehydrogenase activity	1.83E-03
succinate dehydrogenase complex	2.74E-03
glutamate biosynthetic process	3.65E-03
citrate (SI)-synthase activity	4.56E-03
Golgi organization	5.47E-03
cellular biosynthetic process	5.47E-03
peroxisome	6.24E-03
gibberellin 2-beta-dioxygenase activity	6.38E-03
mitochondrial electron transport, succinate to ubiquinone	6.38E-03
succinate dehydrogenase activity	6.38E-03
eukaryotic translation initiation factor 2B complex	7.28E-03

Erythritol (17 SFP sites in 2 genes)

Insufficient genes for over-representation analysis

Fresh weight (177 SFP sites in 38 genes)

GO term	p-value
DNA topoisomerase complex (ATP-hydrolyzing)	4.88E-03
regulation of embryonic development	4.88E-03
ribosome binding	4.88E-03
tRNA-intron endonuclease complex	4.88E-03
cell communication	6.50E-03
glycine-tRNA ligase activity	6.50E-03
glycyl-tRNA aminoacylation	6.50E-03
tRNA-intron endonuclease activity	6.50E-03

Myo-inositol (146 SFP sites in 26 genes)

response to osmotic stress	2.33E-03
glucosylceramidase activity	4.34E-03
glucosylceramide catabolic process	4.34E-03
ATP binding	4.97E-03
DNA helicase activity	7.57E-03
sphingolipid metabolic process	7.57E-03
fruit dehiscence	8.65E-03

Protein (83 SFP sites in 8 genes)

Insufficient genes for over-representation analysis

Starch (8 SFP sites in 2 genes)

Insufficient genes for over-representation analysis

Sucrose (161 SFP sites in 20 genes)

specification of organ position	1.02E-03
cell fate commitment	4.08E-03
UDP-glycosyltransferase activity	4.81E-03
SNARE binding	5.10E-03
abaxial cell fate specification	7.13E-03
anion exchanger activity	7.13E-03

Threonic acid (168 SFP sites in 16 genes)

oxidoreductase activity	2.50E-03

II. Graph property definitions

Biological interpretations are not available for all graph properties. For those with interpretations, the interpretation is very much open to debate and will require a more in-depth analysis at a later date.

Shortest path	The shortest path between two nodes.
Path length	The length of a path. In the current paper, average path length refers to the average length of the shortest paths.
Articulation point	A node that, if removed, will disconnect the graph. **Biological interpretation:** A nucleotide in a dangling end or bridge between two separable secondary structures.
Node betweenness	The number of shortest paths that pass through a node. **Biological interpretation:** This property measures to a small extent the number of base-pairs that exist in the structure and how compact the structure is.
Edge betweenness	The number of shortest paths that pass through an edge.
Cocitation coupling	Two vertices are co-cited if there is another vertex citing both of them.
Bibliographic coupling	The bibliographic coupling of two vertices is the number of other vertices they both cite.
Closeness centrality	The mean shortest path between a vertex and all other vertices reachable from it.
Burt's constraint	A measure of the extent to which a node (v) is invested in people who are invested in other of v's alters (neighbours). The "constraint" is characterised by a lack of primary holes around each neighbour.
Degree	The number of edges connected to a node.
Diameter	The length of the longest shortest path.
Girth	The length of the smallest cycle.
K-core	A sub-graph where every node has k connections.
Coreness	The coreness of a vertex is k if it belongs to the k-core but not to the (k+1)-core **Biological interpretation:** Every node in graphs that represent ncRNA secondary structure has a coreness of two except for those in the dangling ends which have a coreness of one. Thus, the coreness is a measure of the total length of the dangling ends ranging between 1 and 2, where an average coreness of 2 means no dangling ends.
Graph density	The density of a graph is the ratio of the number of edges and the number of possible edges.
Transitivity	Transitivity is defined as the ratio of 3 * the number of triangles and the total number of length-two paths (*i.e.*, paths on 3 nodes) in a graph.

III. ncRNA annotation of intergenic SFPs

A 100 nucleotide region around intergenic SFP sites with significant associations with one of the tested traits were submitted to GRAPPLE for analysis. Sequences were labelled functional with a p-score greater or equal to 0.7. The Rfam assigned to the sequence was considered high confidence with a entropy score less than or equal to 1.5.

Amino acid (145 intergenic SFP sites)

Chr	Position	p-score	Assigned Rfam family	entropy
3	605907	0.79	Signal recognition particle	2.11
3	16877052	0.75	Y RNA	2.16
5	8426125	0.75	Y RNA	1.86
2	13373358	0.72	T-box	2.87
3	605609	0.72	miRNA	1.72
1	25328207	0.71	Y RNA	2.52
2	5639522	0.71	Y RNA	2.19
3	16877502	0.7	Signal recognition particle	2.69

Beta-alanine (123 intergenic SFP sites)

2	2225203	0.78	Y RNA	2.01
5	16299974	0.77	Y RNA	2.27
1	21057970	0.74	HIV primer binding site	2.91
4	8677586	0.74	U5 spliceosomal RNA	2.95
1	21070606	0.73	Y RNA	2.40
5	16300015	0.72	Y RNA	2.34
5	16300273	0.71	U4 spliceosomal RNA	2.90

Erythritol (12 intergenic SFP sites)

4	13408215	0.77	tmRNA	2.76

Myo-inositol (115 intergenic SFP sites)

1	426650	0.82	Riboswitch	2.28
1	432673	0.75	U2 spliceosomal RNA	2.23
5	10262704	0.73	CD-BOX	2.69
1	432824	0.71	Y RNA	2.14
4	463906	0.70	U5 spliceosomal RNA	3.11

Fresh weight (142 intergenic SFP sites)

Chr	Position	p-score	Assigned Rfam family	entropy
1	23465348	0.77	Y RNA	2.62
2	5414902	0.74	Y RNA	2.26
4	5269885	0.74	Y RNA	2.59
4	5271610	0.74	Y RNA	1.92
5	2762552	0.74	Y RNA	2.56
5	8238589	0.74	Y RNA	1.88
3	21223564	0.73	Y RNA	2.17
1	25754294	0.71	5.8S ribosomal RNA	3.08
4	14163546	0.70	U1 spliceosomal RNA	2.66
5	24061950	0.70	Signal recognition particle	2.71

Protein (59 intergenic SFP sites)

2	5444095	0.77	Riboswitch	2.71
5	18204217	0.74	Group I catalytic intron	2.64
5	18219399	0.74	Small subunit ribosomal RNA, 5' domain	2.23
2	13285678	0.71	Signal recognition particle	2.07

Sucrose (149 intergenic SFP sites)

2	5403618	0.77	HIV primer binding site	2.94
2	5504883	0.76	Y RNA	2.59
2	5504604	0.74	T-box	3.00
2	9937653	0.72	U1 spliceosomal RNA	2.40
2	9938584	0.72	miRNA	1.74
2	9944219	0.72	Y RNA	2.26
2	9937466	0.71	T-box	1.29
2	5414250	0.70	Signal recognition particle	1.33
2	5505294	0.70	Y RNA	2.30
2	9937579	0.70	Y RNA	2.33
4	8559212	0.70	Signal recognition particle	2.23

Starch (7 intergenic SFP sites)

No sequences with significant functional p-scores

Threonic acid (134 intergenic SFP sites)

Chr	Position	p-score	Assigned Rfam family	entropy
1	21164182	0.82	T-box	2.43
5	17279784	0.78	Y RNA	2.55
1	21167980	0.73	Riboswitch	3.05
2	3984298	0.72	tmRNA	3.16
2	3990529	0.72	Internal ribosome entry site	2.63
2	3994510	0.71	Enterovirus cis-acting replication element	2.93
2	19361544	0.71	Y RNA	1.83
2	3004264	0.70	Riboswitch	2.68
2	3993785	0.70	tmRNA	2.23

IV. URLs

HapMap project

http://naturalvariation.org/hapmap

Tiling-array Analysis Tool (TAS)

http://www.affymetrix.com/support/developer/downloads/TilingArrayTools/index.affx

Statistica 7.1

http://www.statsoft.com

TAIR 8 Gene Feature Format (GFF)

ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/TAIR8_gff3/TAIR8_GFF3_genes.gff

TAIR Gene Ontology annotation

ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt

TAIR Gene families

ftp://ftp.arabidopsis.org/home/tair/Genes/Gene_families

NCBI Map Viewer (Arabidopsis)

www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3702

Structure 2.2

http://pritch.bsd.uchicago.edu/structure.html

Rfam (Sanger)

http://www.sanger.ac.uk/Software/Rfam

CD-HIT

http://bioinformatics.ljcrf.edu/cd-hi

iGraph

http://cneurocvs.rmki.kfki.hu/igraph

WUBLAST

http://blast.wustl.edu

INFERNAL

http://infernal.janelia.org

HMMER

http://hmmer.janelia.org

MUSCLE

http://www.drive5.com/muscle

ClustalW

http://www.clustal.org

V. Publications

Childs L., Witucka-Wall H., Günther T., Sulpice R., v. Korff-Schmising M., Stitt M., Walther D., Schmid K., Altmann T. (2009). Natural selection mapping of genomic regions affecting growth-related metabolic traits in *Arabidopsis thaliana*. Under review

Childs L., Nikoloski Z., May P. and Walther D. (2008). Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Research* **37(9)**:e66

VI. Curriculum Vitae

Education

2006-2009 PhD student, AG Bioinformatik, University of Potsdam, Max-Planck Institute for Molecular Plant Physiology, Germany. Thesis topic: *"Bioinformatics approaches to analysing RNA mediated regulation of gene expression"*

2003-2004 Honours (1st class) in Biology, University of Sydney, Australia. Thesis topic: *"Effect of tRNA anticodon availability on codon usage in Metazoan mitochondria."*

2000-2003 B. Sc. (Bioinformatics major), University of Sydney, Australia.

Employment

2004-2006 Bioinformatician, CSIRO Plant Industry, Australia.

Awards

2006 CSIRO Plant Industry performance reward for exceptional performance. Issued by CSIRO Plant Industries, Australia.

2004 Professor Spencer Smith-White Prize for the greatest proficiency in the field of genetics in undergraduate biology honours. Issued by the School of Biological Sciences, Sydney University, Australia.

Teaching activities

2009 Seminar supervisor in lecture series "Structural Biology", Potsdam University.

Selected seminars

2009.07.24 *An alternative mechanism for translation initiation in prokaryotes*. Progress seminar. Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2008.09.19 *Arabidopsis across Europe: A tale of sweeping associations and disease resistance*. Progress seminar. Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2008.04.28 *Graph properties reflect ncRNA functional classes*. Non-Coding RNAs: Computational Challenges and Applications. Antalya, Turkey.

2008.04.04 *Capturing RNA functionality in graph properties*. Bioinformatics working seminar. Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2007.11.16 *Characterising polymorphism distribution in the Arabidopsis genome*. Havel-Spree Colloquium. Freie Universität, Berlin, Germany.

2007.09.21 *Discovery of SFPs from tiling array data*. Progress seminar. Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2007.04.24 *Whole genome genotyping*. RIKEN-Umea Workshop. Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

2006.12.07 *Analysis of Arabidopsis tiling arrays for SFP discovery*. Bioinformatics affinity seminar. Max-Planck Institute for Molecular Plant Physiology, Golm, Germany.

Bibliography

Aarssen, L.W. and Clauss, M.J. (1992) Genotypic variation in fecundity allocation in *Arabidopsis thaliana*. *J Ecol*. 109-114.

Allen, R.L., Bittner-Eddy, P.D., Grenville-Briggs, L.J., Meitz, J.C., et al. (2004) Host-Parasite Coevolutionary Conflict Between Arabidopsis and Downy Mildew, *Science*, **306**, 1957-1960.

Altman, S., Guerrier-Tokada, C., Frankfort, H.M. and Robertson, H.D. (1982) RNA-processing nucleases, *Nucleases*, 243.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., et al. (1990) Basic Local Alignment Search Tool, *Journal of Molecular Biology*, **215**, 403-410.

Bakker, E.G., Toomajian, C., Kreitman, M. and Bergelson, J. (2006) A Genome-Wide Survey of R Gene Polymorphisms in Arabidopsis, *Plant Cell*, **18**, 1803-1818.

Balasubramanian, S., Sureshkumar, S., Agrawal, M., Michael, T.P., et al. (2006) The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*, *Nature genetics*, **38**, 711 - 715.

Baltimore, D. (1970) Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses, *Nature*, **226**, 1209-1211.

Bargmann, C.I. (2001) High-throughput reverse genetics: RNAi screens in Caenorhabditis elegans, *Genome Biol*, **2**, REVIEWS1005.

Barneche, F., Gaspin, C., Guyot, R. and Echeverria, M. (2001) Identification of 66 box C/D snoRNAs in *Arabidopsis thaliana*: extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2'-O-methylation sites, *Journal of molecular biology*, **311**, 57-73.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*, **116**, 281-297.

Bashkin, J.K., Sampath, U. and Frolova, E. (1995) Ribozyme mimics as catalytic antisense reagents, *Applied biochemistry and biotechnology*, **54**, 43-56.

Benderoth, M., Textor, S., Windsor, A.J., Mitchell-Olds, T., et al. (2006) Positive selection driving diversification in plant secondary metabolism, *Proceedings of the National Academy of Sciences*, **103**, 9118-9123.

Bergelson, J., Kreitman, M., Stahl, E.A. and Tian, D. (2001) Evolutionary Dynamics of Plant R-Genes, *Science*, **292**, 2281-2285.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, **447**, 799-816.

Bonnet, E., Wuyts, J., Rouze, P. and Van de Peer, Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, *Bioinformatics (Oxford, England)*, **20**, 2911-2917.

Borevitz, J. (2006) Genotyping and mapping with high-density oligonucleotide arrays, *Methods in molecular biology (Clifton, N.J*, **323**, 137-145.

Borevitz, J.O., Hazen, S.P., Michael, T.P., Morris, G.P., et al. (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 12057-12062.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., et al. (2003) Large-scale identification of single-feature polymorphisms in complex genomes, *Genome research*, **13**, 513-523.

Borgwardt, K.M., Ong, C.S., Schonauer, S., Vishwanathan, S.V., et al. (2005) Protein function prediction via graph kernels, *Bioinformatics (Oxford, England)*, **21 Suppl 1**, i47-56.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples, *Bioinformatics (Oxford, England)*, **23**, 2633-2635.

Bradshaw Jr, H.D. and Schemske, D.W. (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers, *Nature*, **426** 176-178.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes, *Science*, **321**, 960-964.

Cao, S. and Chen, S.J. (2006) Predicting RNA pseudoknot folding thermodynamics, *Nucleic acids research*, **34**, 2634-2652.

Cech, T.R., Zaug, A.J. and Grabowski, P.J. (1981) In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence, *Cell*, **27**, 487-496.

Chan, C.S., Guo, L. and Shih, M.C. (2001) Promoter analysis of the nuclear gene encoding the chloroplast glyceraldehyde-3-phosphate dehydrogenase B subunit of Arabidopsis thaliana, *Plant Mol Biol*, **46**, 131-141.

Chang, C.C. and Lin, C.J. (2001) LIBSVM: a library for support vector machines.

Chen, H.M. and Wu, S.H. (2009) Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in Arabidopsis, *Nucleic acids research*, **37**, e69.

Chen, Y.W. and Lin, C.J. (2006) Combining SVMs with various feature selection strategies. In Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L.A. (eds), *Feature Extraction: Foundations and Applications*. Springer.

Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping, *Genetics*, **138**, 963-971.

Clark, J. and Ding, S. (2006) Generation of RNAi Libraries for High-Throughput Screens, *Journal of biomedicine & biotechnology*, **2006**, 45716.

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., et al. (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*, *Science*, **317**, 338-342.

Clausen, J., Keck, D.D. and Hiesey, W. (1940) Experimental studies on the nature of species. I. Effects of varied environments on western North American plants, *Carnegie Inst. Wash. Publ.*, **520**.

Claverie, J.M. (2005) Fewer genes, more noncoding RNA, *Science*, **309**, 1529-1530.

Contento, A.L., Kim, S.J. and Bassham, D.C. (2004) Transcriptome profiling of the response of Arabidopsis suspension culture cells to Suc starvation, *Plant physiology*, **135**, 2330-2347.

Crick, F.H. (1958) On protein synthesis, *Symposia of the Society for Experimental Biology*, **12**, 138-163.

Cross, J.M., von Korff, M., Altmann, T., Bartzetko, L., et al. (2006) Variation of enzyme activities and metabolite levels in 24 Arabidopsis accessions growing in carbon-limited conditions, *Plant Physiology*, **142**, 1574-1588.

Csárdi, G. and Nepusz, T. (2006) The igraph software package for complex network research., *InterJournal Complex Systems*, **1695**.

Dalrymple, G.B. (2001) The age of the Earth in the twentieth century: a problem (mostly) solved, *Geological Society, London, Special Publications*, **190**, 205-221.

Das, R. and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 14664-14669.

Dawson, W., Fujiwara, K., Kawai, G., Futamura, Y., et al. (2006) A method for finding optimal rna secondary structures using a new entropy model (vsfold), *Nucleosides, nucleotides & nucleic acids*, **25**, 171-189.

Dijk, H.V., Boudry, P., McCombre, H. and Vernet, P. (1997) Flowering time in wild beet (*Beta vulgaris* ssp. *maritima*) along a latitudinal cline, *Acta Oecologica*, **18**, 47-60.

Ding, Y. and Lawrence, C.E. (1999) A bayesian statistical algorithm for RNA secondary structure prediction, *Computers & chemistry*, **23**, 387-400.

Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics (Oxford, England)*, **14**, 755-763.

Eddy, S.R. (2001) Non-coding RNA genes and the modern RNA world, *Nature reviews*, **2**, 919-929.

Eddy, S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure, *BMC bioinformatics*, **3**, 18.

Eddy, S.R. (2004) How do RNA folding algorithms work?, *Nat Biotechnol*, **22**, 1457-1458.

Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models, *Nucleic acids research*, **22**, 2079-2088.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic acids research*, **32**, 1792-1797.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., et al. (2005) The Sequence Ontology: a tool for the unification of genome annotations, *Genome Biol*, **6**, R44.

Fera, D., Kim, N., Shiffeldrim, N., Zorn, J., et al. (2004) RAG: RNA-As-Graphs web resource, *BMC bioinformatics*, **5**, 88.

Freyhult, E., Gardner, P.P. and Moulton, V. (2005) A comparison of RNA folding measures, *BMC bioinformatics*, **6**, 241.

Freyhult, E., Moulton, V. and Clote, P. (2007) Boltzmann probability of RNA structural neighbors and riboswitch detection, *Bioinformatics (Oxford, England)*, **23**, 2054-2062.

Freyhult, E.K., Bollback, J.P. and Gardner, P.P. (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA, *Genome research*, **17**, 117-125.

Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep, *Nat Biotechnol*, **26**, 407-415.

Gibon, Y., Blasing, O.E., Palacios-Rojas, N., Pankovic, D., et al. (2004) Adjustment of diurnal starch turnover to short days: depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period, *Plant Journal*, **39**, 847-862.

Gilad, Y. and Borevitz, J. (2006) Using DNA microarrays to study natural variation, *Current opinion in genetics & development*, **16**, 553-558.

Gilbert, W. (1986) The RNA world, *Nature*, **319**, 618.

Gresham, D., Dunham, M. and Botstein, D. (2008) Comparing whole genomes using DNA microarrays. *Nature reviews*. 291-302.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., et al. (2003) Rfam: an RNA family database, *Nucleic acids research*, **31**, 439-441.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., et al. (2005) Rfam: annotating non-coding RNAs in complete genomes, *Nucleic acids research*, **33**, D121-124.

Gross, J.L. and Yellen, J. (2004) *Handbook of Graph Theory*. CRC Press.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., et al. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme, *Cell*, **35**, 849-857.

Hanczyc, M.M. and Szostak, J.W. (2004) Replicating vesicles as models of primitive cell growth and division, *Current opinion in chemical biology*, **8**, 660-664.

Hofacker, I.L. (2003) Vienna RNA secondary structure server, *Nucleic acids research*, **31**, 3429-3431.

Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., et al. (1998) Automatic detection of conserved RNA structure elements in complete RNA virus genomes, *Nucleic acids research*, **26**, 3825-3836.

Hofacker, I.L., Fekete, M. and Stadler, P.F. (2002) Secondary Structure Prediction for Aligned RNA Sequences, *Journal of Molecular Biology*, **319**, 1059-1066.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S.L., et al. (1994) Fast Folding and Comparison of RNA Secondary Structures, *Monatsh. Chem.*, **125**, 167--188.

Hofacker, I.L. and Stadler, P.F. (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes, *Comput Chem*, **23**, 401-414.

Hoffmann, M.H. (2002) Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae), *Journal of Biogeography*, **29**, 125-134.

Holub, E.B. (2007) Natural variation in innate immunity of a pioneer species, *Current opinion in plant biology*, **10**, 415-424.

Janssen, S., Reeder, J. and Giegerich, R. (2008) Shape based indexing for faster search of RNA family databases, *BMC bioinformatics*, **9**, 131.

Jiang, M., Anderson, J., Gillespie, J. and Mayne, M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts, *BMC bioinformatics*, **9**, 192.

Karklin, Y., Meraz, R.F. and Holbrook, S.R. (2005) Classification of non-coding RNA using graph representations of secondary structure, *Pacific Symposium on Biocomputing*, **10**, 4-15.

Keurentjes, J.J.B., Sulpice, R., Gibon, Y., Steinhauser, M.-C., et al. (In Press) Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in *Arabidopsis thaliana*, *Genome Biology*, **X**, X.

Kim, N., Shiffeldrim, N., Gan, H.H. and Schlick, T. (2004) Candidates for novel RNA topologies, *Journal of molecular biology*, **341**, 1129-1144.

Kim, S., Zhao, K., Jiang, R., Molitor, J., et al. (2006) Association mapping with single-feature polymorphisms, *Genetics*, **173**, 1125-1133.

Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D. (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*, *Annual review of plant biology*, **55**, 141-172.

Korves, T.M., Schmid, K.J., Caicedo, A.L., Mays, C., et al. (2007) Fitness Effects Associated with the Major Flowering Time Gene FRIGIDA in *Arabidopsis thaliana* in the Field, *The American Naturalist*, **169**, E141-E157.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., et al. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena, *Cell*, **31**, 147-157.

Kudla, G., Murray, A.W., Tollervey, D. and Plotkin, J.B. (2009) Coding-sequence determinants of gene expression in Escherichia coli, *Science*, **324**, 255-258.

Kurihara, Y., Matsui, A., Kawashima, M., Kaminuma, E., et al. (2008) Identification of the candidate genes regulated by RNA-directed DNA methylation in Arabidopsis, *Biochemical and biophysical research communications*, **376**, 553-557.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., et al. (2007) Clustal W and Clustal X version 2.0, *Bioinformatics (Oxford, England)*, **23**, 2947-2948.

Le Corre, V. (2005) Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits, *Molecular ecology*, **14**, 4181-4192.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, *Cell*, **75**, 843-854.

Leslie, C.S., Eskin, E., Cohen, A., Weston, J., et al. (2004) Mismatch string kernels for discriminative protein classification, *Bioinformatics (Oxford, England)*, **20**, 467-476.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics (Oxford, England)*, **22**, 1658-1659.

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., et al. (2003) The microRNAs of Caenorhabditis elegans, *Genes & development*, **17**, 991-1008.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., et al. (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants, *Nature Protocols*, **1**, 387-396.

Lu, C., Kulkarni, K., Souret, F.F., MuthuValliappan, R., et al. (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant, *Genome research*, **16**, 1276-1288.

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., et al. (2005) Elucidation of the small RNA component of the transcriptome, *Science*, **309**, 1567-1569.

Lu, Y.D., Gan, Q.H., Chi, X.Y. and Qin, S. (2008) Roles of microRNA in plant defense and virus offense interaction, *Plant cell reports*, **27**, 1571-1579.

Lyngso, R.B. and Pedersen, C.N. (2000) RNA pseudoknot prediction in energy-based models, *Journal of Computational Biology*, **7**, 409-427.

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., et al. (2006) Biological function of unannotated transcription during the early development of Drosophila melanogaster, *Nature genetics*, **38**, 1151-1158.

Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics, *Trends Genet*, **24**, 133-141.

Mathews, D.H. and Turner, D.H. (2006) Prediction of RNA secondary structure by free energy minimization, *Current opinion in structural biology*, **16**, 270-278.

Mattick, J.S. (2007) A new paradigm for developmental biology, *Journal of Experimental Biology*, **210**, 1526-1547.

Mauricio, R., Stahl, E.A., Korves, T., Tian, D., et al. (2003) Natural Selection for Polymorphism in the Disease Resistance Gene Rps2 of *Arabidopsis thaliana*, *Genetics*, **163**, 735-746.

Meyer, R.C., Steinfath, M., Lisec, J., Becher, M., et al. (2007) The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4759-4764.

Meyers, B.C., Matzke, M. and Sundaresan, V. (2008) The RNA world is alive and well, *Trends in plant science*, **13**, 311-313.

Miller, S.L. and Urey, H.C. (1959) Organic compound synthesis on the primitive earth, *Science*, **130**, 245-251.

Mironov, A.S., Gusarov, I., Rafikov, R., Lopez, L.E., et al. (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria, *Cell*, **111**, 747-756.

Mitchell-Olds, T. and Pedersen, D. (1998) The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis, *Genetics*, **149**, 739-747.

Mondragon-Palomino, M., Meyers, B.C., Michelmore, R.W. and Gaut, B.S. (2002) Patterns of Positive Selection in the Complete NBS-LRR Gene Family of *Arabidopsis thaliana*, *Genome Res.*, **12**, 1305-1315.

Mourier, T., Carret, C., Kyes, S., Christodoulou, Z., et al. (2008) Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum, *Genome research*, **18**, 281-292.

Myslyuk, I., Doniger, T., Horesh, Y., Hury, A., et al. (2008) Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in trypanosomatid genomes, *BMC bioinformatics*, **9**, 471.

Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., et al. (2002) Genetic control by a metabolite binding mRNA, *Chemistry & biology*, **9**, 1043-1049.

Nakamoto, T. (2009) Evolution and the universality of the mechanism of initiation of protein synthesis, *Gene*, **432**, 1-6.

Nakashima, A., Takaku, H., Shibata, H.S., Negishi, Y., et al. (2007) Gene silencing by the tRNA maturase tRNase ZL under the direction of small-guide RNA, *Gene therapy*, **14**, 78-85.

Nissen, P., Hansen, J., Ban, N., Moore, P.B., et al. (2000) The structural basis of ribosome activity in peptide bond synthesis, *Science*, **289**, 920-930.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., et al. (2005) The pattern of polymorphism in *Arabidopsis thaliana*, *PLoS Biology*, **3**, e196.

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., et al. (2006) Identification and classification of conserved RNA secondary structures in the human genome, *PLoS computational biology*, **2**, e33.

Powner, M.W., Gerland, B. and Sutherland, J.D. (2009) Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions, *Nature*, **459**, 239-242.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945-959.

Qu, L.H., Meng, Q., Zhou, H. and Chen, Y.Q. (2001) Identification of 10 novel snoRNA gene clusters from *Arabidopsis thaliana*, *Nucleic acids research*, **29**, 1623-1630.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**, 257-286.

Rivas, E. and Eddy, S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs, *Bioinformatics (Oxford, England)*, **16**, 583-605.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure, *Nature*, **419**, 832-837.

Schauer, N., Semel, Y., Roessner, U., Gur, A., et al. (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement, *Nature Biotechnology*, **24**, 447-454.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., et al. (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples, *Febs Letters*, **579**, 1332-1337.

Schemske, D.W., Bierzychudek, Paulette (2007) Spatial differentiation for flower color in the desert annual *Linanthus parryae*: Was Wright right?, *Evolution*, **61**, 2528-2543.

Schilling, O., Langbein, I., Muller, M., Schmalisch, M.H., et al. (2004) A protein-dependent riboswitch controlling ptsGHI operon expression in Bacillus subtilis: RNA structure rather than sequence provides interaction specificity, *Nucleic acids research*, **32**, 2853-2864.

Schmid, K.J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., et al. (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism, *Genetics*, **169**, 1601-1615.

Schmid, K.J., Torjek, O., Meyer, R., Schmuths, H., et al. (2006) Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers, *TAG. Theoretical and applied genetics*, **112**, 1104-1114.

Schopf, J.W., Kudryavtsev, A.B., Agresti, D.G., Wdowiak, T.J., et al. (2002) Laser--Raman imagery of Earth's earliest fossils, *Nature*, **416**, 73-76.

Sessions, A., Burke, E., Presting, G., Aux, G., et al. (2002) A high-throughput Arabidopsis reverse genetics system, *Plant Cell*, **14**, 2985-2994.

Sharbel, T.F., Haubold, B. and Mitchell-Olds, T. (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe, *Molecular Ecology*, **9**, 2109-2118.

Shine, J. and Dalgarno, L. (1975) Determinant of cistron specificity in bacterial ribosomes, *Nature*, **254**, 34-38.

Smith, S.E. and Macnair, M.R. (1998) Hypostatic modifiers cause variation in degree of copper tolerance in Mimulus guttatus, *Heredity*, **80**, 760-768.

Song, D., Yang, Y., Yu, B., Zheng, B., et al. (2009) Computational prediction of novel non-coding RNAs in Arabidopsis thaliana, *BMC bioinformatics*, **10 Suppl 1**, S36.

Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., et al. (1999) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis, *Nature*, **400**, 667-671.

Starmer, J., Stomp, A., Vouk, M. and Bitzer, D. (2006) Predicting Shine-Dalgarno sequence locations exposes genome annotation errors, *PLoS computational biology*, **2**, e57.

Steigele, S., Huber, W., Stocsits, C., Stadler, P.F., et al. (2007) Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions, *BMC biology*, **5**, 25.

Stinchcombe, J.R., Weinig, C., Ungerer, M., Olsen, K.M., et al. (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4712-4717.

Stuitje, A.R., Verbree, E.C., van der Linden, K.H., Mietkiewska, E.M., et al. (2003) Seed-expressed fluorescent proteins as versatile tools for easy (co)transformation and high-throughput functional genomics in Arabidopsis, *Plant biotechnology journal*, **1**, 301-309.

Sudarsan, N., Barrick, J.E. and Breaker, R.R. (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes, *RNA (New York, N.Y*, **9**, 644-647.

Szostak, J.W. (2009) Origins of life: Systems chemistry on early Earth, *Nature*, **459**, 171-172.

Temin, H.M. and Mizutani, S. (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus, *Nature*, **226**, 1211-1213.

Thilmony, R., Underwood, W. and Sheng Yang He (2006) Genome-wide transcriptional analysis of the *Arabidopsis thaliana* interaction with the plant pathogen *Pseudomonas*

syringae pv. *tomato* DC3000 and the human pathogen *Escherichia coli* O157:H7, *The Plant Journal*, **46**, 34-53.

Tian, D., Araki, H., Stahl, E., Bergelson, J., et al. (2002) Signature of balancing selection in Arabidopsis, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11525-11530.

Tinoco, I., Jr., Uhlenbeck, O.C. and Levine, M.D. (1971) Estimation of secondary structure in ribonucleic acids, *Nature*, **230**, 362-367.

Toerjek, O., Berger, D., Meyer, R.C., Muessig, C., et al. (2003) Establishment of a high-efficiency SNP-based framework marker set for Arabidopsis, *Plant Journal*, **36**, 122-140.

Toomajian, C., Hu, T.T., Aranzana, M.J., Lister, C., et al. (2006) A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome, *PLoS Biology*, **4**, e137.

Tuerk, C. and Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, *Science*, **249**, 505-510.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.

Warthmann, N., Fitz, J. and Weigel, D. (2007) MSQT for choosing SNP assays from multiple DNA alignments, *Bioinformatics (Oxford, England)*, **23**, 2784-2787.

Washietl, S., Hofacker, I.L. and Stadler, P.F. (2005) Fast and reliable prediction of noncoding RNAs, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2454-2459.

Weinstock, G.M. (2007) ENCODE: more genomic empowerment, *Genome research*, **17**, 667-668.

Winkler, W., Nahvi, A. and Breaker, R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression, *Nature*, **419**, 952-956.

Winkler, W.C., Cohen-Chalamish, S. and Breaker, R.R. (2002) An mRNA structure that controls gene expression by binding FMN, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15908-15913.

Yao, Z., Weinberg, Z. and Ruzzo, W.L. (2006) CMfinder--a covariance model based RNA motif finding algorithm, *Bioinformatics (Oxford, England)*, **22**, 445-452.

Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nature genetics*, **38**, 203-208.

Zhang, Y. (2005) miRU: an automated plant miRNA target prediction server, *Nucleic acids research*, **33**, W701-704.

Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic acids research*, **31**, 3406-3415.