A genomic study of the nuclear matrix attachment region recognition signature

Alasdair Anthony

Thesis presented in accordance with the requirements for the degree of Doctor of Philosophy

University of Edinburgh

2008

Declaration

I declare that this thesis has been composed by myself and, except where otherwise stated, is entirely my own work.

Alasdair Anthony

Contributions of co-authors

Some of the work described here forms the basis of a published paper:

Anthony A and Blaxter M (2007). Association of the Matrix Attachment Region Recognition Signature with coding regions in *Caenorhabditis elegans. BMC Genomics,* **8:***418.*

Alasdair Anthony carried out the analyses and wrote the manuscript. Mark Blaxter assisted with the analyses and writing the manuscript and supervised the project.

Abstract

Matrix attachment regions (MAR) are the sites on genomic DNA that interact with the nuclear matrix. A complex bipartite motif, the MAR recognition signature (MRS), has been proposed as a DNA sequence marker for MAR but its specificity and sensitivity remain unresolved.

I describe here the distribution of the MRS in the genomes of a number of species from across animals and plants. The MRS is shown to have a distinctive, non-random distribution, with a particular relationship to genes. This relationship was studied in detail in the genome of *Caenorhabditis elegans*, revealing striking peaks of average MRS frequency in the regions flanking *C. elegans* genes. The occurrence of similar peaks in *C. briggsae, Danio rerio, Arabidopsis thaliana, Drosophila melanogaster* and *Homo sapiens* was also investigated. The nucleotide content in the vicinity of genes is examined and it too is shown to have striking peaks in regions surrounding genes. *C. elegans* genes associated with MRS were found to be significantly enriched for receptor activity annotations but not for some other features.

Using this analysis of the genomic distribution of the MRS, the relevance of the MRS as a marker for MAR is discussed. The potential for MRS to play a functional role, as indicated by their peculiar frequency in the vicinity of genes, is also explored.

Abbreviations used

AQL: Ace Query Language

CDS: Coding Sequence

GO: The Gene Ontology

IEA: Inferred from Electronic Annotation

mRNA: messenger RNA

MAR: Matrix Attachment Region

MRS: MAR Recognition Signature

ncRNA: non-coding RNA

OR: Odds Ratio

rRNA: ribosomal RNA

SAGE: Serial Analysis of Gene Expression

SL: Spliced Leader

snRNA: small nuclear RNA

tRNA: transfer RNA

UTR: Untranslated Region

Contents

Table Index

Figure Index

Chapter 1 – Introduction

1.1 Non-coding DNA

The publishing of the complete genome sequence of the bacterium *Haemophilus influenzae* in 1995 heralded a new era in genome biology [1]. A year later the first eukaryotic genome to be sequenced, that of the yeast *Saccharomyces cerevisiae* [2], was completed, followed by the first animal genome, *Caenorhabditis elegans* [3], in 1998 and the draft human genome sequence in 2001 [4]. Over 700 genomes from all kingdoms of life have now been sequenced and the list continues to grow [5]. The study of an organism's genome provides a platform for the characterisation of its genes and the genomic landscape in which they reside.

Until recently, most genomic studies have focused on the identification and analysis of genes - the parts of the genome that are transcribed and translated. Indeed, the central dogma of biology, that DNA makes RNA makes protein, makes this is a logical course of action. However, whole genome sequencing and subsequent annotation has revealed that the translated portion of the genome is only a small fraction of the total, as little as 1-2% in mammals [4, 6]. Furthermore, the surprising revelation that the human genome not only has a similar number of genes to other mammals but also to that of the nematode worm *Caenorhabditis elegans* indicates that there is more to genomes than genes [3, 4, 6, 7] . Once often referred to as “junk DNA” it is now apparent that the non-protein coding regions of the genome have an important role to play. The regulation of genes, and therefore the genomic sequence involved in that regulation, is of equal importance to the genes themselves.

Comparative genomics is a powerful tool in modern functional characterisation of genomes. It relies on the principle that mutations in functional genomic elements are likely to be deleterious and therefore eliminated by purifying selection. Purifying selection results in evolutionary constraint on functional sequence and can be

detected as sequence conservation between species. Comparisons with other mammalian genomes have revealed that at least 5% of the human genome is under purifying selection, and thus likely to be functional [6-8]. The presence of ultra-conserved non-coding sequence has been revealed by whole genome comparisons, for example between humans and *Fugu rubripes* [9], murids and humans [10], 14 mammals [11], insects [12] and nematodes [13]. Surprisingly, the level of conservation between these elements often exceeds that found between orthologous protein sequences.

Inter-species sequence conservation can therefore, when observed in excess of that predicted by a neutral model, imply functionality. However, lack of conservation does not preclude a sequence from functionality. Large numbers of functional sequences for which no detectable conservation exists have been reported in several species, including human [14], fly [15], and zebrafish [16]. Additionally, an in-depth analysis of the human genome, as part of the ENCODE project, found that as much as 80% of the genome is transcribed, although transcription alone is not sufficient to infer function [14, 17].

Functional non-coding sequence can be assigned to a wide range of roles. One is that the conserved sequences form some kind of non-protein coding RNA (ncRNA) [18]. Although several types of ncRNA can be reliably annotated (e.g. rRNA, tRNA, snRNAs), microRNAs were only recently discovered and other novel ncRNAs may be the source of some conserved sequences. It is also likely that a significant portion of functional non-coding sequence represents *cis* regulatory regions. The evidence for this is conflicting, perhaps reflecting the wide range of different types of element that have been identified. For instance, the gene-independent distribution of some conserved non-coding regions makes it less likely that they are involved in direct control of gene expression [19, 20]. In addition, the level of sequence conservation of conserved non-coding regions is generally much greater than that seen for many experimentally proven *cis* regulatory elements. Nonetheless, certain conserved non-coding sequences, such as the ultra-conserved elements, tend to be associated with

particular classes of genes particularly those involving transcriptional regulation and development, suggesting that these elements may be distal enhancers of genes [9, 10, 13, 19] Furthermore, it has been shown that some conserved sequences exhibit a clustering of conserved sites, in a pattern reminiscent of protein binding motifs [11]. The interesting possibility that some sequence may function through *trans* regulation of gene expression has also been raised [19, 20]. A purely structural functional role for sequence in genomic architecture has been proposed, potentially involving formation of protein bridges between chromosomes, which subsequently facilitate the movement or distribution of chromosomes around the nucleus [9, 19]. Finally, it has been postulated that a portion of functional sequence may be involved in the attachment of chromatin to the nuclear matrix. Glazko *et al.* found that a significant fraction of human -mouse "homologous intergenic tracts" occurred in regions predicted to contain nuclear matrix attachment regions [21]. This interaction of DNA and the nuclear matrix is the aspect of non-coding DNA functionality that this thesis is concerned with.

1.2 Matrix attachment regions

The organization of DNA into highly a condensed structure is important for a range of genome functions, including gene regulation, DNA synthesis, recombination and DNA repair. Several hierarchical levels are involved in the compaction of DNA in nuclei. The elementary building block of DNA packing is the nucleosome, the repeating unit of DNA and histone proteins. Further compaction results in a chromatin fibre of about 30 nm in diameter, although precise characteristics of this fibre are influenced by the spacing between nucleosomes [22]. Higher order packaging involves the formation of distinct chromatin loops by periodic attachment to the nuclear matrix [23].

The nuclear matrix is a complex association of over 500 proteins that form a three dimensional network throughout the nucleus [24]. Although many nuclear matrix proteins have been identified, less is known about how these proteins assemble to make the fibres, filaments and other assemblies that constitute the nuclear matrix

[25]. DNA attachment to the nuclear matrix is not random; it occurs at specific sites known as matrix (or scaffold) attachment regions (MAR). It has been estimated that there are about 100,000 MAR in mammalian genomes, ranging in length from 300 bp to several kilobases [26]. Experimentally, MAR have been defined as either DNA fragments that remain bound to the nuclear matrix after chromatin proteins and other DNA have been removed, or DNA that binds to extracted nuclear matrix in the presence of competitor DNA [27, 28]. The most common experimental method for identifying MAR uses re-association assays to define DNA fragments that bind to the nuclear matrix [29].

Despite the large body of evidence surrounding MAR, the validity of their existence as a true biological entities *in vivo* has been called into question. Issues have been raised surrounding the ease with which nuclear marcomolecules interact in different ways according to the prevailing environmental conditions [30]. It has been suggested that MAR actually bind various DNA binding proteins, such as replication and transcription machinery, during preparation of the matrix but not *in vivo* [31]. The co-localisation of MAR with regions of DNA specifically designed to bind multi-protein complexes, e.g. replication origins and transcription regulatory elements, has been cited as evidence to support this theory [31] . However, recent observations of the nuclear matrix *in vivo* by electron microscopy and in living cells using florescently tagged proteins have essentially confirmed the validity of the nuclear matrix concept [25, 32].

In addition to their involvement in DNA packaging, MAR have been found to coincide with origins of replication, centromeres and telomeres [33, 34]. MAR have also been implicated in a number of other functional roles, mainly relating to gene expression. For example, when the human gene HLA-G was positioned near a MAR its expression was down-regulated and when positioned away from the base of the chromatin loop, its expression was increased [35]. Restriction of the effects of long-range enhancers by MAR can also cause inhibition of gene expression [36]. Conversely, MAR have also been associated with enhancement of gene expression,

by anchoring DNA in such a way as to shorten chromatin loops [37]. The ability of MAR to enhance gene expression has been exploited in transgene constructs, particularly in plants [38]. One route through which MAR are thought to act is in the reduction of gene silencing, for example by shielding transgenes from RNA silencing [39]. MAR sequence added to transgenic constructs have also been shown to increase transgene expression by both increasing the percentage of cells expressing the transgene and by increasing level of expression per cell [40]. MAR are also an important component in setting up functional domains in the nucleus [41]. As such they may also act as a boundary between functional domains [24, 42]. Similarly, MAR have been proposed as “boundary elements”, segregating genes into separate regulatory modules [43]. MAR have also been implicated in the positioning of chromosomal territories [44]. This allows coordinated spatial positioning of sequences on different chromosomes to facilitate interactions in *trans*. For example, active genes from different chromosomes have been shown to migrate through the nuclear space to converge on "transcriptional factories" [45]. Localisation of genes in this way is likely to involve control of higher order chromosome structure. There is increasing evidence for the dynamic association of MARs with the nuclear matrix. One model of matrix attachment suggests that there are two types of MAR: those attached to DNA permanently and those interacting with DNA dynamically in response to functional demands [46]. This is supported by evidence that some chromatin loop attachments are under developmental control and that transgenic MAR in mice are selected and used as nuclear matrix anchors in a discriminatory manner [41, 47]. Furthermore, the selective use of MAR appears to be directly linked to the movement of the loops.

Genome-wide mapping of MAR would provide an invaluable tool for further dissection of MAR function. The traditional experimental methods of MAR identification are poorly amenable to genome wide analysis, though recent application of high-throughput cosmid and olgionucleotide array technologies may prove useful in large-scale identification of MAR [48]. However, the use of such methods in MAR identification is in its infancy and recent years have seen a focus on

the development of *in silico* methods. These methods rely on detection of the sequence characteristics of MAR, such as is catalogued for approximately 500 experimentally defined MAR in the MAR transaction Database [49]. Analysis of MAR sequence has revealed that the overriding feature of many MAR is that they are AT rich, but several other more specific sequence motifs are also characteristic of MAR. A compilation of sequence motifs often associated with MAR is presented in Table 1.1.

A number of methods for the computational identification of MAR exist. One of the first attempts at automated MAR prediction was a rule based method, MAR-finder [50]. Motifs known to occur within MAR, including origin of replication, TG-rich sequences, curved DNA, kinked DNA and topoisomerase sites, were used to formulate rules. A score, based on the statistical significance of the presence of these patterns in relation to the surrounding sequence, is then calculated. Around the same time, a method of predicting MAR based on the stabilization of the DNA duplex was developed [51]. The stress-induced duplex DNA destabilisation (SIDD) method predicts where torsional stress in DNA is relieved through strand separation. On the basis of a thermodynamic model, SIDD produces a graph showing the energy required for a given base pair to separate. Another method, ChrClass, is based on multivariate linear discriminant analysis to compare the sequences of experimentally defined MAR [52]. The classification is augmented by the use of random sequences as non- matrix related controls. SMARTest is based on 97 position weight matrices that describe MAR associated motifs 10-21 bases in length [26]. The position weight matrices were generated from 34 animal and plant MARs, aligned using DiAlign [53]. For each of the methods described above, generally favourable reports on their effectiveness have been published. MAR-finder is reported to give 80% precision and 32% sensitivity [26], ChrClass a precision of 50% and sensitivity of 85% [54], SMARTest a precision of 68% and sensitivity of 38%. However, some researchers have commented that there is no correlation between MAR prediction methods and experimentally defined MAR [56], and a recent study stated that all of the available

Sequence motif	Description	Reference
ATTA, ATTTA, ATTTTA	Origin of replication sites	[52,56]
$(TG)_{rich}$, TGTTTTG, TGTTTTTTG, TTTTGGGG	TG-rich signal	[52]
$A_4N_7A_{3/4}N_7A_{4,}$ TTTAAA	Curved DNA	[52]
$TAN_{3\text{-}8}TG$, $CAN_{3\text{-}8}TA$, $TGN_{3\text{-}8}TA$	Kinked DNA	[52]
$(RY)_n$, GTNWAYATTNATNNR	Topoisomerase-II binding signal for *D. melanogaster*	[57,58]
$(GC)_n$, $(AT)_n$	Z-form DNA	[23]
A_n,C_n,G_n,T_n	Homopolytracts - lamin binding	[23]
R_n, Y_n, S_n, K_n, M_n	Short repeats (>6)	[23]
AATATATTT	Base un-pairing sequence	[23,28]
$(TTAGGG)_n$	Vertebrate telomeric repeats	[23]
TCTTTAATTTCTAATAT ATTTAGAA	SATB1 binding motif	[28,52]
$(A/C/T)_n$	H-rule, helix destabilization	[52,59]

Table 1.1 Sequence motifs associated with MAR

MAR prediction methods had very little predictive power [57]. Nonetheless, genome-wide mapping of MAR using *in silico* methods has been undertaken, specifically, SMARTest was used to predict MAR in the genome of *Arabidopsis thaliana* [58].

1.3 MAR recognition signature

A. thaliana has been subjected to several attempts to identify MAR experimentally, albeit on smaller genome regions than in the computational study. In one such study a 16 kb region around the plastocyanin gene was investigated, revealing three MAR [59]. The sequence making up the MAR containing restriction fragments was aligned to facilitate the identification of sequence shared across all three MAR. This lead to the identification of a degenerate 21 bp sequence, shared between the MAR but not found elsewhere on the 16 kb fragment. An additional four MAR containing fragments were identified elsewhere in the *A. thaliana* genome. In each of these fragments the 21 bp sequence was found as two closely spaced sequences. Further analysis of all seven MAR revealed that all had a closely spaced combination of a 7 bp and 12 bp sequence, in a number of different configurations. The concept that two sequence elements could together identify MAR was researched further, leading to the identification of an additional conserved element. This element consisted of an extension of the previously identified 12bp sequence to 16bp, found in conjunction with an 8 bp sequence. Both the 16 bp and 8 bp sequences exist in numerous locations through the *A. thaliana* genome, but the combination of the two sequences within 200 bp was only found in the experimentally defined MAR and not elsewhere in the genome. This new element was termed the MAR recognition signature (MRS) [42]. Surprisingly, when a 33 kb fragment of *C. elegans* genomic DNA was screened, nine MRS were identified, all of them mapping to one of five experimentally defined MAR. In the *C. elegans* genome fragment one additional experimentally defined MAR was found that did not contain any MRS. Similarly, the MRS was found in some, but not all, of the MAR previously identified in yeast, chicken, mouse, rabbit and human (MRS were found in 20 of 27 MAR). Therefore, the MRS appeared to be a reliable marker of a subset of MAR. The possible existence of two types of MAR,

dynamic and constitutive, was raised above [46]. In this scenario it is easy to imagine that MRS are representative of one type of MAR. The specific function of the MRS remains unresolved, although preliminary experiments suggested that the MRS itself does not bind the nuclear matrix [59].

Since its discovery, the MRS has been used to successfully predict MAR *de novo.* MRS were identified in both of two selected genomic fragments from the protozoan parasite, *Entamoeba histolytica,* a species for which no MAR had previously been identified [60]. Experimental procedures showed that both fragments bound nuclear matrix extracts, confirming that they were MAR. In another study, five MRS were identified near the LMP/TAP gene cluster in the human genome [61]. The five genomic fragments containing these MRS were all found to bind the nuclear matrix, while two random fragments from the same region did not. Three of the MRS containing fragments were then shown to actively recruit the mRNA processing protein hn RNP-A1 during transcriptional activation of nearby genes.

Post hoc analysis of previously identified MAR has shown the predictive power of the MRS to be less robust. For example, MAR mapping studies in mammals have shown that MRS are sometimes identified outside known MAR [62]. In their analysis of 1 Mb of the mouse genome, Purbowasito *et al*. reported that MAR prediction based on MRS had a specificity of 41%, with 29 of 49 predictions lying outside experimentally defined MAR [54]. In addition, in their review of MAR prediction tools, Evans *et al.* found the MRS, along with all the other tools studied, to be poor predictors of MAR [57]. There is, therefore, some doubt as to the effectiveness of the MRS as a marker for MAR.

1.4 Genomes

An understanding of genomic environment in which any DNA sequence motif analysis is conducted is necessary to place the research in context. In this thesis, analyses covering six species is presented. Chapters two and three are concerned solely with the genomes of the nematodes *Caenorhabditis elegans* and

Caenorhabditis briggsae with a particular focus on the former. In chapter 4, four additional species are studied: the flowering plant, *Arabidopsis thaliana*, the zebrafish *Danio rerio*, the fruit fly *Drosophila melanogaster* and the mammal, *Homo sapiens*. All these species have been intensively studied. The anatomic simplicity of *C. elegans* lead to its use as model for animal development, and ultimately contributed to it being the first animal to have its genome sequenced. As a close relative, the genome sequence of *C. briggsae* complements that of *C. elegans*. The ease of cultivation of *A. thaliana* has contributed to it becoming representative of the plant kingdom. Similarly, ease of growth in the laboratory has allowed *D. melanogaster* to become one of the most studied organisms. As a vertebrate model, *D. rerio* is considered particularly useful because of the availability of a large number of mutations affecting development and physiology [63]. Finally, many of these species are used as a model to further our understanding of our own species, *H. sapiens*. The salient features the genomes of each of theses six species are compared here.

The genomes of both *C. elegans* and its close relative *C. briggsae* are composed of approximately 100 Mb, comparable with the 125 Mb genome of *A. thaliana* and the 180 Mb genome of *D. melanogaster* [64, 65]. However, these genomes are dwarfed by the 1,500 Mb genome of *D. rerio* and the 3,000 Mb genome of *H. sapiens* [65]. The two nematode genomes are split, fairly evenly, across six chromosomes. Unlike most other animals, their chromosomes have no cytologically defined centromeres [3]. These two nematodes also lack heterochromatic regions (highly repetitive portions of the genome, often associated with centromeric and telomeric regions in other species). Instead, the highly repetitive sequence characteristic of centromeres in other organisms, is replaced by tandem repeats scattered along the chromosome, particularly the arms [3]. Although the chromosome number of *D. melanogaster* is similar, it differs greatly from *C. elegans*, and many other organisms, in the structure of its chromosomes. All of the *D. melanogaster* Y chromosome, most of chromosome 4, half of the X chromosome and the centromeric portions of the other chromosomes is composed of tightly packed, gene poor DNA - heterochromatin [66].

The euchromatic portion of the *D. melanogaster* genome makes up two thirds of the nucleotides but harbours 98% of the 13,500 genes [66].

Gene distribution in *C. elegans* is also non-uniform. The majority of its genes are found in the central two thirds of the chromosomes, except in the X chromosome, where the distribution is more even. Of the 20,000 *C. elegan*s genes, about 2,800 are contained within 1,053 trans-spliced operons [67]. This transcriptional feature is unusual amongst metazoans, although operons are found in *C. briggsae*.

With 27,000 protein coding genes across its five chromosomes the genome of *A. thaliana* is the most gene dense of those studied here. The most notable feature of genome organisation in *A. thaliana* is the prevalence of tandem gene arrays and segmental duplications which have created a degree of redundancy [68]. Up to 17% of the genes are arranged in tandem arrays and a high proportion (37%) can be assigned, on the basis of similarity, to gene families [68]. Both of these factors likely contribute to the large number of genes in *A. thaliana.*

Perhaps surprisingly, despite their large size and the complexity of the organisms themselves, the genomes of *D. rerio* and *H. sapiens* do not contain large numbers of genes. This does, however, mean that both these genomes have much lower gene densities than the other species. In *D. rerio*, the 17,500 genes are spread evenly across 25 similarly sized chromosomes [63]. In the 26 chromosomes of *H. sapiens* there is a strong correlation between G+C% and gene density [4, 6]. Studies into G+C profile of have lead to the identification of isochores - long range sections (>300 kb) of uniform GC content. A large proportion of genes are found in the GC rich isochores [69]. The high G+C% makes the DNA duplex more stable and decreases nucleosome formation potential, leading to the theory that isochrores function by optimising genome structure for epigenetic control [70].

The frequency and distribution of repeats is another feature that distinguishes nematode, insect and mammalian genomes. Repeats make up 16.5% of the *C. elegans* genome and slightly more of the *C. briggsae* genome. In both species the repeats are most prevalent on the arms of the chromosomes [67]. The vast majority of

repeats in the *D. melanogaster* genome are found in the heterochromatin, although ~4% of the euchromatin is made up of transposable elements [66]. A striking feature of mammalian genomes is the high proportion of the genome formed by repeats, up to 46% in humans [6]. A similar proportion of the *D. rerio* genome is also composed of repeats.

Analysis of repeats have given insight into the evolution of the nematode genomes. *C. elegans* and *C. briggsae* share very few repeat families, indicating that most of the repeats were acquired after the species diverged or that they are undergoing rapid evolution [67]. Half of the *C. elegans* and *C. briggsae* genomes were aligned at the nucleotide level using the WABA algorithm [71]. Surprisingly, just one third of the aligned bases lay in coding exons, another third lay in introns and the final third in intergenic regions [67].

The evolution of genomes is directed by large scale chromosomal rearrangements as well as point mutations. Again there are both differences and similarities between the species studied here. A common feature across most genomes is the high rate of rearrangements within chromosomes rather than between chromosomes [6, 67]. However, the rate of chromosomal rearrangements is significantly different. It is estimated that there have been 57 breaks/Mb since the speciation of *C. elegans* and *C. briggsae*, while the breakpoint rate in *H. sapiens* is significantly lower [67].

One of the most remarkable features of the Caenorhabditis genomes is the distinction between the arms and the centres of the chromosomes. The centres of the chromosomes are gene rich, with a particularly high frequency of essential genes. The centres also exhibit a lower rate of meiotic recombination. By contrast, the chromosomal arms are characterised by a high density of repeats and are subject to more frequent chromosomal rearrangements. The chromosomal separation of these features suggests that the arms are the major site of evolution in *C. elegans* and *C. briggsae*. In contrast, analysis of *A. thaliana* genome has revealed that low gene expression and high repeat density is correlated with low recombination rates [68].

1.5 Aims and thesis plan

There is increasing evidence for the extensive involvement of MAR in the control of gene expression. However, current methods for experimentally predicting MAR are not amenable to large-scale analyses, and only a small fraction of the anticipated total number of MAR has been experimentally defined. Computational methods have facilitated reliable genome-wide annotation in fields such as gene finding but *in silico* MAR prediction has suffered from a lack of accuracy and sensitivity. The MRS was initially shown to be exclusive to at least a sub set of MAR, and thus a good marker for MAR. Some subsequent studies have successfully used the MRS to identify MAR, yet others have brought its validity as a reliable marker of MAR into question. The primary focus of this thesis is a study of the MRS with the initial aim of resolving its ability to predict MAR.

The main obstacle to measuring the effectiveness of MRS, or any *in silico* method, to predict MAR is the lack of experimentally verified MAR. The six MAR identified by van Drunen *et al.* (1999) that lead to the creation of the MRS remain the only MAR to have been experimentally defined in *C. elegans* [42]. However, in the intervening years another valuable resource has become available, the complete genome sequence of *C. elegans* and many other species. This allows the incidence of every MRS in a genome to be identified. We can ask questions of this 'map' of MRS occurrence to determine if it matches the identifying characteristics of experimentally defined MAR. In this way, this thesis aims to make a qualified judgement as to how effective a predictor of MAR the MRS is.

In addition to some understanding of characteristics, such as their location in the genome, MAR have also been described as evolutionarily conserved [21, 72]. Furthermore, the MRS has also been demonstrated to reside in MAR from a wide range of species, including plants, invertebrates and vertebrates [42]. It is therefore reasonable to expect a degree of conservation of MRS. Thus conservation of the MRS was studied in a number of different ways, depending on the level of conservation expected. Fortunately, the genome of *C. briggsae,* one of the closest

relatives of *C. elegans*, has been sequenced. This means a comparison of MRS incidence between these two related species could be made, and the degree of conservation assessed. Comparison of MRS incidence across much greater evolutionary distances was also achieved by making use of other fully sequenced genomes. When considering inter-phylum comparisons, we would not necessarily expect conservation of individual MRS but conservation of general genomic positioning patterns would be indicative of a functional role for the MRS. Comparison of MRS pattern between the relatively compact genomes of *C. elegans* and *A. thaliana,* in which the MRS was originally derived, and the much larger genomes of, for example *H. sapiens,* also inform on the characteristics of MRS distribution.

This thesis also presents an analysis of the MRS itself, rather than its distribution. Not surprisingly, the MRS replicates the AT richness of MAR. As an AT rich (and degenerate) motif its incidence will be influenced by the background AT content of the genomes studied. Assessing the degree to which the MRS occurs due to the base composition of the surrounding sequence forms part of this analysis. Another way to assess the ability of the MRS to predict MAR is to compare it with other *in silico* MAR prediction methods. However, this is not a trivial task, as the MRS does not predict MAR as such, it simply denotes sequence belonging to a MAR.

As described above, MAR have been implicated in control of gene expression. But the mechanisms through which this is achieved are less clear, and many varied means of gene regulation by MAR have been proposed. Whatever their precise role, it is likely that their position in relation to genes will give some clues as to how MAR act. Current evidence has indicated that MAR are found flanking genes and in intronic sequence, particularly the first intron. However, there may be a certain ascertainment bias, as sequence screened for MAR tends to come from intensively studied gene loci. Therefore, the aims served in investigating the relationship that the MRS has with genes are twofold. Firstly, we can assess if the MRS complies with our expectations of MAR incidence in relation to genes. Secondly, the characteristics of

the distribution of the MRS itself may provide insight into its specific functions.

The research carried out to fulfil these aims is presented in the next three chapters of this thesis. In chapter two, the distribution of the MRS in the complete genome of *C. elegans* is described. Extensive use of various sequence randomisation methods is used to show that the observed pattern of MRS incidence is different to what would be expect of a randomly occurring motif. The distribution of MRS along the chromosomes and its relative frequency in the various genic and non-genic portions of the genome implied a specific relationship with genes. Closer inspection revealed a distinctive spike in frequency of MRS in the regions immediately flanking *C. elegans* coding sequences (CDS).

The analysis of the MRS in *C. elegans* is extended in three directions in chapter three. Firstly, the effect of the complex nature of the bipartite MRS motif on its characteristics, such as size and spacing are investigated. Secondly, the MRS is compared with another MAR prediction tool, involving the transformation of the MRS towards MAR predictions. Lastly, an attempt is made to identify common characteristics of the genes contributing to the distinctive spike in frequency of MRS flanking CDS.

In chapter four the investigation of the MRS is broadened to encompass five additional genomes: *C. briggsae, D. rerio, D. melanogaster, A. thaliana* and *H. sapiens.* The primary focus was to study frequency of the MRS surrounding genes and make comparisons between the genomes of this diverse range of species. Efforts were also made to define the precise location of the MRS frequency spikes in relation to gene transcript and translation start and stop sites.

The final chapter forms a general discussion of the results presented across chapters two, three and four. The implications of the findings on the role of the MRS as a predictor of MAR prediction are considered, along with reflection on the specific functions of the MRS.

Chapter 2 - The Matrix Attachment Region Recognition Signature in the genome of *Caenorhabditis elegans*

2.1 Abstract

The MAR recognition signature (MRS) has been reported to be associated with a significant fraction of MAR in *C. elegans* and has also been found in MAR from a wide range of other eukaryotes. However the effectiveness of the MRS in specifically and sensitively identifying MAR remains unresolved. In an effort to clarify the role of the MRS, the incidence of MRS across the entire *C. elegans* genome was mapped. The MRS was found to have a distinctive chromosomal distribution, in which it appears more frequently in the gene-rich chromosome centres than in arms. Comparison to distributions of MRS estimated from chromosomal sequences randomised using mono-, di- tri- and tetra-nucleotide frequency patterns showed that, while MRS are less common in real sequence than would be expected from nucleotide content alone, they are more frequent than would be predicted from short-range nucleotide structure. In comparison to the rest of the genome, MRS frequency was elevated in 5' and 3' UTRs, even after accounting for AT content. Striking peaks of average MRS frequency were found to flank *C. elegans* coding sequence (CDS). Analysis of the genome of the closely related nematode, *C. briggsae*, revealed that it had a similar peak of average MRS frequency at the end of the CDS but not at the start. A degree of conservation of MRS between *C. elegans* and *C. briggsae* orthologs was observed. Due to their association with untranslated regions, it is possible that MRS could have a post-transcriptional role in the control of gene expression.

2.2 Introduction

The matrix attachment region (MAR) recognition signature (MRS) was conceived using 33 kb of *C. elegans* DNA [42]. The complete genome sequence of *C. elegans* is

now available and well annotated. This resource is made use of in this chapter, in which the incidence of the MRS in the entire nuclear genome of *C. elegans* is studied in an attempt to determine the validity if the MRS. There were a number of reasons for choosing *C.elegans* for this study, in addition to furthering the work of van Drunen *et al.* [42]. At about 100 Mb, the genome is relatively small and as two thirds of it is protein coding, there is a likely to be high concentration of functional non-coding DNA, compared to, for example, *H. sapiens*. The *C. elegans* genome is also well annotated, particularly with respect to the protein coding genes.

The only other published account of genome-wide prediction of MAR is the use of the SMARTest MAR prediction software with the genome of *A. thaliana* [58]. In this study a total of 21, 705 MAR were predicted, positioned uniformly along the 5 chromosomes. However, the 620 kb mitochondrial DNA insertion on *A. thaliana* chromosome 2 was found to be devoid of MAR predictions. About 8% of genes were found to overlap with MAR predictions, representing a 4-fold under-representation. Where MAR were found in genes, they were preferentially located in the first intron, with lower abundance in introns towards the 3' end and exons. Genes containing MAR were found to negatively correlate with transcriptional abundance, thus indicating MAR had the effect of down-regulating genes in *A. thaliana.* A follow-up study used the same SMARTest MAR predictions to investigate expression of genes containing MAR predictions further [73]. Using high resolution expression datasets MAR prediction containing genes were confirmed as being generally less expressed and were also shown to be more likely to be differentially expressed. A disproportionate number of transcription factor genes were found to harbour MAR predictions. Further, transcription factor genes with a MAR prediction had a greater degree of differential expression than those without. Differential expression was also found to be greater for genes in which the MAR predictions lay in introns compared to other gene regions.

The initial aim was to produce a comprehensive description of all the loci in the *C. elegans* genome in which the MRS resides. If MRS constitutes a feature with real

biological meaning then its distribution would be expected to be non-random with respect to other genome features. One way of measuring this is to compare the MRS distribution pattern with that of characteristics such as the disparity between *C. elegans* chromosome centres and arms. Another way is to compare the MRS frequency in various genic and non-genic portions of the genome with what would be expected of a randomly distributed pattern. The relative enrichment of MRS in various classes of genomic sequence can be used to determine if it complies with our expectations of where MAR lie in the genome.

The comparison of MRS incidence in randomised sequence with that found in real genomic sequence is also used here. If the pattern of MRS in real genomic sequence is shown to be different to that found in randomised sequence, then we can infer that its occurrence results from selection pressure on the genome. The alternative hypothesis is that the distribution of MRS is dictated purely by the base composition of the sequence in which it is found. For example, as an AT rich motif, the MRS may simply represent AT rich regions of the genome. Several methods of randomising the genome are employed in this chapter. The most simple is the randomisation of the order of bases in a section of genome sequence, referred to as mono-nucleotide randomisation. Various section lengths were used so that the effect of base composition variation over different scales could be studied. More complex randomisation protocols involving Markov chain processes were also used. In Markov chain randomisation the random sequence string is extended by adding new bases according to what the previous bases are. In this way the frequency of oligonucleotides in the real genomic sequence can be replicated in the Markov chain randomised sequence. This is important in genome sequence randomisation, for example to account for the triplet pattern of amino acid codons. In this study first, second and third order Markov chains processes were used, allowing di-, tri- and tetra- nucleotide patterns to be taken into account.

Having established a relationship between MRS and genes, a secondary aim of this chapter was to clarify the positioning of the MRS in the vicinity of genes. By

obtaining information about where the MRS were sited, insight into potential functions could be gained. To facilitate the study of MRS frequency in the regions immediately surrounding genes, the concept of aligning genes based on their start and stop positions was introduced. Similar techniques have previously been used to study nucleotide frequency variation among genes and conserved sequences [74-77]. This technique allows general patterns of average MRS frequency to be identified that would be difficult to identify by studying single genes.

The final aim of this chapter was to make an initial investigation of the level of evolutionary conservation of MRS. Conservation of individual MRS between species would indicate selective pressure to maintain those sequences, and thus be strong evidence of a specific functional role for MRS. If the MRS is under less constraint it may still be possible to identify common frequency patterns between species that may point towards some kind of functionality for the MRS. For a preliminary study of the conservation of the MRS, the closely related nematode, *C. briggsae* is ideal. The two species are thought to have diverged about 100 million years ago and share many characteristics, such as genome size, chromosome number and ecological niche. In addition, the genome sequence of *C. briggsae* has been completed and good quality gene predictions are available.

2.3 Methods

2.3.1 MRSfinder

The identification of MRS on a genome-wide scale was automated through the use of a custom Perl program, MRSfinder. Using the description of the MRS given by van Drunen *et al*. [42], MRSfinder locates all occurrences of the MRS in a given sequence in either orientation and reports their start and stop positions. The program is freely available, along with the coordinates of all the MRS found in the genome of *C. elegans*[78].

2.3.2 Genome sequence data

Version WS150 of the *C. elegans* genome was downloaded from the WormBase ftp site [79]. The associated gene annotation for WS150 was downloaded using WormMart [80] and additional annotation was downloaded from the WormBase genome browser [81].

Version cb25 of the *C. briggsae* genome was downloaded from WormBase ftp site [82]. This version is assembled into 578 contigs. The associated annotation was downloaded using WormMart [80].

2.3.3 MRS and gene distribution in 2 Mb windows

Each chromosome was divided into consecutive, non-overlapping 2 Mb windows, with the first window starting at chromosome base position 1. Where the final window did not contain 2 Mb, the counts for that window were scaled proportionally. For each window, the number of MRS (from MRSfinder) and gene start positions (from WormBase) were assessed. Where a gene was annotated as having more than one transcript or gene model, one transcript and model was randomly selected.

2.3.4 Mononucleotide randomisation of the genome sequence in variety of window sizes

For randomisation of sequences >= 32,000 bp, a roulette wheel selection algorithm was used where a nucleotide's chance of selection was based on its frequency in the original sequence. Due to the stochastic nature of this randomisation method the nucleotide frequency was verified to ensure it fell within 0.2% of that found in the original sequence.

For sequences <32,000 bp, the sequence was randomised using a Fisher-Yates shuffle. Each sequence was randomised 1000 times. Each chromosomal sequence was split into consecutive, non-overlapping windows of the appropriate length with correction for shorter end windows as above. Following randomisation, MRSfinder was used to identify all the MRS in the randomised sequence. The mean and standard deviation of the MRS counts for each randomised version of the sequence were

calculated. The validity of the randomisation process was checked using a measure of compositional heterogeneity, implemented in the program gc_index [69]. GC_index calculates the average difference in GC content between two adjacent windows normalised by the standard error expected under the assumption of random distribution of nucleotides in a window. Random sequence should have low variability between windows and therefore a low compositional heterogeneity index.

2.3.5 Randomisation of the genome using Markov chain processes

First, second and third order Markov chain processes were used to randomise the genome sequence following the algorithm of Workman and Krogh [83]. In a first order Markov chain process, the first nucleotide is chosen by sampling from the mono-nucleotide frequency. Subsequent nucleotides are added by sampling the probability distribution derived from the frequency of the four di-nucleotides that start with the previous nucleotide. Second and third order Markov chain process randomisation was carried out in a similar fashion, using tri- and tetra-nucleotide frequencies respectively.

2.3.6 Number of MRS in genome features

Genes, introns, exons, 3' UTR and 5' UTR were identified based on the GFF file for the appropriate *C. elegans* chromosome. Intergenic regions were defined as all sections of DNA not annotated as belonging to a gene. Where two or more incidents of a single feature type overlap, they were joined to form a single incident of that feature. The genomic coordinates of each feature were used to identify MRS that lay wholly within and partially overlapping a unit of that feature.

The number of MRS expected to lie wholly within each feature type (i.e. complete overlap) was calculated using the formula:

$$M(F((f-m)+1))/c$$

The expected number of MRS expected to partially overlap a feature:

$$M(F(2(m-w)))/c$$

When the average size of the MRS exceeds that of the feature, a complete overlap is defined as a feature lying wholly within an MRS. The expected number was calculated using the formula:

$$M(F(m-f)+1))/c$$

The expected number of partial overlaps when the average size of the MRS exceeds that of the feature:

$$M(F(2(f-w)))/c$$

where M = number of MRS, F = number features of specific type, f = average length of feature, m = average length of MRS, w = minimum number of nucleotides required for a partial overlap and c = total sequence length.

Three different scoring methods were used to combine the number of partial and complete overlaps to give an overall score. In method 1 complete overlaps = 1 point, partial overlaps = 0 points, method 2 complete overlaps = 1 point, partial overlaps = 1 point, method 3 complete overlaps = 1 point, partial overlaps = 1/2 point. In all scoring methods, the minimum number of nucleotides required for a partial overlap was 12. The AT content correction factor was calculated based on the ratio of the number of MRS found in random sequence with the same AT content as each feature to the number of MRS found in random sequence with the same AT content as the genome. The number of MRS found in random sequence of specific AT content is shown in Figure 2.8.

2.3.7 Number of MRS by GC%

The *C. elegans* genome sequence was split into non-overlapping sections of 1000 bp. For each section the average GC%, the number of MRS mid-points and the genic class, (e.g. gene, intergenic) according to the WS150 genome annotation was recorded. Simulated data was generated by creating 1224 bp (to account for the size of the MRS) of random sequence (mononucleotide) with a specified GC% and counting the number of MRS mid-points that lay in the central 1000 bp. This was

repeated 1000 times for GC % ranging from 10 – 50%. Investigations using longer random sequence lengths (up to 2 Mb) showed that the number of MRS/ bp for each specific GC value was unaffected by choice of sequence length.

2.3.8 AT and MRS frequency across CDS

In this analysis, one CDS per gene was used: where a gene was annotated with multiple transcripts and/or gene models, a single transcript/model was randomly selected to represent the gene. The CDS were then subjected to quality filters to remove poor quality sequence (containing Ns), CDS with insufficient sequence upstream or downstream and CDS that did not start with ATG or end with a stop codon. Of the 20,052 *C. elegans* CDS originally identified, 20,032 passed these filters. The 19,528 *C. briggsae* CDS were reduced to 12,954 after filtering. Each successfully filtered CDS was then split into consecutive, non-overlapping 50 bp windows, starting 1000 bp upstream of the CDS start site and continuing to 1000 bp downstream of the CDS stop site. The total number of MRS mid-points occurring in each window across all CDS was divided by the number of CDS used to produce a frequency of MRS occurrence in that window.

For AT analysis, the CDS sequences were split into consecutive, non-overlapping 50 and 10 bp windows. For each window the AT content was calculated as a percentage of the window length. The mean AT% for each position across all CDS was calculated.

2.3.9 MRS in *C. briggsae* orthologs

The cb25 version of the *C. briggsae* genome sequence and annotated orthologs to *C. elegans* were downloaded from WormBase. After subjecting the 11,953 orthologs to filtering for length (i.e. sufficient sequence upstream and downstream for further analysis), poor quality (sequence containing Ns), and CDS not starting with ATG or ending in a stop codon, 4,132 genes remained. MRSfinder was used to detect MRS within 200 bp of the CDS stop for each of these filtered genes in *C. elegans* and *C. briggsae*. Association between a *C. elegans* gene having an MRS and the *C. briggsae*

ortholog having an MRS was tested using the log odds ratio (a × d)/(b × c) where a is the number of orthologs with an MRS within 200 bp of the CDS stop codon in *C. elegans* and *C. briggsae*, b is the number of orthologs where an MRS is only found within 200 bp of the CDS stop codon in *C. briggsae*, c is the number of orthologs where an MRS is only found within 200 bp of the CDS stop codon in *C. elegans* and d is the number of orthologs where neither organism has an MRS within 200 bp of the CDS stop codon.

2.4 Results

2.4.1 MRS distribution in C. elegans

The MRS is a degenerate bipartite motif consisting of a 16 bp pattern, AWWRTAANNWWGNNNC (where W = A or T, R = A or G, N = A,C,G or T), within which one mismatch is allowed, and an 8 bp pattern, AATAAYAA (where Y = C or T) [42]. To be scored as an MRS, both these sequences must lie within 200 bp of each other, although they may overlap and they may be on either strand of the DNA duplex [42]. Existing MRS finding programs were designed to under-report closely apposed MRS [84]. To allow full control over data reported, a custom program, MRSfinder, was designed. MRSfinder was used to map the location of MRS across the entire *C.elegans* genome. MRS were found across all 6 *C.elegans* chromosomes at an average frequency of 249 per Mb. This is similar to the frequency of genes, which is 228 per Mb. At small scales (<500 kb), the motif distribution was noisy (Figure 2.1 and Figure 2.2). As would be expected of an AT-rich motif, there was some correlation with regions of high AT% (see below).

However, at a chromosomal level distinct patterns emerged. Analyses of non-overlapping 2 Mb windows along the chromosomes showed that MRS were significantly more abundant in the centres than in the arms of all chromosomes except chromosome IV (Figure 2.3 and Table 2.1). The division between chromosome arms and centres is characteristic of several genomic features in *C. elegans*. Centres tend to be gene rich, with a high concentration of essential, well

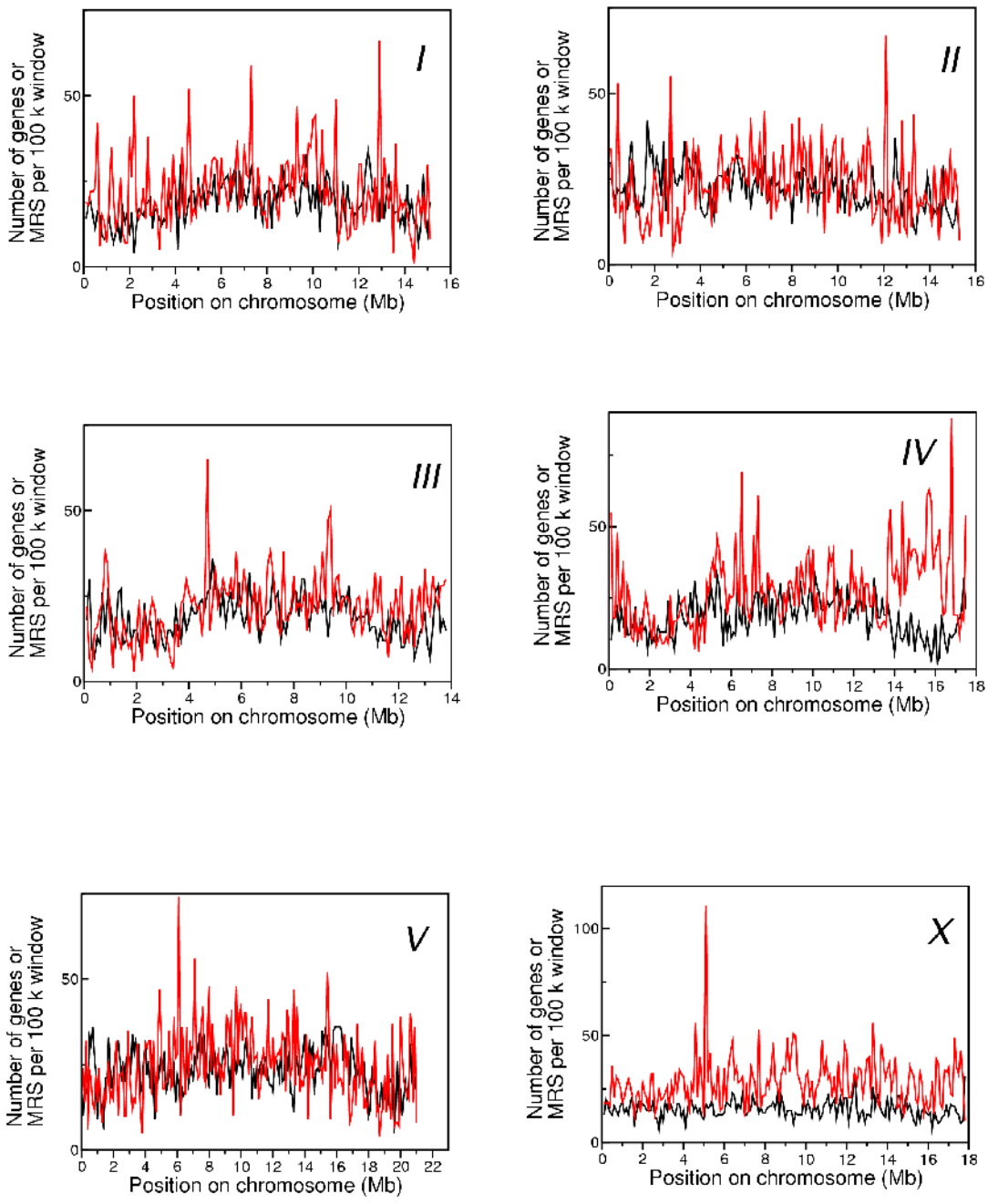


Figure 2.1 Distribution of genes and MRS in C.elegans chromosomes at window sizes of 100 kb

Number of gene (black) and MRS (red) start positions in non-overlapping 100 kb windows. To account for short sequence length in the end window, the number of genes and MRS in the last window was scaled.

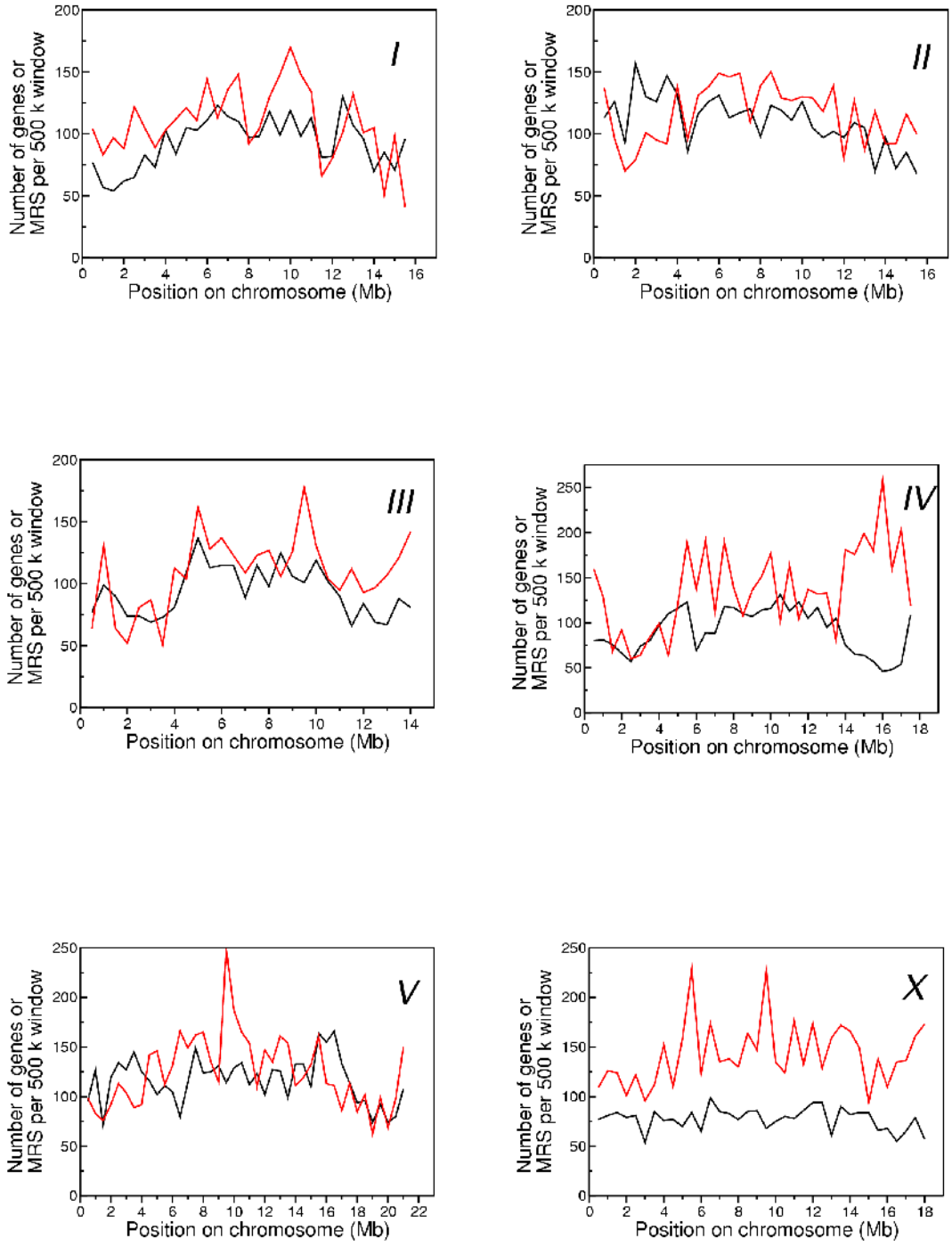


Figure 2.2 Distribution of genes and MRS in C.elegans chromosomes at window sizes of 500 kb

Number of gene (black) and MRS (red) start positions in non-overlapping 500 kb windows. To account for short sequence length in the end window, the number of genes and MRS in the last window was scaled.

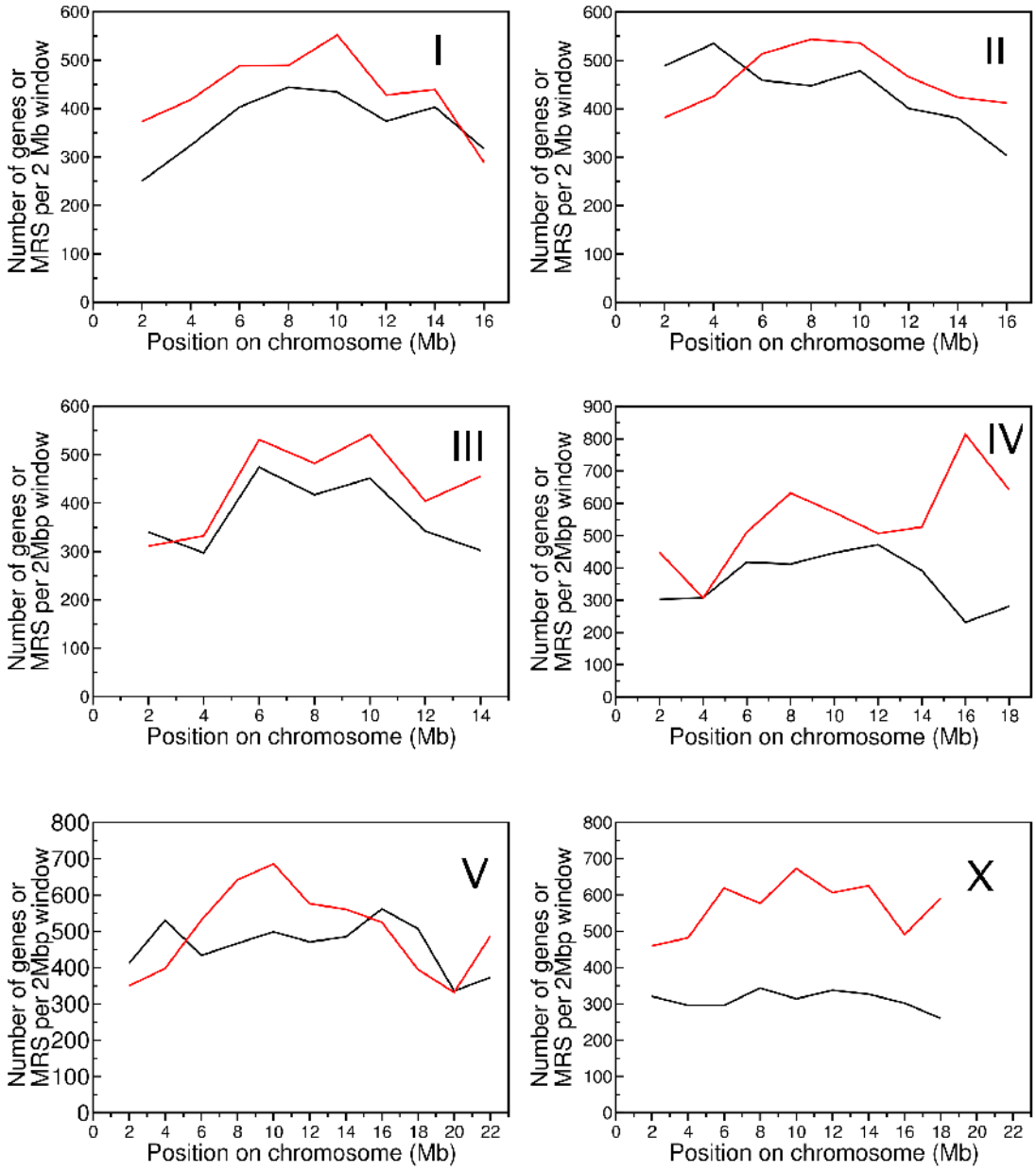


Figure 2.3 Distribution of genes and MRS in C. elegans chromosomes, 2 Mb window

Number of gene (black) and MRS (red) start positions in non-overlapping 2 Mb windows. To account for short sequence length in the end window, the number of genes and MRS in the last window has been scaled to 2 Mb.

Chromosome	I	II	III	IV	V	X
correlation co-efficient	0.89	0.96	0.67	0.05	0.75	0.72
p-value	0.001	<0.0001	0.049	0.443	0.004	0.015

Table 2.1 Correlation between MRS frequency and distance to centre of chromosome

Each 2 Mb chromosome window was given a number based on its distance from the centre of the chromosome. The windows at the far ends chromosome were assigned 1, the next windows towards the chromosome centre were assigned 2 and so on until all windows had been assigned a number. The correlation between the MRS frequency in each window and its number was then calculated using Pearson's r correlation coefficient.

conserved and highly expressed genes [3, 67]. By comparison, the chromosome arms exhibit a higher meiotic recombination rate, and are enriched for transposons and repeats [67]. Thus, at the chromosome level, MRS are more likely to be found in the vicinity of highly expressed and essential genes.

2.4.2 MRS distribution in randomised sequence

Although the distribution of MRS appeared to correlate broadly with several other genome features, the specific nucleotide composition of each sequence window will influence the number of MRS. By randomising the genome sequence whilst maintaining nucleotide composition (mononucleotide randomisation), the number of MRS expected in the sequence due to nucleotide composition alone was estimated. Additional randomisation models were used in order to account for relationships between adjacent bases. The mononucleotide randomisation model generated sequence in which the frequency of each of the four nucleotides matched that observed in the chromosomal sequence. More complex first, second and third order Markov chain randomisation processes reflected the di-, tri- and tetra-nucleotide content of the chromosomal sequence. For each 2 Mb non-overlapping window used in Figure 2.3, the nucleotide sequence was randomised 1000 times, and MRSfinder was used to map and count the number of MRS in each randomised sequence. A comparison of MRS counts for chromosome I under each randomisation process is shown in Figure 2.4, results for second order Markov chain randomisation of the other chromosomes can be found in Figure 2.5. The observed number of MRS in mononucleotide randomised sequence was similar to that found in real sequence, while the first, second and third order Markov chain randomised sequence yielded far fewer MRS. As MRS occurrence was best modelled by the mononucleotide randomisation process, subsequent analyses focussed on this method of randomisation.

Figure 2.6 shows the difference in observed MRS count for each 2 Mb window from the mean count in the mononucleotide randomised sequences, in terms of standard deviations from the mean. Throughout the length of each chromosome, the number

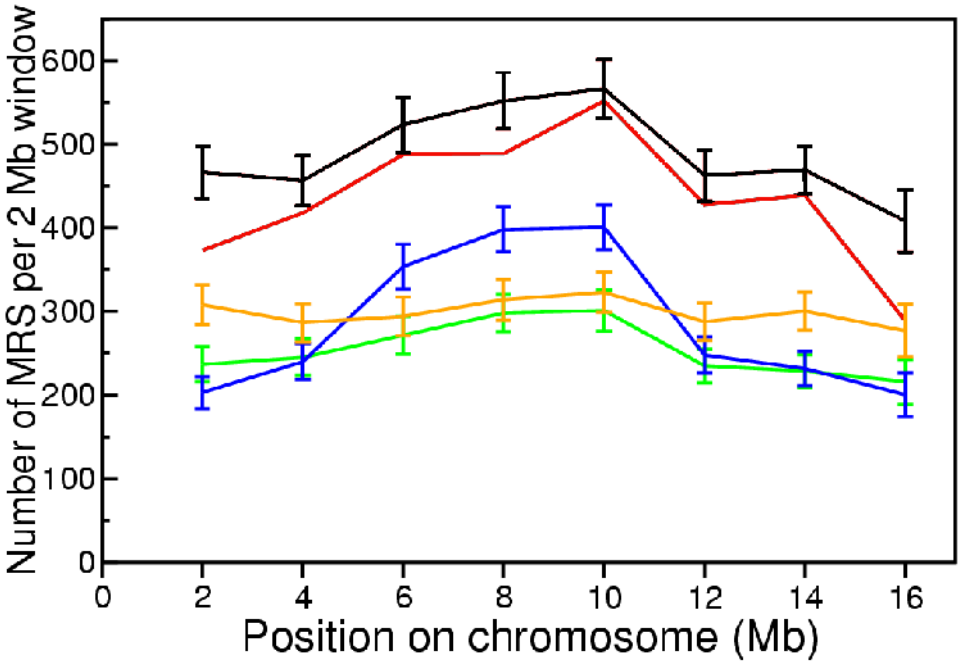


Figure 2.4 Comparison of MRS distribution in C. elegans chromosome I under various randomisations

The number of MRS in non-overlapping 2 Mb windows in real *C. elegans* chromosome I sequence is shown in red. The chromosome was randomised in non-overlapping 2 Mb sections using four different Markov chain processes. The average number of MRS +/- one standard deviation for the 2 Mb windows for zero (mononucleotide, black), first (orange), second (green) and third (blue) order Markov chain process randomisation is shown.

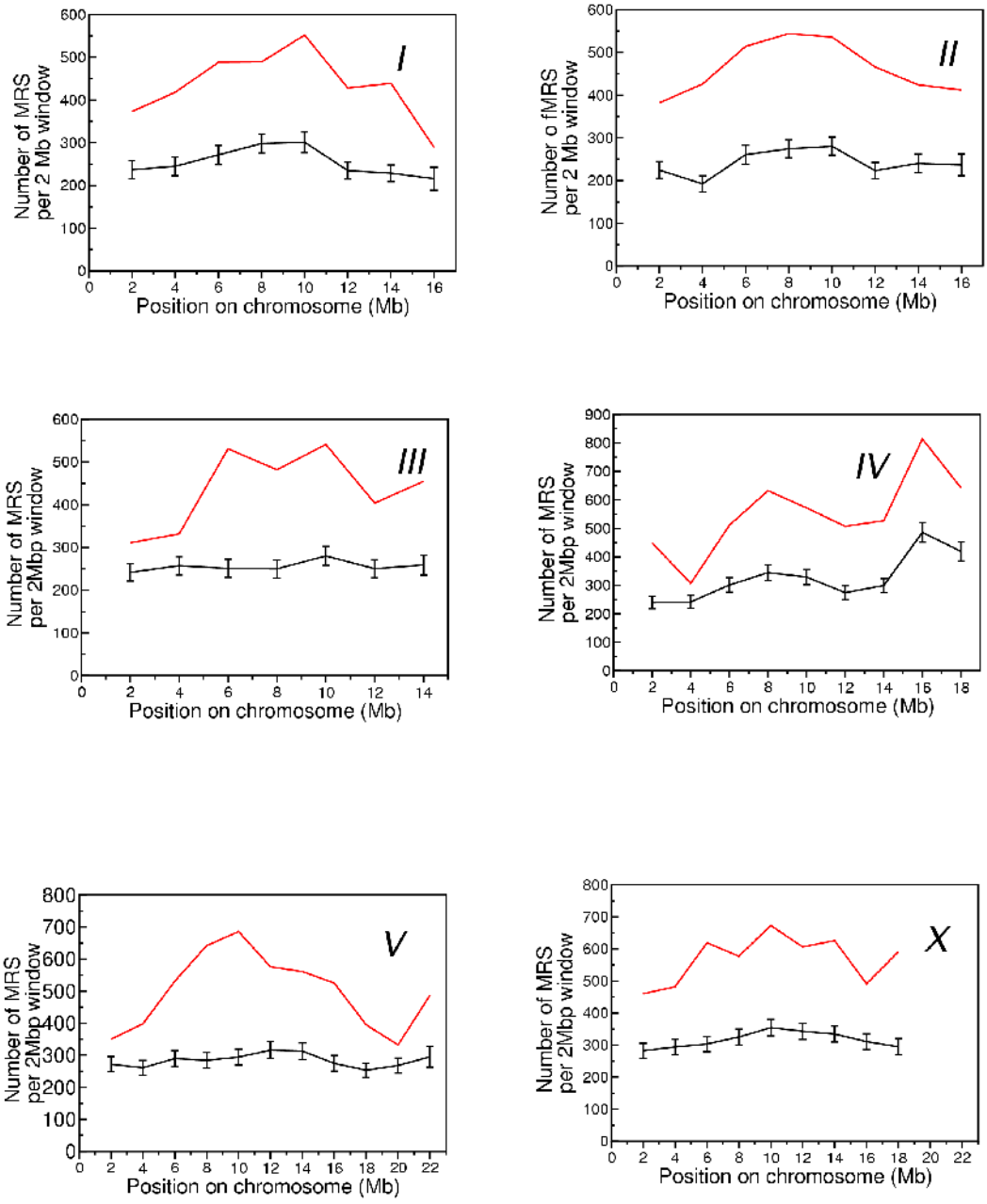


Figure 2.5 MRS in second order Markov chain randomised chromosome I, II, III, IV, V and X.

The chromosomes were randomised in non-overlapping 2 Mb windows using a second order Markov chain process. The average number of MRS over 1000 randomisations (+/- one standard deviation) in the 2 Mb windows (black) is compared with the number of MRS in real sequence (red).

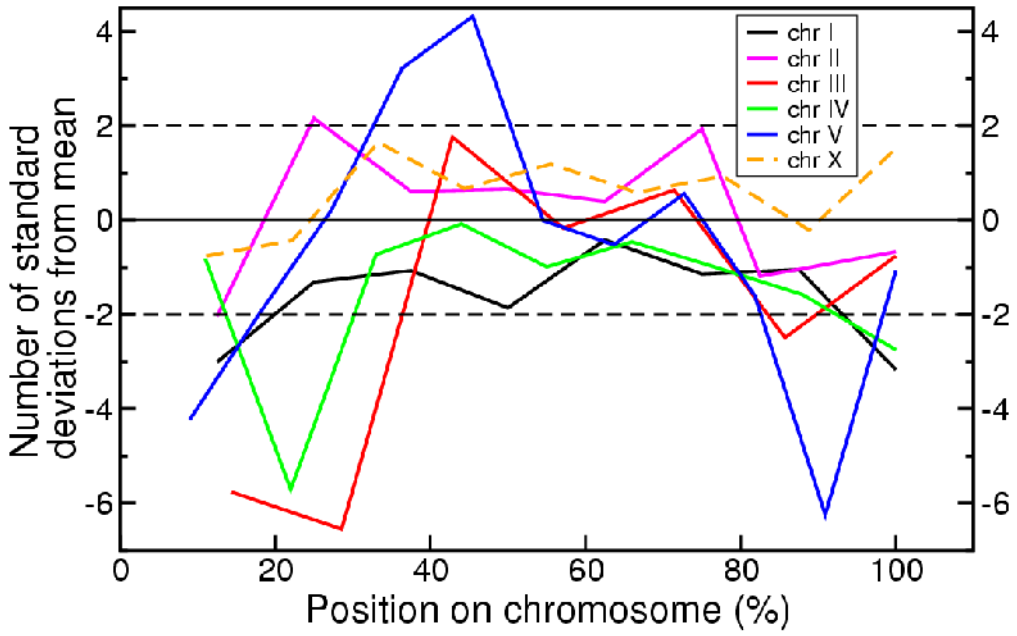


Figure 2.6 Distribution of MRS along C. elegans chromosomes, relative to average number of MRS in chromosome sequence randomised in 2 Mb sections

The sequence of each chromosome was randomised using a mononucleotide process in non-overlapping sections of 2 Mb, MRS were then mapped in this sequence using MRSfinder. This was repeated 1000 times and the average and standard deviation of MRS frequency in the 2 Mb sections was obtained. This graph shows the distribution of MRS in actual *C. elegans* sequence, as the number of standard deviations from the mean MRS frequency in the randomised sequence.

of MRS in real sequence was generally lower than in the mononucleotide randomised sequence. The arms were particularly poor in MRS and the chromosome centres were at most only slightly enriched for MRS. In contrast to the autosomes, the distribution of MRS along chromosome X (Figure 2.6, broken line) was much more even and similar to that found in mononucleotide randomised chromosome X sequence.

One effect of randomising the genome sequence in relatively large sections of 2 Mb is that nucleotide content (or nucleotide local pattern) becomes more uniform across each section, eliminating, for example, local peaks of very high AT%. To identify the effects of local areas of extreme nucleotide composition, mononucleotide randomisation was applied to smaller sections of sequence (10 bp, 100 bp, 1 kb, 50 kb, 2 Mb and the whole chromosome length) in *C. elegans* chromosome I. The number of MRS found in the whole chromosome under each mononucleotide randomisation regime, averaged over 1000 iterations, is shown in Figure 2.7. The numbers of MRS found when the chromosome was randomised along its entire length in one section and in 50 kb sections were very similar to the 2 Mb randomised sequence (about 10% higher than in the actual sequence). However, at randomisation sections of less than 50 kb the total number of MRS found rose dramatically. A similar effect was observed in the second order Markov chain process randomised sequence (data not shown). Compared to actual genomic sequence, the average number of MRS observed in mononucleotide randomised sequence doubled when the chromosome was randomised in sections of 10 bp.

2.4.3 MRS distribution in relation to genic classification

The above results show that the number and distribution of MRS in the *C. elegans* genome is distinct from that found in random sequence. To investigate how this distribution is related to other genome features, the degree of overlap between MRS and different functional parts of the genome was assessed. The number of MRS occupying the same genome space as exons, introns, 3' untranslated regions (UTR), 5' UTR, genes and intergenic regions, is given in Table 2.2. The expected score

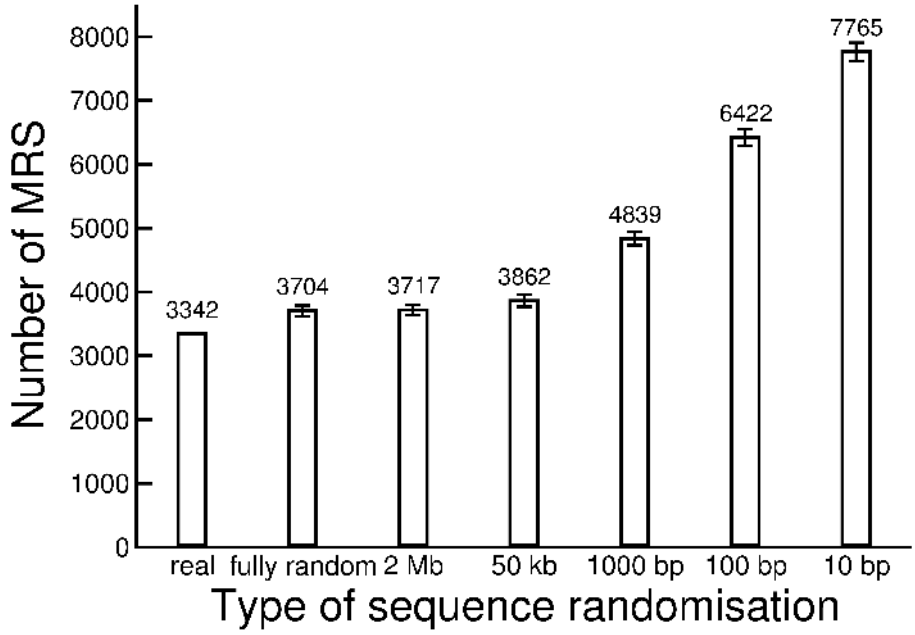


Figure 2.7 Frequency of MRS in C. elegans chromosome I randomised in various section lengths

Chromosome I was randomised using a mononucleotide process in non-overlapping sections of various lengths 10 bp, 100 bp, 1 kb, 50 kb, 2 Mb and the entire length of the chromosome, and MRSfinder used to identify MRS in each sequence. The randomisation and MRS mapping was repeated 1000 times for each section length. The bar height shows the average number of MRS in the chromosome and the error bars represent +/- 1 standard deviation. The actual number of MRS in *C. elegans* chromosome I is shown for comparison.

	Genes	Exons	Introns	5' UTR	3'UTR	Intergenic
Size of feature (kb)	58,735	25,497	30,587	457	1,616	41,741
Number of features in genome	18,719	124,049	100,853	8,293	9,103	18,832
Feature AT%	63	57	68	60	68	66
Actual number of MRS in feature	11,368	1,955	7,094	139	691	12,683
Expected number of MRS in feature	14,303	4,218	5,883	33	246	10,070
Ratio (actual/expected) (score system 1)	0.79	0.46	1.21	4.22	2.81	1.26
AT% corrected ratio (score system 1)	1.05	1.66	0.74	8.80	1.71	1.02
AT% corrected ratio (score system 2)	1.09	1.83	0.64	2.08	1.34	1.03
AT% corrected ratio (score system 3)	1.07	1.77	0.68	2.98	1.46	1.02
AT correction factor	0.76	0.28	1.64	0.48	1.64	1.24

Table 2.2 Number of MRS in genic and non-genic portions of the genome.

The number of MRS overlapping genes, exons, introns, 3' UTR, 5' UTR and intergenic regions of the genome was used to calculate an overlap score as described in Methods. The expected overlap score was calculated assuming a uniform distribution of MRS across the genome, using the formulae described in Methods. The ratio of the actual to expected score is shown. The expected number of MRS was multiplied by the AT correction factor (see Methods) and the ratio re-calculated to give the AT corrected ratio. For details on score system, see Methods.

indicates how many MRS would be expected to lie in a feature, based on the total size of the feature and assuming a uniform distribution of MRS across the genome.

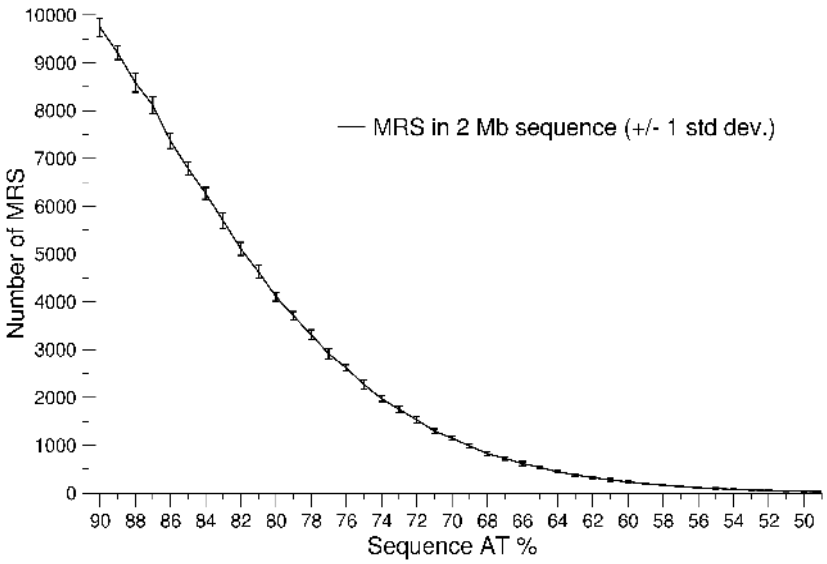
The ratios of actual and expected MRS numbers showed large differences in MRS abundance in each of the genome features. MRS were particularly rare in exons, which contained less than half the MRS expected. As a result, the number of MRS in genes was also lower than expected, despite enrichment for MRS in introns and untranslated regions. Intergenic regions had slightly more MRS than expected. However, the 5' UTR and 3' UTR were by far the most MRS-enriched parts of the genome, by factors of 4.2 and 2.8 respectively. The relative enrichment of introns, 5' UTR and 3' UTR for MRS provides an explanation for the spatial relationship between genes and MRS described in Figure 2.3.

The MRS is AT rich and so is more likely to occur in AT rich sequence. The extent to which the AT content of the sequence can influence the frequency of MRS is demonstrated in Figure 2.8. Random sequence 2Mb long was generated (using a mono-nucleotide randomisation process) with specific AT content, ranging from 50 to 90%. In sequence with 90% AT, there was nearly 10,000 MRS, an average of 1 every 20 bp. The MRS density drops exponentially with decreasing AT content. At 50% AT, there are virtually no MRS in the sequence.

To control for this bias, an AT-correction factor was used to adjust the expected number of MRS. The correction factor was based on the number of MRS found in mononucleotide random sequence with AT content equivalent to that of each feature, as a proportion of the number of MRS found in random sequence with AT content equivalent to that of the whole genome. When this correction is applied, the AT-poor exons appeared enriched for MRS, while the AT-rich introns had fewer than expected. Both genes and intergenic regions had approximately the number of MRS expected.

However, even with AT correction, the untranslated regions, particularly the 5' UTR, showed strong enrichment for MRS. Alternative overlap scoring systems that take into account partial MRS-feature overlaps did not affect these results. Although UTR

A



B

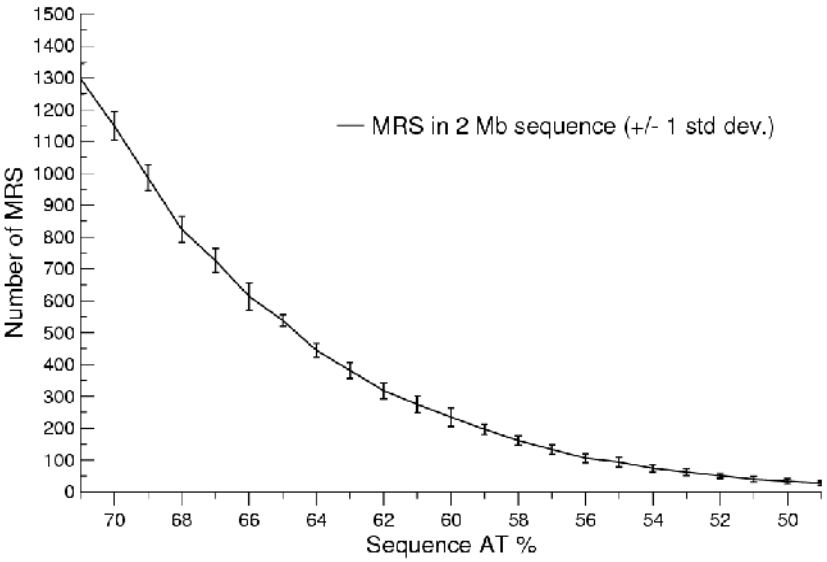


Figure 2.8 Number of MRS in random sequence of defined AT content

The number of MRS in 2 Mb of random sequence with AT content ranging from 90% to 50% (A) and 70% to 50% (B) was calculated. Random sequence for each AT value was generated 1000 times, error bars show +/- 1 standard deviation.

form only a small part of the genome and contain only a small proportion of the total MRS, the degree of MRS enrichment and their proximity to genes points to a functional role for MRS.

Above, the number of MRS in a sequence is shown to be influenced not only by the nucleotide content of the sequence but also its functional state (e.g. genic, UTR, exonic etc.).That analysis is extended here by comparing the number of MRS in sequence of specific GC content from genes, intergenic regions and all genomic sequence. For each of these sequence categories, the mean number of MRS for 1000 bp of sequence with a specific GC content is shown in Figure 2.9. For comparison, the profile of MRS frequency in (mononucleotide) randomised sequence over the same GC content range is included (adapted from figure 2.8). All genomic sequence for each GC graduation was used. In some cases, particularly at the upper and lower end of the GC range, the number of instances was very low.

Genomic sequence has a lower frequency of MRS at low GC levels (<36%) than observed in randomised sequence and slightly more MRS at higher GC levels (>36%). There was a slightly higher frequency of MRS in intergenic sequence than in genic sequence at all GC levels. There was a big spike in MRS frequency in intergenic sequence of about 47% GC. This may be functionally significant or be the result of a random fluctuation in MRS frequency in a few windows, exaggerated by the low number of windows at that GC level. In broad terms, compared to randomised sequence, genomic sequence has a lower frequency of MRS at low GC levels and slightly higher frequency at higher GC levels.

2.4.4 MRS frequency surrounding genes

To clarify the relationship between genes, especially their 5' and 3' UTRs and MRS, the frequency of MRS in the regions surrounding gene boundaries was investigated. Using the data from MRSfinder, MRS locations were plotted on a section of sequence extending 1000 bp upstream of the translation start site (ATG codon) through the first 400 bp of the coding sequence (CDS) from each *C. elegans* gene.

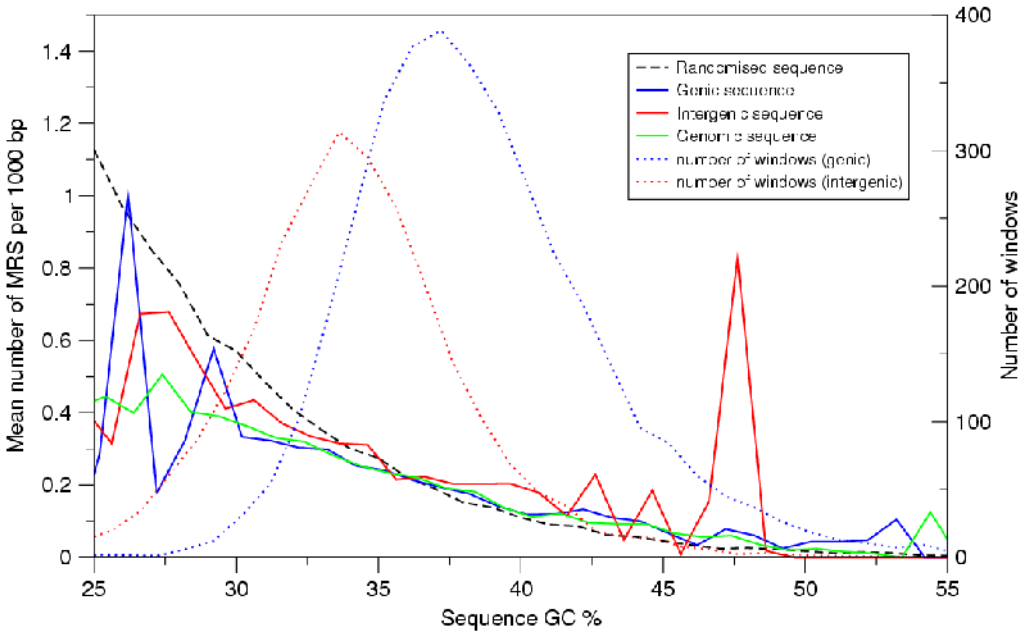


Figure 2.9 Number of MRS for genome sequence of specific GC content

The lines indicate the number of MRS in 1000 bp sections of *C. elegans* genomic sequence, ordered by GC content.

The same analysis was carried out on sequence from the last 400 bp of the CDS through to 1000 bp downstream of the stop codon (Figure 2.10).

As expected from the overlap of MRS with genes and intergenic regions reported in Table 2.2, the frequency of MRS in regions outside the CDS was higher than in the CDS itself. The enrichment of MRS in the 5' and 3' UTRs shown in Table 2.2 correlates with striking increases in MRS frequency in the regions immediately flanking genes. The MRS frequency sharply rose and fell over a span of 350 bp, peaking 50–100 bp upstream of the CDS start. At the 3' end of the CDS the MRS frequency spike had an even greater amplitude, increasing by more than 3 fold in 200 bp.

One explanation for the MRS spikes bounding CDS is that they are related to AT content of these areas. For example, in the case of 3' UTR the apparent over-representation of MRS was reduced when AT content was taken into account (Table 2.2). Plotting AT content in the region surrounding CDS revealed a pattern of sharp spikes similar to that observed for MRS frequency (Figure 2.10). However, on closer inspection there were subtle differences between the MRS frequency and AT content variation. Firstly, the upstream AT peak occurred in the 50 bp immediately preceding the start codon, 50–100 bp after the MRS peak. Similarly at the downstream end, the AT peak occurred in the 50 bp immediately following the stop codon, again 50–100 bp separate from the MRS peak.

Another difference was that the AT content dropped to 58% in the first 50 bp of the CDS, then rose to about 62% for the middle part of the CDS. The pattern was similar at the end of the CDS, where the AT dropped to near 58% in the last 50–100 bp. In both locations this AT dip was not matched by a dip in the MRS frequency. The variation in AT content in the vicinity of gene boundaries is an intriguing observation. Similar patterns have been described previously [74, 75, 77].

An analysis of the MRS frequency surrounding gene boundaries was also performed on a related nematode, *Caenorhabditis briggsae* (Figure 2.11). As in *C. elegans*, the frequency of MRS was higher in *C. briggsae* intergenic regions near genes than in

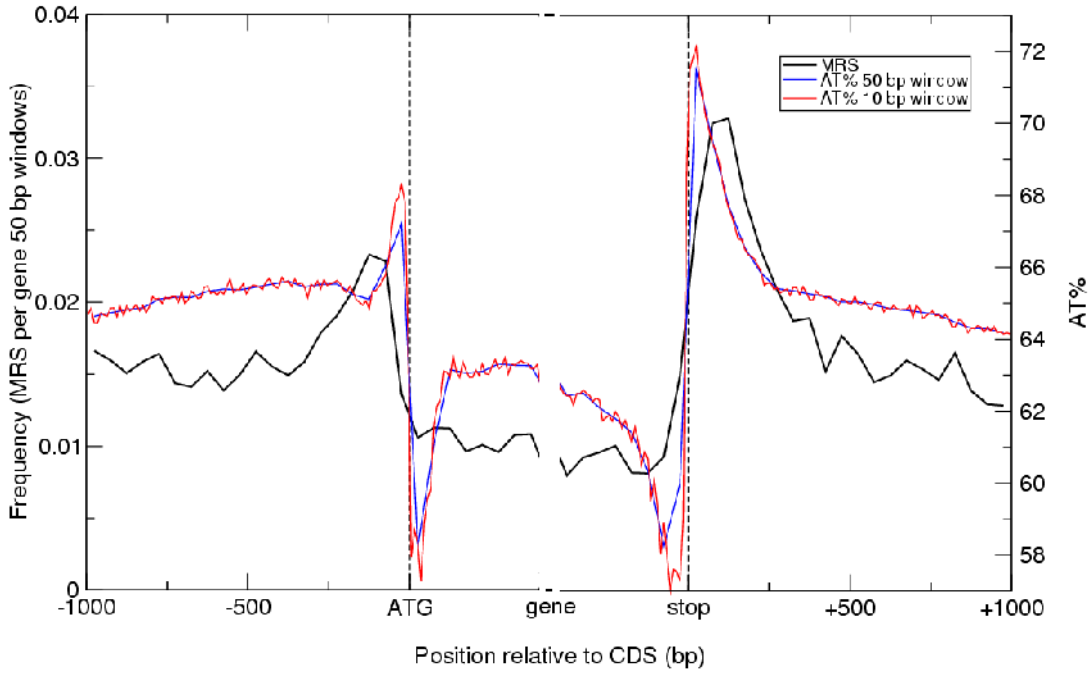


Figure 2.10 MRS distribution and AT content near genes in C. elegans

Average AT% in 50 bp (blue line) and 10 bp (red line) non-overlapping windows and number of MRS per CDS in 50 bp non-overlapping windows (black line) is displayed. The windows extend from 1000 bp upstream of the translation start site (ATG codon) through the first 400 bp of the CDS and from the last 400 bp of the CDS, through to 1000 bp downstream of the translation stop site (stop codon).

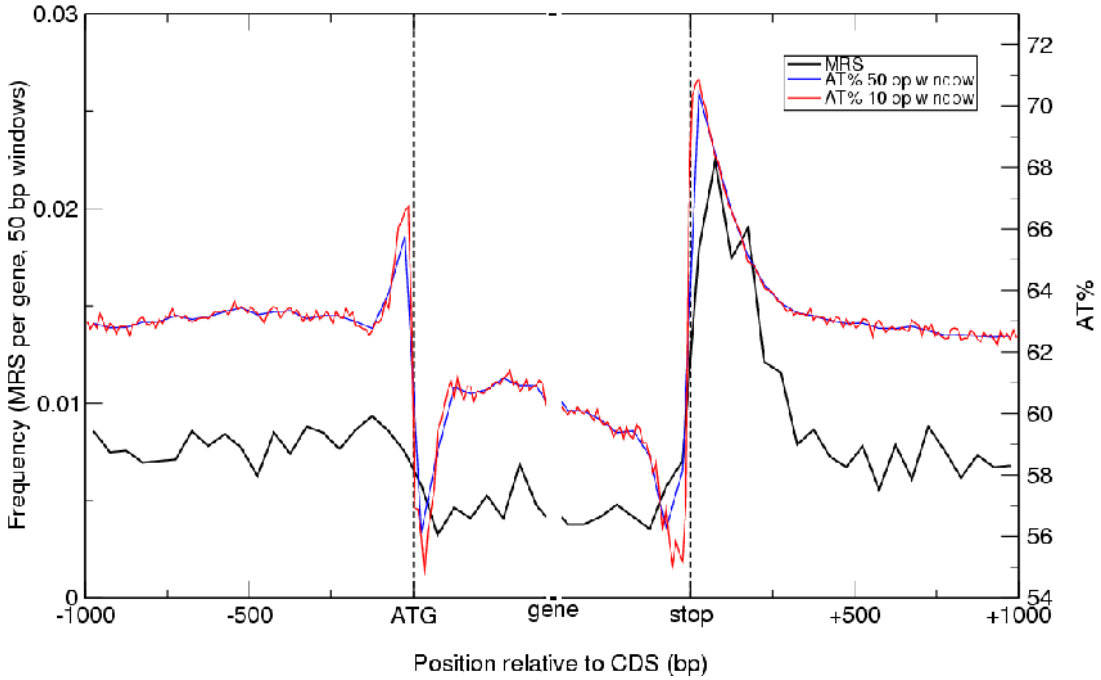


Figure 2.11 MRS distribution and AT content near genes in C. briggsae

Average AT% in 50 bp (blue line) and 10 bp (red line) non-overlapping windows and number of MRS per CDS in 50 bp non-overlapping windows (black line) is displayed. The windows extend from 1000 bp upstream of the translation start site (ATG codon) through the first 400 bp of the CDS and from the last 400 bp of the CDS, through to 1000 bp downstream of the translation stop site (stop codon).

CDS. However, from 1 kb upstream to 1 kb downstream of the CDS, the frequency of MRS was generally lower in *C. briggsae* than in *C. elegans*. The main difference in the pattern of MRS frequency between the species was that while *C. briggsae* displayed the same striking increase in average MRS frequency at the 3' end of the CDS, it lacked any increase in frequency at the 5' end. The possibility that less robust gene annotation in *C. briggsae* could have lead to this discrepancy was addressed by filtering the dataset to ensure all CDS started with ATG and ended with a stop codon, and that the selected sequence was complete and of high quality (i.e. no Ns). However, the possibility that the *C. briggsae* gene set is systematically lacking upstream exons cannot be excluded.

The difference between MRS frequency and AT content is even more marked in *C. briggsae* than in *C. elegans*. Although *C. briggsae* lacked an upstream MRS peak, an increase in AT content from about 63% to 66% was evident in the 50 bp immediately preceding the CDS start. In common with *C. elegans*, the downstream AT peak occurred 50 bp before the MRS peak and the AT dip at the start and end of the CDS was not matched by a dip in MRS frequency.

The distinctive increase in MRS frequency at the downstream end of both *C. elegans* and *C. briggsae* CDS could be due to conservation of MRS in specific genes, or simply a reflection of a general tendency. To investigate this, the occurrence of MRS within 200 bp of the CDS stop codon in *C. elegans* genes was compared to MRS occurrence in the same region of the corresponding *C. briggsae* ortholog (Table 2.3). Surprisingly, of the 224 *C. briggsae* genes annotated as orthologs of *C. elegans* genes with an MRS within 200 bp of the CDS stop codon, only 18 had an MRS in a similar position. Nonetheless, a small but significant degree of correlation between *C. elegans* genes and their *C. briggsae* orthologs for the presence or absence of MRS was detected (log odds ratio = 0.641, p value = 0.006). Therefore, the peak of average MRS frequency at the downstream end of *C. elegans* and *C. briggsae* CDS was due partly to apparent conservation of MRS in specific genes.

		***C. elegans* genes in ortholog set**	
		MRS within 200 bp of CDS stop	No MRS within 200 bp of CDS stop
***C. briggsae* genes in ortholog set**	MRS within 200 bp of CDS stop	18	172
	No MRS within 200 bp of CDS stop	206	3,736

Table 2.3 MRS within 200 bp downstream of translation stop sites of C. briggsae orthologs of *C. elegans* genes.

The filtered set of *C. elegans* and *C. briggsae* orthologs were assessed to identify the number of genes in the set from each organism that had an MRS within 200 bp of the CDS stop codon. The association between orthologs for the presence or absence of 3' MRS was significant (log odds 0.641, p value = 0.006).

2.5 Discussion

In describing and analysing MRS frequency in the genome of *C. elegans*, these sites have been shown to have a specific distribution, particularly in relation to genes. These observations support the validity of the MRS as a real genomic feature, though not necessarily indicative of MAR, and may also provide an insight to specific roles for MRS.

At the chromosomal level, MRS density had features similar to that of protein-coding genes, with more MRS per kilobase in chromosome centres compared to arms. Chromosome X was distinct in having no such pattern in gene density, and MRS on the X also had a flat distribution. The MRS signature is AT rich, and thus some correlation with local AT% of the genome would be expected (Figure 2.8). The question of whether the distribution of the MRS signature was merely a by-product of the local nucleotide content of the genome, and/or of the local content of di-, tri- and tetra-nucleotides was investigated. When genome sequence was randomised in 2 Mb sections, the frequency of MRS observed in the real chromosomal DNA was less than that predicted from simple (mononucleotide) randomisation, and approximately double that found in second and third order Markov model randomisations. Thus, the distribution of the MRS signature in the *C. elegans* genome is not simply a product of small- or large-scale base-compositional biases. MRS frequency in some classes of genomic regions was elevated compared to the surrounding sequence. Coincidence of MRS and genes was apparent from their similar chromosomal distributions (as shown in Figure 2.3). By analysing the overlap of MRS with different functional parts of the genome, it was found that MRS had relatively high incidence in the non-coding parts of genes, specifically 5' and 3' UTRs. These results contrast with experimental identification of a high incidence of MAR in intergenic and intronic regions, rather than UTRs. This suggests that MRS may not be representative of a large portion of MAR.

There were striking peaks of average MRS frequency at the 3' and 5' ends of *C. elegans* CDS, which were distinct from similar peaks in average AT content in the

same regions. Interestingly, the average MRS frequency surrounding *C. briggsae* CDS showed no peak at the 5' end, though the pattern of average AT content was very similar to *C. elegans*. However, the peak at the 3' end of CDS was maintained in *C. briggsae* and there was evidence for conservation of MRS in this region.

Although *C. briggsae* orthologs of *C. elegans* genes that had 3' MRS were more likely also to have an MRS than were orthologs of genes that lacked an MRS, it was surprising that the MRS was conserved in only 10% of orthologs. It is possible that the MRS, as currently defined, does not accurately represent the potential functional element. The non-conserved MRS from both *C. elegans* and *C. briggsae* could represent a high 'false positive' rate, giving rise to a background level of MRS that masks the degree of conservation of the underlying functional element. Alternatively, the apparent low level of conservation of MRS could reflect rapid evolution of the MRS. The association of MRS with the start and stop of genes means they are in a position to influence the control of transcription. However, if, as discussed above, the MRS does not accurately represent an underlying functional element and is subject to a high false positive rate, then the true degree of association with specific annotations may be underestimated. The presence of MRS in *C. elegans* 5' and 3' UTRs suggest that they may be transcribed and therefore also have a role in mRNA stability or translational control. The MRS is therefore an element that is perhaps of limited value in predicting MAR, but serves as a clear marker of some CDS boundaries.

Chapter 3 - Further Analyses of the Matrix Attachment Region Recognition Signature in *C. elegans*

3.1 Abstract

The findings from the previous chapter cast some doubt over the effectiveness of the MRS as an indicator of MAR but hinted at some kind of functional role for the MRS. In this chapter, three approaches were used to resolve the role of the MRS. In the first, the influence of configuration of the MRS in genomic sequence, such as its size and the degree of overlaps between and within MRS, was studied. This analysis also showed that both motifs were found to influence the MRS distribution pattern. Secondly, the MRS was compared to MAR predictions made using another method, SMARTest. The MAR predictions derived from MRS and from SMARTest were found to have little correlation. Finally, the relationship between the MRS and genes in *C. elegans* was studied. Genes associated with an MRS in the regions immediately flanking the CDS were found to be significantly enriched for the GO term 'receptor activity' and to have elevated expression levels. However, no correlation was found between genes associated with MRS and operons or *trans* splicing of a leader sequence to mRNA.

3.2 Introduction

As implied by its name, the MRS was originally proposed as a signature of MAR. However, the analysis of the frequency and distribution of the MRS in the genome of *C. elegans* and *C. briggsae*, described in the previous chapter, presented evidence that questioned the relationship between MRS and MAR. Rather than abundance in genomic regions associated with MAR, the MRS was instead found to have a distinctive relationship with genes. The spikes in average MRS frequency found flanking coding sequence hint at a role in transcription, but provided no firm evidence of a specific function. The objective of this chapter is to try and resolve the

role of the MRS: Is it a feature peculiar to MAR or a potentially functional characteristic of genes?

This chapter tackles this issue in three parts. In the first section the characteristics of the MRS are analysed in detail to determine the effect, if any, of the complex structure of the MRS has on its frequency and distribution. The following sections study each potential MRS role in turn. Firstly, the power of MRS to detect MAR is examined. Next, the ability for the MRS to function in the control of genes is explored.

3.2.1 Analysis of the MRS

The MRS is comprised of two different, degenerate motifs that are separated by a variable amount of spacer DNA. Motif 1 (16 bp, AWWRTAANNWWGNNNC) and motif 2 (8 bp, AATAAYAA) may occur on either strand of the DNA duplex and they may overlap [42]. This complexity permits many different nucleotide sequences to contain an MRS. Information about the specific characteristics of how the MRS occurs in genomic sequence allow the positioning and abundance of MRS to be put in context. The maximum size of the MRS is constrained by the limit of 200 bp separation between the two motifs. By permitting the motifs to overlap a minimum size of 16 bp is possible, where the 8 bp motif occurs entirely within the larger motif. However, this situation is only possible in a small set of specific nucleotide combinations and the frequency of these in genomic sequence is unknown. Investigation of the size profile of the MRS may aid in inference of their function. For example, it is important to know how frequently overlapping motifs occur. In addition to motifs overlapping within an MRS, overlapping may also occur between multiple MRS. This could happen through motifs being shared between MRS, partial overlapping of motifs or one MRS lying in the spacing between the motifs of another MRS. The degree to which a section of DNA includes multiple MRS has implications for understanding the density of MRS over short regions. To tackle these questions surrounding MRS overlaps and the potential density of MRS at one site,

analyses of MRS size and the distance between MRS were employed.

Analysing the distribution of motif 1 and 2 separately allows the contribution each one makes to the MRS distribution to be assessed. The simple probability (ignoring overlaps) of finding motif 1 and 2 in randomised sequence with equal proportions of the four nucleotides (i.e. 50% AT) is 2.59 $x10^{-4}$ (518 in 2 Mb) and 3.05 x 10^{-5} respectively. The probability in 64% AT sequence (the mean AT% of the *C. elegans* genome) is 9.63 x 10^{-4} (1,926 in 2Mb) and 1.72 x 10^{-4} (344 in 2Mb) for motif 1 and 2 respectively. If one motif is found to be particularly rare in actual genome sequence then it may be possible that it has a greater influence over the occurrence of the MRS than the other motif. These analyses may help our understanding of how the MRS functions. For example, if one motif is found to make no appreciable difference to the MRS distribution then it may not be necessary to consider it as having a functional role.

In the previous chapter, sequence randomisation was used to measure the effect of nucleotide composition of the genomic sequence on MRS frequency. By randomising the sequence in sections of varying lengths, the influence of local variations in sequence nucleotide composition could also be assessed. In this chapter, the frequency of randomised MRS in real genomic sequence is studied. The complex structure of the MRS means that the effect of randomising it is difficult to predict. The empirical experiments described in this chapter allow MRS randomisation to be compared with genomic sequence randomisation.

3.2.2 MRS-MAR

In the previous chapter, the analysis of the distribution of the MRS indicated they do not act as a good marker of MAR. MAR would be expected to be most prevalent in intronic and intergenic regions of a genome, a distribution pattern not reflected by MRS. However, the MRS was not designed to predict MAR in their entirety: they represent sequence motifs that are likely to be part of a MAR. The MRS and the

MAR predicted from it will almost certainly have different boundaries. Therefore, to make a fairer assessment of the power of MRS to predict MAR, MRS need to be distinguished from MAR. By processing the MRS data in some way to generate MRS-MAR, a more accurate representation of MAR than may be predicted by MRS should be obtained. MRS-MAR can then be used as the unit to test MRS prediction of MAR, for example by determining if MRS-MAR frequency and distribution matches what is known about MAR frequency and distribution. Creating MRS-MAR will also allow the MRS to be compared to other MAR prediction methods. In this chapter, MRS-MAR are compared to predictions of MAR from the SMARTest program (SMART-MAR), which has been used to predict MAR in a number of studies [56, 58]. The level of agreement between MRS-MAR and SMART-MAR will give an indication of whether both methods are predicting the same type of element. However, in the absence of a much larger dataset of experimentally defined MAR in *C. elegans,* the problem of determining which method, if either, faithfully predicts MAR remains.

Before MRS-MAR predictions can be investigated, the issue of how best to create them must be tackled. In the original description of the MRS, van Drunen *et al*. describe several MAR that contain multiple MRS [42]. The often short distance between MRS and the not uncommon occurrence of overlapping MRS is described in detail below. Therefore, one way of processing the raw MRS data to create MRS-MAR, is to merge closely apposed MRS. This method should create a unit with a size more closely related to that of MAR, although it will not define the actual boundaries of MAR. By treating closely apposed MRS as a single unit, the number of MRS-MAR should be closer to the number of true MAR. The major obstacle in this method is selection of the correct distance allowed between MRS to create accurate MAR predictions. There is no single correct value, so whatever distance parameter is chosen will be a generalised estimate. The number of MRS-MAR created by different values for the distance parameter was investigated before selection of the most appropriate value for the final MRS-MAR dataset. The frequency and distribution of

the MRS-MAR and a comparison against SMART-MAR was used to make a judgement as to how appropriate the MRS is for the prediction of MAR.

3.2.3 MRS relationship with genes

As described in the previous chapter, the abundance of MRS in UTR and the regions immediately flanking CDS suggest that rather than serving as a signature of MAR, the MRS may be more appropriately thought of as a marker of genes. Additionally, the proximity of the spike in average MRS frequency to CDS suggests that significant numbers of MRS are likely to be transcribed. Therefore, MRS may play a role in either transcription, post-transcription or more generally in gene expression. The final part of this chapter therefore focuses on genes with neighbouring MRS (termed MRS-genes). Several categories of MRS-genes were identified, based on the position of the MRS (upstream or downstream of the CDS) and its proximity to the coding sequence. In this section, analysis was guided by the assumption that if MRS play a role in controlling the expression of genes, then MRS-genes could be distinguished from other genes either functionally or through some other identifiable characteristic. To this end, MRS-genes were studied with respect to their functional annotation, position in operons, correlation with post-transcriptional addition of spliced leader sequence and their temporal and spatial expression patterns.

The Gene Ontology (GO) is a system of providing functional annotation to genes using a hierarchical set of terms, universal to all organisms [85]. GO can be used to identify a disguising characteristic of genes associated with a particular genomic feature. For example, highly conserved non-coding elements have recently been found to frequently occur near genes involved in developmental control using GO annotation [9, 13, 86]. Similarly, one possible role for the MRS is that it acts as a kind of *cis* control element for a specific class of genes. Over-representation of a particular GO term (or terms) in the MRS-gene set would provide strong evidence that the MRS plays a functional role in the expression of those genes.

Operons are a common feature of gene organisation in *C. elegans* and other nematodes [87]. In an operon, a group of genes is under the control of a single promoter. Transcription of the operon produces a poly-cistronic pre-mRNA which, in nematodes, is spit into separate mature mRNAs. The average frequency of MRS surrounding CDS showed a clear correlation with genome regions likely to be fundamental to transcriptional initiation and termination. If MRS do have a role in transcription then we would expect to see a distribution of MRS surrounding genes in operons that reflects their distinct transcriptional organisation. In this chapter, the positioning of various categories of MRS-genes in operons (e.g. external or internal) was used to investigate the relationship of MRS with operons and therefore test whether there was any evidence for a role for MRS in transcription.

Trans splicing of a short exon, the spliced leader (SL), to the 5' end of mRNAs has been identified in several eukaryotes, including cnidaria [88], urochordates [89], rotifers [90] and nematodes [91]. Over half of *C. elegans* mRNAs are *trans*-spliced to SL1, a 22 nucleotide non-coding sequence [92]. *Trans*-splicing of SL1 is closely related to *cis*-splicing. A second spliced leader sequence, SL2, is *trans*-spliced to most downstream genes in *C. elegans* operons [87]. The SL acceptor sites tend to be close or immediately adjacent to the start of the CDS [92]. The proximity of these sites with the upstream spike in average MRS frequency opens the possibility that the two features are functionally related. For example, the MRS may be involved in aiding recognition of the SL1 acceptor site by the SL1 splicing machinery. Alternatively, a negative correlation between SL1 acceptor sites and MRS would support the suggestion that the MRS was required post-transcriptionally, as it lies in the region of sequence that is cleaved during ligation of the SL1 RNA. The relationship between MRS and SL1 was studied by measuring the correlation between the presence of an SL1 acceptor site with MRS-genes.

If the MRS directly affects gene expression, then this may be detectable through the use of Serial Analysis of Gene Expression (SAGE) data. SAGE involves the use of

short DNA tags to measure the level of each mRNA in a cell. SAGE provides similar information to that gained from microarray experiments, although as the SAGE observations are not based on hybridization, more qualitative values are obtained. This allows direct comparisons to be easily made between SAGE libraries, whereas comparing microarray experiments is more difficult [93]. By using cDNA libraries from different cell types and at different developmental stages, SAGE has been used to build up a spatial and temporal picture of gene expression levels in *C.elegans* [94]. In this chapter, the mean expression level of MRS-genes in each of the SAGE libraries was used to detect any developmental stages or cell types where MRS-genes were either over- or under-expressed. Significant over- or under expression of MRS-genes would provide evidence that MRS had an influence over the expression of nearby genes.

3.3 Methods

3.3.1 Analysis of the MRS

Distance between MRS

The distance between MRS was calculated as the number of nucleotides between the mid-point of an MRS to the mid-point of the nearest MRS. The MRS mid-point was defined as the position coordinate of the nucleotide that was equidistant from the start and end positions of the MRS. Where the mid-point lay between nucleotides, the nucleotide nearer the MRS start position was chosen as the mid-point.

Simulated data was generated by selecting X random numbers between 1 and Y (where X = the number of MRS in the genome and Y = the length of the genome), simulating random distribution of MRS throughout the genome with respect to other genome features. Treating the numbers as mid-points, the distance between each one was calculated as for MRS. Mean values were calculated over 1000 simulations.

Analysis of MRS motif 1 and 2

MRSfinder (see Chapter 2, Methods) was adapted to find every instance of motifs 1 and 2 in throughout the *C. elegans* genome. The distribution of each motif was calculated for non-overlapping 2 Mb windows, with the last window scaled where appropriate.

Counts of motif 1 and 2 in locally randomised genomic sequence were obtained in a similar fashion. Ten bp locally randomised sequence was generated using a Fisher-Yates shuffle, 2 Mb locally randomised sequence was generated using a roulette wheel selection algorithm (see Chapter 2, Methods for more details). The *C. elegans* genome was subjected to each randomisation 1000 times.

Random MRS

To generate randomised versions of the MRS, the sequences of motif 1 and 2 were individually randomised using a Fisher-Yates shuffle. A modified version of MRSfinder was then used to search for the randomised MRS in the sequence of the *C. elegans* chromosomes. The mean and standard deviation of the number of MRS for each 2 Mb window was calculated over 1000 randomisations.

3.3.2 MAR from MRS and SMARTest

MRS-MAR were created by aggregating MRS separated by less than a chosen distance parameter.

SMART-MAR predictions were obtained via the genomatix website [95]. As the web interface will only report a maximum of 1000 MAR predictions per input sequence, the *C. elegans* chromosomes were split into 2.2 Mb sections, each with a 5000 bp overlap at both ends. After ensuring that the limit of 1000 predictions had not been reached in any of the 2.2 Mb sections, the SMART-MAR predictions were merged to form a continuous set of predictions for each chromosome.

Each MRS-MAR was assessed to determine if it overlapped with a SMART-MAR. MRS-MAR lying entirely within a SMART-MAR defined a complete overlap. If the overlap was not complete but involved at least 12 bp of the MRS-MAR, it was scored as a partial overlap. The expected number of overlaps was the sum of the expected complete and partial overlaps, calculated using the formula described earlier (Chapter 2, Methods). Briefly:

$$\text{expected number of complete overlaps} = M.F((f-m)+1)/g$$

$$\text{expected number of partial overlaps} = M.F(2(m-w))/g$$

where M = number of MRS-MAR, F = number of SMART-MAR, f = mean length of SMART-MAR, m = mean length of MRS-MAR, w = minimum number of nucleotides required for a partial overlap (here 12), g = genome/chromosome length.

3.3.3 MRS relationship with genes

MRS-genes

MRS-genes were defined based on the presence of an MRS near the translation start or stop site. Several categories of MRS-genes were created, based on the location of MRS with respect to the translation boundaries of genes. Genes with an MRS within 200 bp of the translation start /stop site were classed as ‘close’, ‘1k’ was used to identify genes with an MRS within 1 kb of the translation start/stop. Genes were classified further based on the presence of an MRS near the translation start site, stop site or both start and stop. The ‘1k start or stop’ set includes genes with an MRS within 1 kb of the translation start or translation stop and is inclusive of genes with MRS within 1 kb of both the start and stop site. The number of genes in each category is shown in Table 3.1.

MRS-gene type	Number in C. elegans genome (% of total)
‘1k start or stop’	6124 (30%)
‘1k start and stop’	622 (3%)
‘1k start’	3340 (16%)
‘1k stop’	3407 (17%)
‘close start’	1078 (5%)
‘close stop’	1057 (5%)

Table 3.1 MRS-genes

GO terms associated with MRS-genes

Two MRS-gene sets were selected for GO-term analysis. The 'close stop' set represents genes with the strongest bias in MRS distribution, i.e. MRS within 200 bp of the translation stop site. The '1k start or stop' set encompasses the widest set of genes associated with an MRS. Reference sets were also created that contain all the *C. elegans* genes not contained in the corresponding MRS-gene set, e.g. the 'non close stop' set contained all genes that did not have an MRS within 200bp of the translation stop site.

Perl scripts were used to map GO terms from GO annotation/association file dated 11/05/2007 to the MRS-gene and reference sets. As noted by Vavouri *et al.*, there is a heavy bias towards RNAi phenotypes in GO annotations of *C. elegans* genes [13]. For this reason, both the compete GO annotation set and a set consisting only of those terms inferred through automatic electronic annotation (evidence code IEA) were used. The perl script map2slim from the go-perl package was used to map these GO terms to their parent GO slim terms and give gene counts for each term. The number of annotated genes for each gene set is given in Table 3.2.

The log odds ratio (OR) was used to compare the relative representation of genes from the MRS-gene sets and reference gene sets. For each GOslim term log (OR) was calculated thus:

$$\text{Log } ((a.b)/(c.d))$$

Where a is the number of genes in the MRS-genes set with the specified annotation, b is the number of genes in the reference set without the specified annotation, c is the number of genes in the reference set with the specified annotation and d is the number of genes in the MRS-genes set without the specified annotation.

Gene set	‘Close stop’ MRS-genes	‘1k start or stop’ MRS-genes	'close stop' reference set	'1k start or stop' reference set
Number of genes	1,057	6,124	18,982	13,915
Full GOslim annotation	603	3,317	10,308	7,540
IEA only GOslim annotation	509	2,976	9,102	6,635

Table 3.2 GO annotation of MRS-genes

The p- value for each log (OR) was derived from the z-score, calculated according to the following:

$$\text{standard error of log (OR)} = \sqrt{(1/a + 1/b + 1/c + 1/d)}$$

$$z = \log(OR)/\text{standard error of log (OR)}$$

The p-value associated with z is determined by referring to the cumulative distribution of the standard normal curve. The confidence limits were calculated as:

$$\log(OR) \pm 1.96 \times \text{standard error of log (OR)}$$

where 1.96 is the z value defining the two-sided 95% confidence interval.

In order to correct for multiple testing, a 5% false discovery rate threshold was calculated according to the false discovery rate method of Benjamini and Hochberg [96].

MRS-genes in operons

Using operon annotation from the BioMart section of WormBase (release WS150, [97], each *C. elegans* gene was designated as being one of: first gene of an operon, last gene of an operon, internal position in an operon, or not associated with an operon. Each gene was then assigned an MRS-gene status, according to the previously complied lists of MRS-genes.

Spliced leader acceptor sites

The Ace Query Language (AQL) was used to search the Acedb wormbase database (version WS180 [98]) for SL acceptor sites and the genes that they were associated with. The total number of annotated SL acceptor sites and the number associated with genes is summarised in Table 3.3. the number of genes associated with a SL acceptor site is some way short of the number of genes reported to be *trans*-spliced (about 70% in total, [92]). Therefore, the set of genes with an annotated SL acceptor site is likely to be incomplete.

	Acceptor site type		% of all genes associated with splice acceptor site
	Annotated	**Associated with a gene**	
Spliced leader 1	7671	3948	20
Spliced leader 2	2054	1086	5
Spliced leader 1 and 2	9725	5034	25

Table 3.3 Number of annotated SL1 sites assigned to a gene.

The log odds method (described above) was used to measure the degree of association between genes that have a spliced leader acceptor site and MRS-genes.

Gene expression levels using SAGE

SAGE data were obtained from the Genome BC *C. elegans* Gene Expression Consortium [94]. For each available SAGE library, tags, their associated genes and the tag counts were collated. The tag counts were normalised to a standard library size of 100,000. The gene- tag associations were used to match tag counts to either MRS-genes or non MRS-genes. Differences in expression levels between the gene sets was measured by calculating the mean tag count for each SAGE library. The Student's t-test was used to measure the significance of the difference between the means. In order to correct for multiple testing, a 5% false discovery rate threshold was calculated according to the false discovery rate method of Benjamini and Hochberg [96].

3.4 Results

3.4.1 Analysis of the MRS

MRS characteristics

The MRS is unusual for a sequence signature in that its size can range from 16 bp (where motif 2 lies wholly within motif 1) to 224 bp (where the nearest ends of motif 1 and 2 are separated by 200 bp). The size distribution of the MRS is useful, for example when considering its positioning in relation to genes. Figure 3.1 summarises the size of each MRS in the *C. elegans* genome. In about 12% of MRS, motif 2 lay wholly within motif 1. The degeneracy of the motifs (particularly motif 1), the possibility that the motifs may occur on different strands, and the AT richness of both motifs mean that there are many different configurations that allow motif 1 and motif 2 to occupy the same locus. For the same reasons the number of MRS in which the two motifs overlap to some degree was also large: an additional 18% of the total MRS fell into this category. The remaining 70% of MRS were greater than 23 bp (i.e.

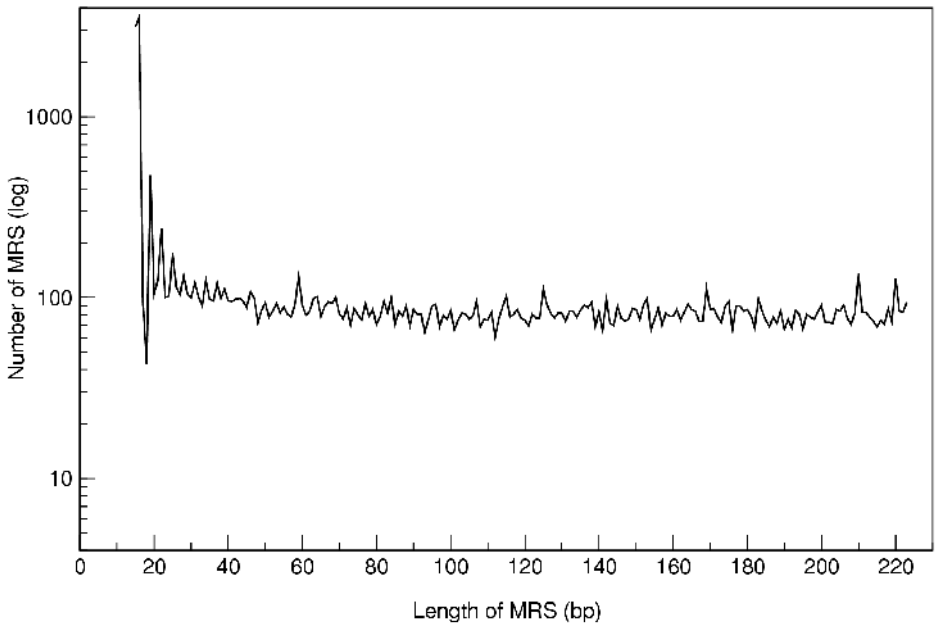


Figure 3.1 Length of MRS in C. elegans

no overlap between the motifs) and had an even distribution of sizes across the remaining range.

As reported in Chapter 2, MRS have non-random distribution across the genome of *C. elegans,* with a tendency to cluster around genes. One method of testing if the distribution is significantly different from random is to measure the distance between MRS. The distances between a randomly distributed pattern would be expected to have an exponential decay. In Figure 3.2, the distance between MRS is compared to simulated data representing the same number of elements randomly spread over the genome. As expected, the frequency of simulated data points decreases exponentially for increasing distance between them. In contrast, the distances between MRS followed a more complex pattern that lay almost entirely beyond the upper and lower 2.5% limits of the simulated data. The most extreme departure from the simulated data occurred for separation distances of 0-200 bp and 201-400bp. Many fewer than expected MRS were separated by 201-400 bp. In contrast many more than expected MRS were separated by less than 201 bp and it was calculated that about three quarters of MRS overlapped with at least one other MRS. There are a number of features about the structure of the MRS that increase the chance of overlaps. Firstly, the van Drunen *et al*. definition of the MRS allows different MRS to share the same motif [42]. So, for example, a single motif 1 could have multiple instances of motif 2 within 200 bp and each one would be classed as a separate MRS. Furthermore, for the same reasons discussed above that allow motif 1 and 2 to overlap, namely the degeneracy of the motifs and their relatively homogenous nucleotide content, it is also possible that multiple motifs of one type will overlap.

As so many MRS are within 200 bp of each other it is not surprising that the number of MRS separated from each by more than 200 bp is much lower than found in the simulated data. However, it is interesting that towards the extreme end of the distance between MRS (i.e. >11,000 bp) the number of MRS is actually greater than found in the simulated data. Although there are only a few hundred MRS in this range, these

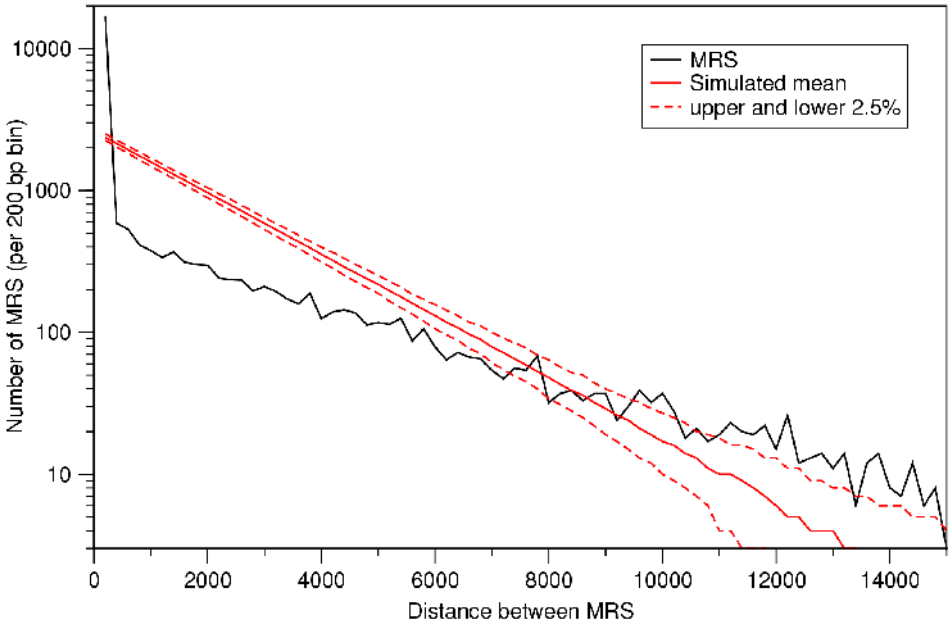


Figure 3.2 Distance between MRS in the C. elegans genome

The distance between MRS (black) is compared to the distances between a simulated set of randomly spaced elements (red).

data indicate that they are spaced further apart than expected under a random distribution model. In summary, the spacing between MRS shows that their distribution on the *C. elegans* chromosomes is different to that expected of a purely random pattern.

Motif 1 and 2

The frequency and distribution of the two motifs that make up the MRS may be expected to have a strong influence on the frequency and distribution of the MRS itself. Using an adapted version of MRSfinder, each instance of motifs 1 and 2 was located, regardless of whether it formed part of an MRS or not. The chromosomal distribution of instances of motifs 1 and 2 are compared to that of the MRS in Figure 3.3. Across all chromosomes, motif 1 was more frequent than motif 2 by a factor of about 3, which itself was more frequent than MRS by a factor of about 1.5. Although in absolute terms the variation of motif 1 is greater than for motif 2 and the MRS, proportionally it varies slightly less. Taking this into account, the distribution of motif 1 and 2 along each of the chromosomes is broadly similar to each other and to the MRS.

As the distributions of motif 1 and 2 are so similar and they both occur more frequently than the MRS, it does not appear that one motif has consistently more influence on MRS frequency than the other. Close inspection of Figure 3.3 reveals examples where MRS frequency was influenced primarily by each of the motifs at specific locations. Between 8 and 10 Mb on chromosome I there was an increase in both motif 2 and MRS, while motif 1 decreased slightly over this region. Conversely, the increase in MRS from 8 to 10 Mb on chromosome III was matched by motif 1 but not by motif 2. However, these cases may just represent a certain level of stochasticity in MRS frequency in relation to motif 1 and 2. This is evident at the beginning of chromosome II where the motif 1 and 2 decreased in frequency from 2 to 4 Mb but MRS frequency increased.

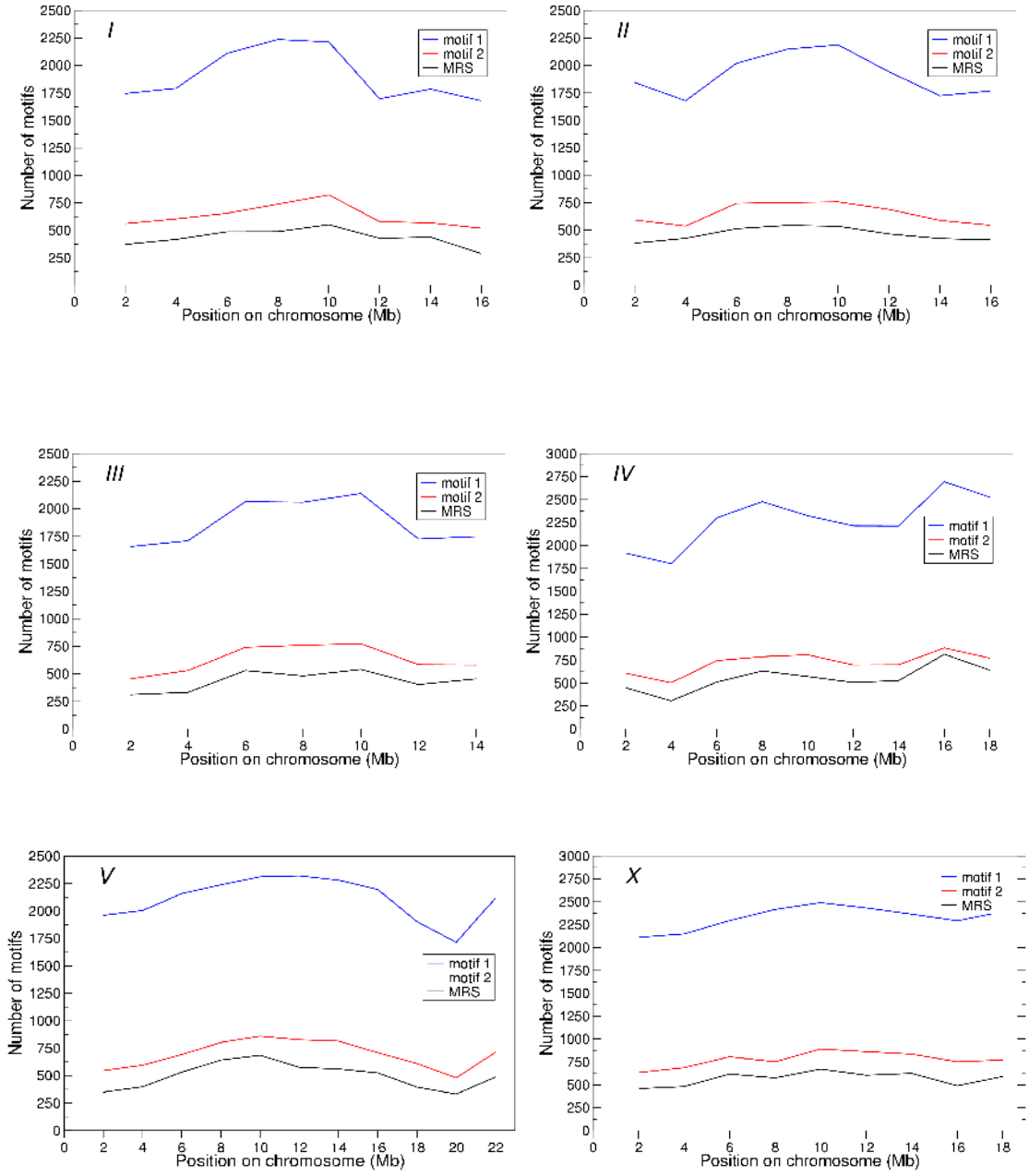


Figure 3.3 MRS, motif 1 and motif 2 in C. elegans chromosomes

Frequency of MRS (black), motif 1 (blue) and motif 2 (red) in 2 Mb windows along each of the *C. elegans* chromosomes.

Figure 3.4 shows the distribution of motif 1 and motif 2 in chromosome I sequence randomised in 2 Mb sections, using mononucleotide randomisation. In sequence randomised in 2 Mb sections MRS were slightly more frequent than in the real sequence and had a flatter distribution. This was also true for both motif 1 and 2, although there were proportionally more motif 1 in the 2 Mb random sequence than motif 2 and MRS.

As discussed in Chapter 2, the number of MRS in sequence randomised in sections of 10 bp was more than 2 fold greater than both real sequence and sequence randomised in 2 Mb sections (see Figure 2.7). Figure 3.5 shows that the frequency of motif 2 increased by a similar proportion to MRS but that motif 1 increased to a lesser extent in sequence randomised in 10 bp sections. The differential increase means that the ratio of MRS to motif 1 in chromosome I dropped from 4.3:1 in real sequence to 2.7:1 in sequence randomised in 10 bp sections, while the ratio of MRS to motif 2 only dropped slightly from 1.4:1 to 1.2:1. The distribution along the chromosome of MRS, motif 1 and motif 2 is also very different in sequence randomised in 10 bp sections compared to the equivalent distributions in real sequence. In the randomised sequence, although they all have quite flat distributions, the distribution of the MRS was clearly matched by that of motif 2 but not motif 1.

In summary, in real sequence the MRS frequency appeared to be dictated by both motif 1 and 2. The same was true when chromosome I was randomised in 2 Mb sections, although this did result in an increase in abundance of MRS and both motifs. The 2-fold increase in MRS previously observed when the sequence was randomised in sections of 10 bp was matched by motif 2 but not motif 1 which increased by a smaller amount. However, motif 1 was still the most abundant and could have been near the maximum level possible in sequence with that nucleotide content. This could explain why the distribution of motif 1 did not match that of the MRS in sequence randomised in 10 bp sections. On the whole, there is little or no evidence that one or other of the motifs had a greater influence over the MRS

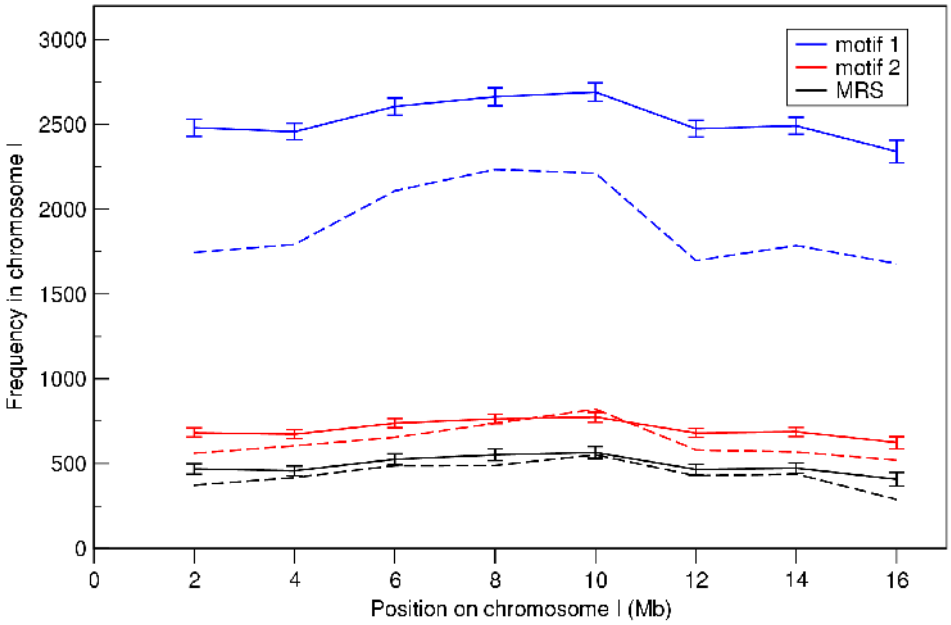


Figure 3.4 MRS, motif 1 and motif 2 in chromosome I, randomised in 2 Mb sections

The frequency of MRS (black) motif 1 (blue) and motif 2 (red) in *C. elegans* chromsome I sequence, randomised in 2 Mb sections using a mononucleotide process. Broken lines show frequencies in real chromosome I sequence. The error bars show +/- one standard deviation

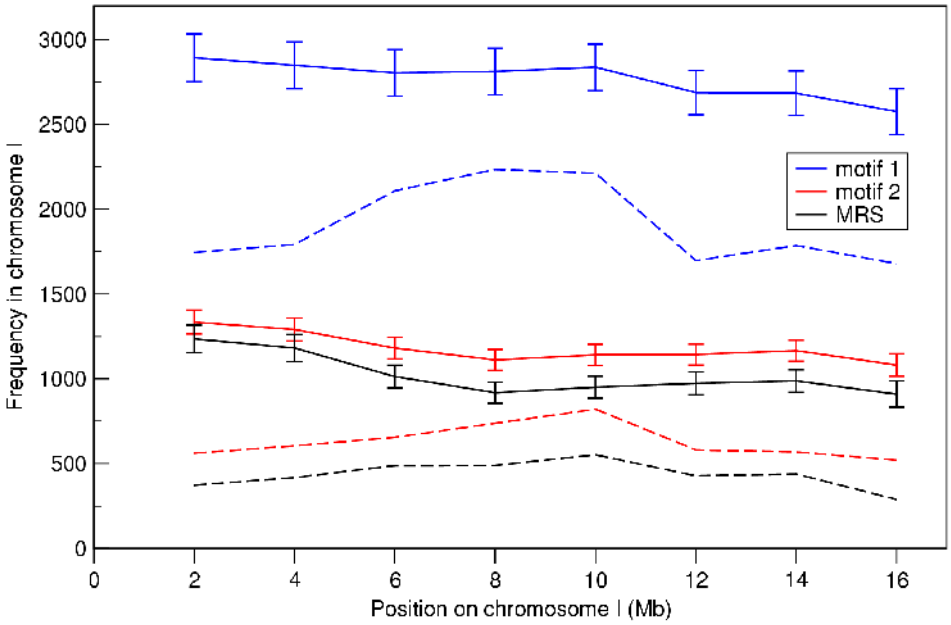


Figure 3.5 MRS, motif 1 and motif 2 in chromosome I, randomised in 10 bp sections

The frequency of MRS (black) motif 1 (blue) and motif 2 (red) in *C. elegans* chromsome I sequence, randomised in 10 bp sections using a mononucleotide process. Broken lines show frequencies in real chromosome I sequence.The error bars show +/- one standard deviation

frequency pattern in real or randomised chromosome I sequence.

MRS randomisation

By locally randomising a sequence in discrete sections we can observe the contribution made to the frequency of a pattern by the relative composition of the nucleotides in that section of sequence. Similarly, the pattern itself can be randomised to reveal the effect of its nucleotide content in the context of real genomic sequence. As shown in Figures 3.6 and 3.7, the number and distribution of randomised MRS found in chromosomes I and III was very similar to the number and distribution of actual MRS found in the same sequence randomised in 10 bp sections. The overall abundance and distribution of MRS in both these datasets were much closer to each other than either is to actual MRS in real sequence. The main difference is that the frequency of randomised MRS in real chromosome sequence had a much greater variance. The mean number of randomised MRS was consistently slightly lower than the mean number of MRS in sequence randomised in10 bp sections. From Figure 2.7 we know that the number of MRS increases as the randomisation section is shortened. It is reasonable to expect, then, that if the sequence was randomised in longer sections of ~24 bp (the total number of bases in the MRS motifs), then the number of MRS observed in the randomised sequence would be lower, and therefore even closer to the number found for randomised MRS. The experiment was only performed on chromosome I and III but there is no reason to suspect that the other chromosomes would produce a different pattern of results. The consistent finding that the frequency of real MRS in real sequence is significantly less than that of randomised MRS in real sequence, or real MRS in randomised sequence suggests that the MRS is a non-random complex pattern, and thus is likely to have functional significance.

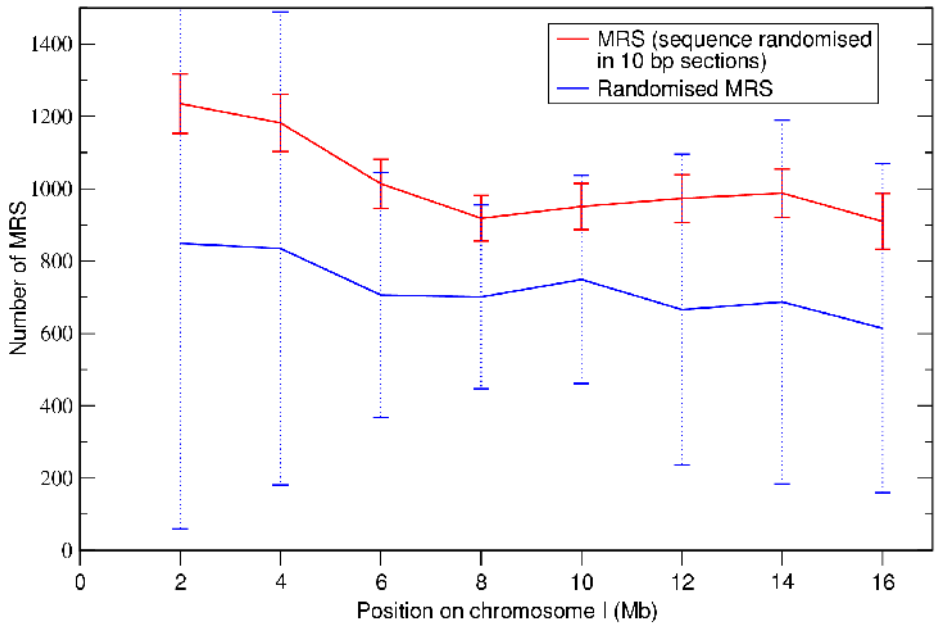


Figure 3.6 MRS in chromosome I sequence randomised in 10 bp sections compared to randomised MRS in actual chromosome I sequence.

The frequency of MRS (red) in chromosome I sequence randomised in 10 bp sections using a mononucleotide process is compared to the frequency of randomised MRS in actual chromosome I sequence. The error bars show +/- standard deviation.

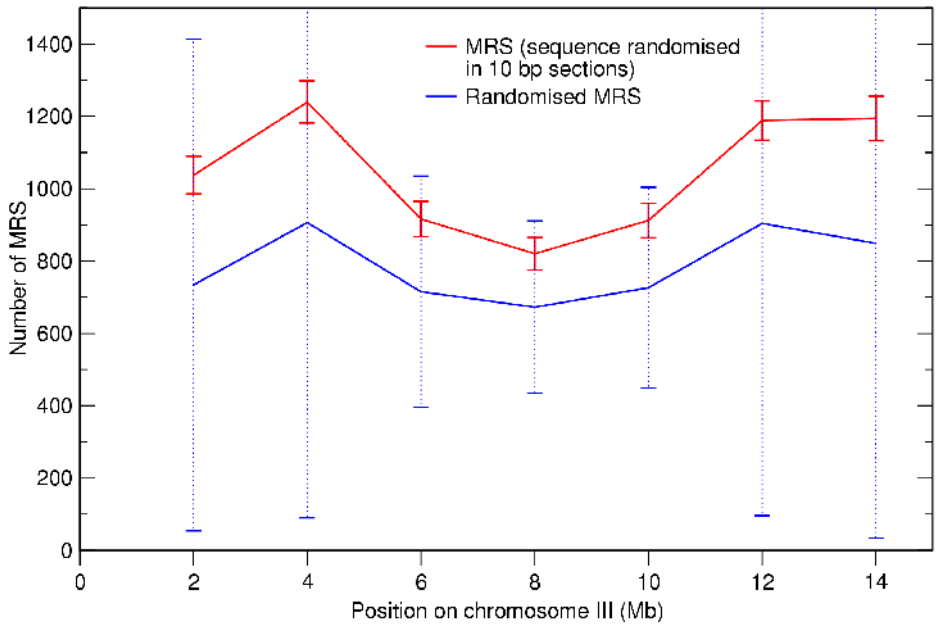


Figure 3.7 MRS in chromosome III sequence randomised in 10 bp sections compared to randomised MRS in actual chromosome III sequence.

The frequency of MRS (red) in chromosome III sequence randomised in 10 bp sections using a mononucleotide process is compared to the frequency of randomised MRS in actual chromosome III sequence. The error bars show +/- standard deviation.

3.4.2 MRS-MAR

MRS-MAR

The presence of an MRS is suggested to indicate that a MAR is present in the surrounding sequence, but does not define the boundaries of the MAR. As described above, MRS are commonly found in close proximity to each other and multiple MRS may lie in a single MAR. To get an idea of MAR abundance and distribution and to allow comparison of the MRS with other MAR prediction methods, the raw MRS data is not sufficient. One way of processing the data to give a more accurate reflection of MAR is to merge closely apposed MRS to form MRS-MAR. The main purpose of MRS-MAR is to reduce the effect of the large number of overlapping MRS, that biologically would be likely to represent a single MAR. The relationship between the distance parameter (the spacing between MRS, below which they were grouped into one MRS-MAR) and the number of putative MAR generated is shown in Figure 3.8. For values of less than 200 bp the number of MAR remained relatively stable and, as this approximates to the minimum size of MAR, for the results described below, MRS-MAR were created by merging MRS that lay within 200 bp of each other. However, MAR can span up to several kilobases, so this is a conservative estimate that is likely to over-predict the number of MAR.

The size profile of the 200 bp-merged MRS-MAR (Figure 3.9) confirmed that a range parameter of 200 bp may have been too conservative and highlighted the limitations of using this method to obtain accurate predictions of true MAR. Most MRS-MAR were found to be short (<200 bp). The large number of MRS-MAR <= 20 bp derived from isolated MRS in which motif 1 and 2 overlapped. As many MAR may only have one MRS, increasing the range parameter to make them form larger MRS-MAR would not necessarily give a more accurate representation of the true situation. Although MRS-MAR may not accurately reflect true MAR, by combining overlapping and closely spaced MRS into one MRS-MAR prediction the representation of MAR is likely to be improved over the use of MRS alone.

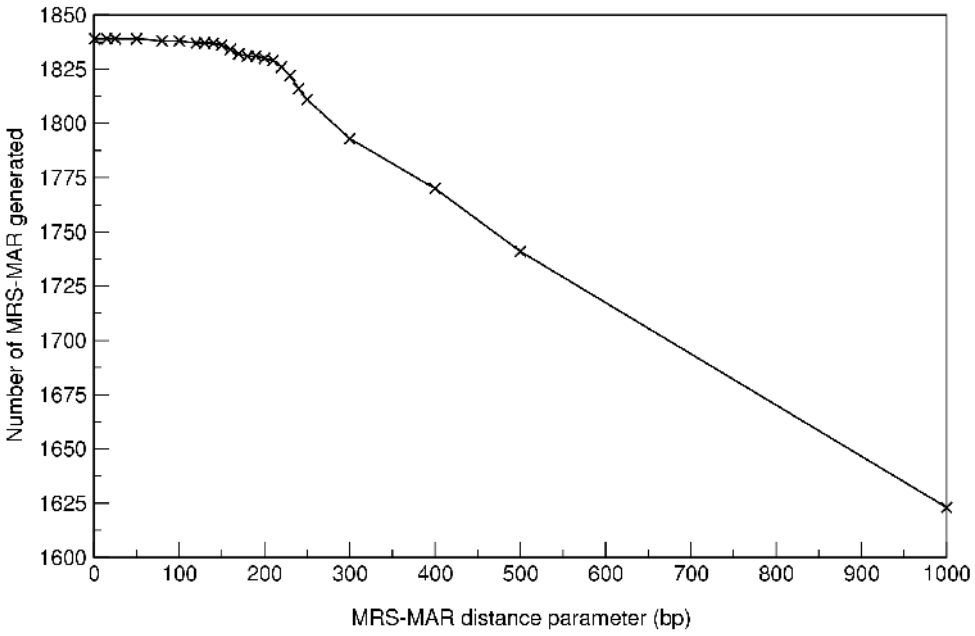


Figure 3.8 Number of MRS-MAR created when the MRS-MAR distance parameter is varied

MRS-MAR were created by aggregating MRS separated by less than the number of bases specified by the MRS-MAR distance parameter. This figure shows the number of MRS-MAR created for a range of distance parameter values. A distance parameter of 200 bp was used to create the MRS-MAR used in subsequent analyses.

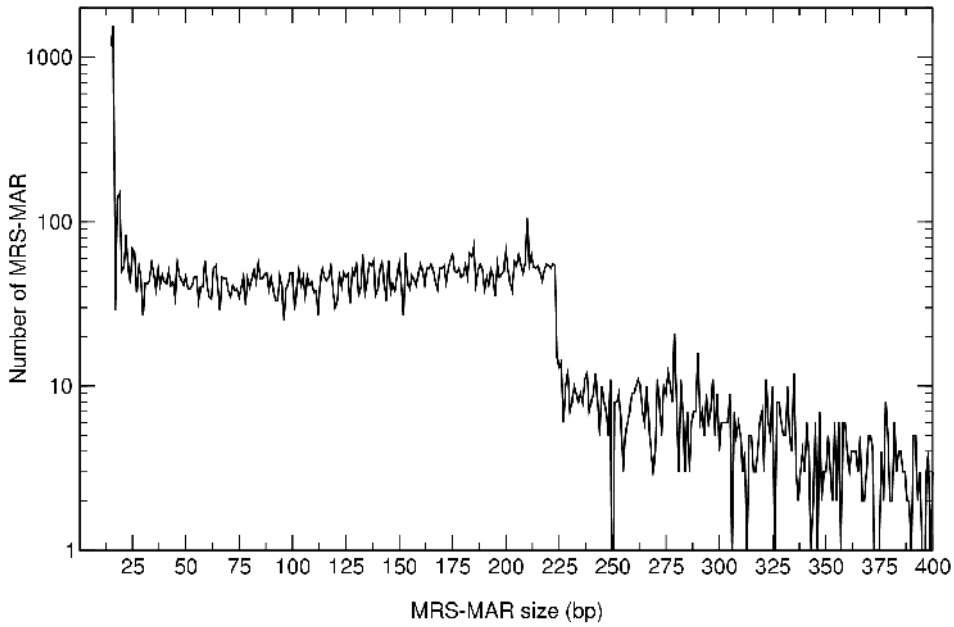


Figure 3.9 MRS-MAR size distribution

This figure shows the size distribution of MRS-MAR created using a distance parameter of 200 bp. MRS-MAR less than 25 bp long consist of isolated MRS in which motif 1 and 2 are adjacent or overlap. MRS-MAR less than 200 bp are mostly composed of isolated MRS with up to 200 bp separating motif 1 and 2.

From the 24,967 *C. elegans* MRS, 13,761 MRS-MAR were derived. The ratio of MRS to MRS-MAR was similar between all six chromosomes and indeed remained generally consistent along the length of each chromosome (Figure 3.10). However, where the MRS frequency exceeded about 500 per 2 Mb window, the ratio of MRS to MRS-MAR dropped slightly. The relatively high MRS frequency over 2 Mb was reflected in closer MRS spacing over the range of a few hundred base pairs, which in turn resulted in more MRS per MRS-MAR than elsewhere in the genome.

Comparison with SMARTest

Creation of MRS-MAR allows a more meaningful comparison with other MAR prediction methods than do MRS alone. Here the MRS-MAR predictions are compared with those from SMARTest [26]. SMARTest identifies potential MARs by performing a density analysis based on a S/MAR matrix library of MAR-associated sequences described as weight matrices. The distribution of SMART-MAR was compared to MRS-MAR (Figure 3.11). Interestingly, there were well over twice as many SMART-MAR as MRS-MAR, despite the conservative distance parameter used to derive MRS-MAR. Indeed, the number of SMART-MAR exceeds the number of MRS. There was no obvious correlation in chromosomal location between the distributions of the two predictions, even though both prediction methods are based on AT-rich sequences. In contrast to the enrichment of MRS-MAR in the centres of all autosomes, there was no common chromosomal distribution of SMART-MAR.

The disparity between SMART-MAR and MRS-MAR predictions is further underlined in Table 3.4. Across the genome, only 48% of MRS-MAR overlapped with a SMART-MAR and just 18% of SMART-MAR overlapped with an MRS-MAR. Estimation of the expected number of overlaps (assuming random and independent distribution) revealed that the number of overlaps between MRS-MAR and SMART-MAR was about one third greater than the number expected by chance. Thus it appears that while there is some correlation between MRS-MAR and SMART-MAR, they largely describe different entities.

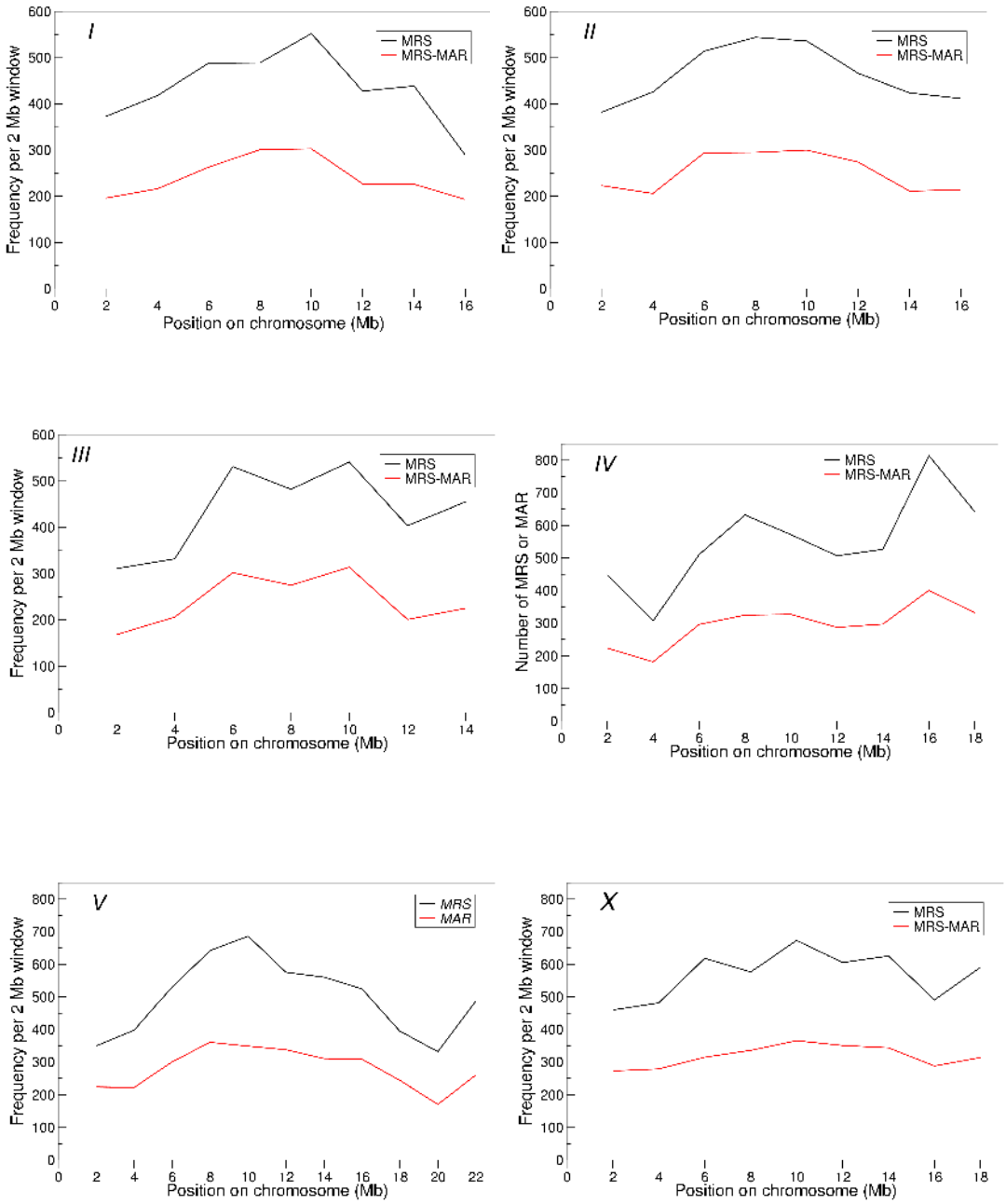


Figure 3.10 Frequency of MRS-MAR (red) in 2 Mb windows compared to MRS (black) for each of the C. elegans chromosomes

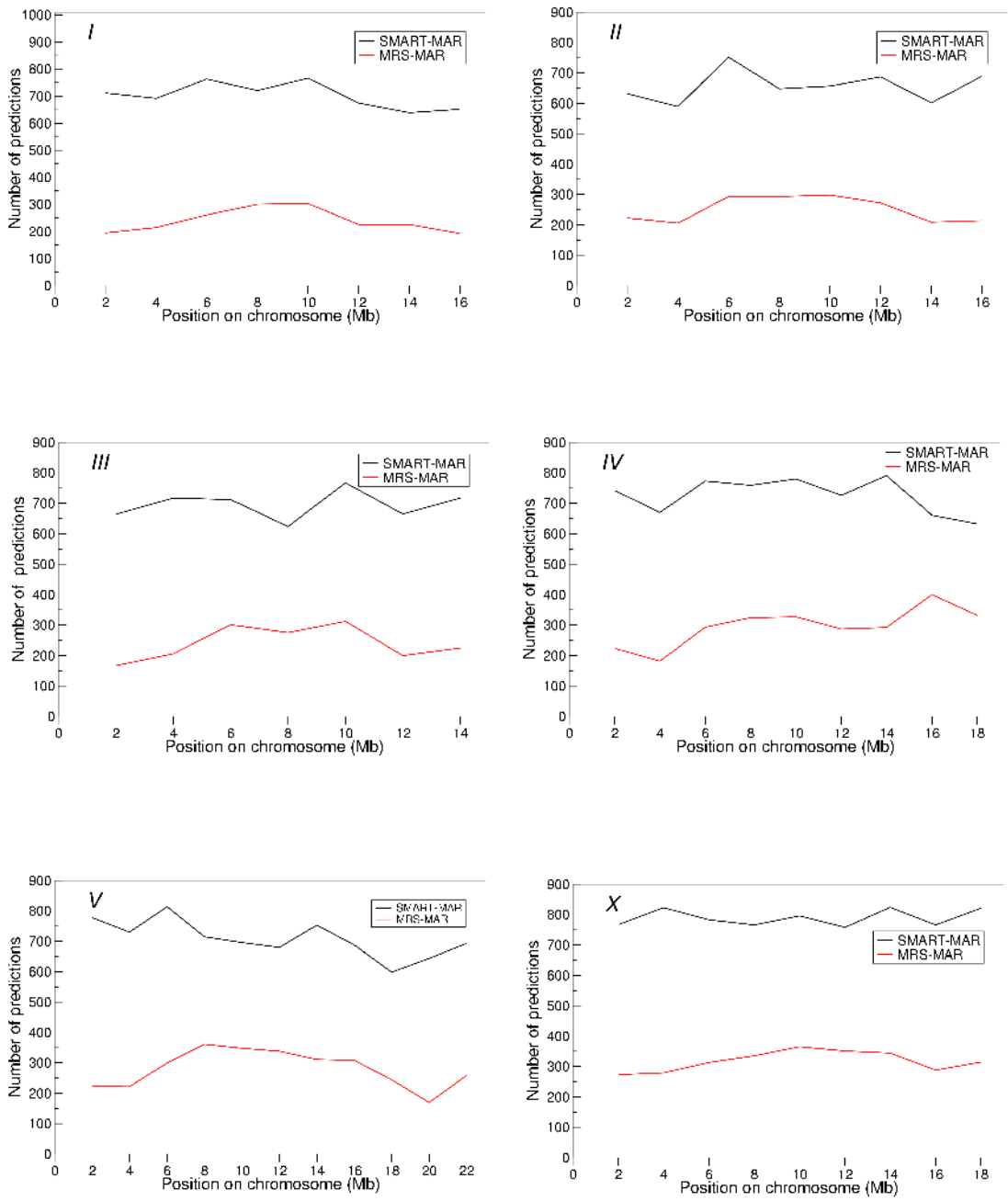


Figure 3.11 Frequency of MRS-MAR (red) in 2 Mb windows compared to SMART-MAR (black) for each of the C. elegans chromosomes.

Chromosome	MRS-MAR (MM)	SMART-MAR (SM)	MM that overlap SM (%)	SM that overlap MM (%)
I	1,830	5,314	42	14
II	1,931	5,009	43	17
III	1,662	4,789	52	18
IV	2,577	6,375	59	24
V	2,943	7,417	43	17
X	2,818	6,989	47	19
Genome	13,761	35,893	48	18
Expected			36	14

Table 3.4 Number of MRS-MAR and SMART-MAR and percentage of overlaps

To help understand the relationship between MRS-MAR, SMART-MAR and biochemically defined MAR, their positions on the experimentally investigated *C. elegans* cosmid M88 are depicted in Figure 3.12. This ~32 kb sequence was the cosmid used by van Drunen *et al*. to identify the MRS and is the only section of the *C. elegans* genome for which MAR have been experimentally defined [42]. In this figure, the restriction enzyme fragments that gave a positive result in the matrix re-association assay determine the coordinates of the experimentally defined MAR. However, the restriction enzyme fragments do not necessarily represent the exact boundaries of the MAR. As indicated by the previously discussed analyses, there was a high degree of overlap between MRS in cosmid M88. Figure 3.12 shows how the overlapping signatures were consolidated into one MRS-MAR. Despite this the MRS-MAR still only span a small proportion of the MAR-containing restriction fragments. Furthermore, two of the six MAR contained multiple MRS-MAR, confirming the earlier suggestion that parameters used to define MRS-MAR are conservative. In contrast to MRS-MAR, the SMART-MAR were both more numerous and had a more uniform in their distribution. Although there was some agreement between SMART-MAR and experimental MAR, the SMART-MAR also covered a large fraction of genomic sequence not assigned MAR status in the matrix re-association assay.

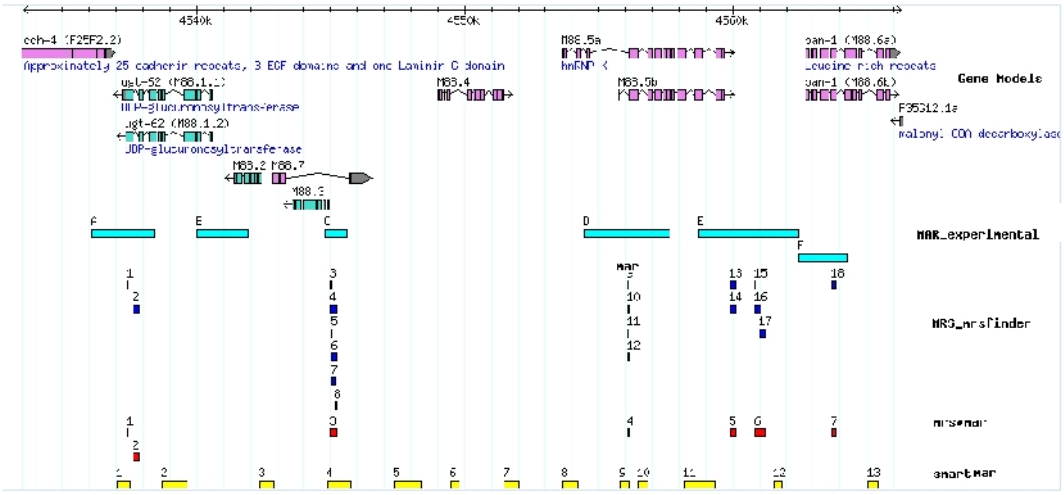


Figure 3.12 Graphical illustration of MRS, experimentally defined MAR, MRS-MAR, SMART-MAR and gene models in C. elegans cosmid M88.

3.4.3 MRS relationship with genes

MRS genes - Gene Ontology

To find out if MRS-genes are enriched (or depleted) for particular functional annotations, an analysis was carried out to compare the GO terms associated with MRS-genes to GO terms associated with non MRS-genes. As described above two MRS-gene sets, ‘close stop’ and ‘1k start or stop’, were analysed using the full set of *C. elegans* GO annotations and the IEA-only annotation set. Other MRS-gene sets were also analysed but the general pattern of enriched GO terms did not differ to those found for the MRS-gene sets presented in detail here.

Figures 3.13a-d show the log odds ratios and associated p-values for the most common GOslim terms for each MRS-gene set and annotation set analysis. The log odds ratio is a measure of the degree of enrichment of a particular term in the MRS-gene set compared to all the genes not in that set. The log odds ratio for each term has been assigned a p-value indicating the likelihood of the odds ratio occurring by chance. The GO terms considered statistically significant after applying the multiple testing correction are displayed in red font.

The ‘1k start or stop’ MRS-gene set was largely unaffected by the type of annotation set used (full or IEA only), both sets returned near identical top GOslim terms. The two annotation datasets also gave similar results for ‘close stop’ MRS-genes, with some notable differences. Only when analysed with the 'full' annotation dataset did the terms ‘embryonic development’ and ‘regulation of biological process’ appear for the ‘close stop’ MRS-gene set. Both these terms can arise through annotation resulting from embryonic lethal phenotypes produced by RNAi screens. Therefore, the ‘close stop’ MRS-gene set does appear to be susceptible to the annotation bias discussed by Vavouri *et al*. [13]. This may be because the ‘close stop’ MRS-gene set is a relatively small dataset (603 genes with full GOslim annotation, compared to 3317 for the ‘1k start or stop’ dataset).

A common feature of all the analyses shown in Figures 3.13a-d was the prominence of the ‘receptor activity’ annotation term. It remained significant after multiple testing correction in each case, except for ‘close stop’ MRS-genes when analysed with the full annotation set, where it was second to ‘embryonic development’. No other GOslim terms were found to be statistically significantly over-represented in any of the datasets. However, several other terms had log odds ratios indicating some degree of enrichment. GOslim terms for functions, processes and locations relating to gene expression rank highly across the MRS-gene sets. The term 'transcription factor activity' was highly ranked in all the datasets. In the close-stop MRS-gene sets the biological process terms 'translation' and transcription' were highly ranked and the cellular component term 'ribosome' was ranked behind 'receptor activity'. In the ‘1k start or stop’ MRS-gene sets the molecular function 'nucleic acid binding’ was highly ranked. In this MRS-gene-set, the molecular functions 'lipid binding' and 'ion channel activity' were among the top terms.

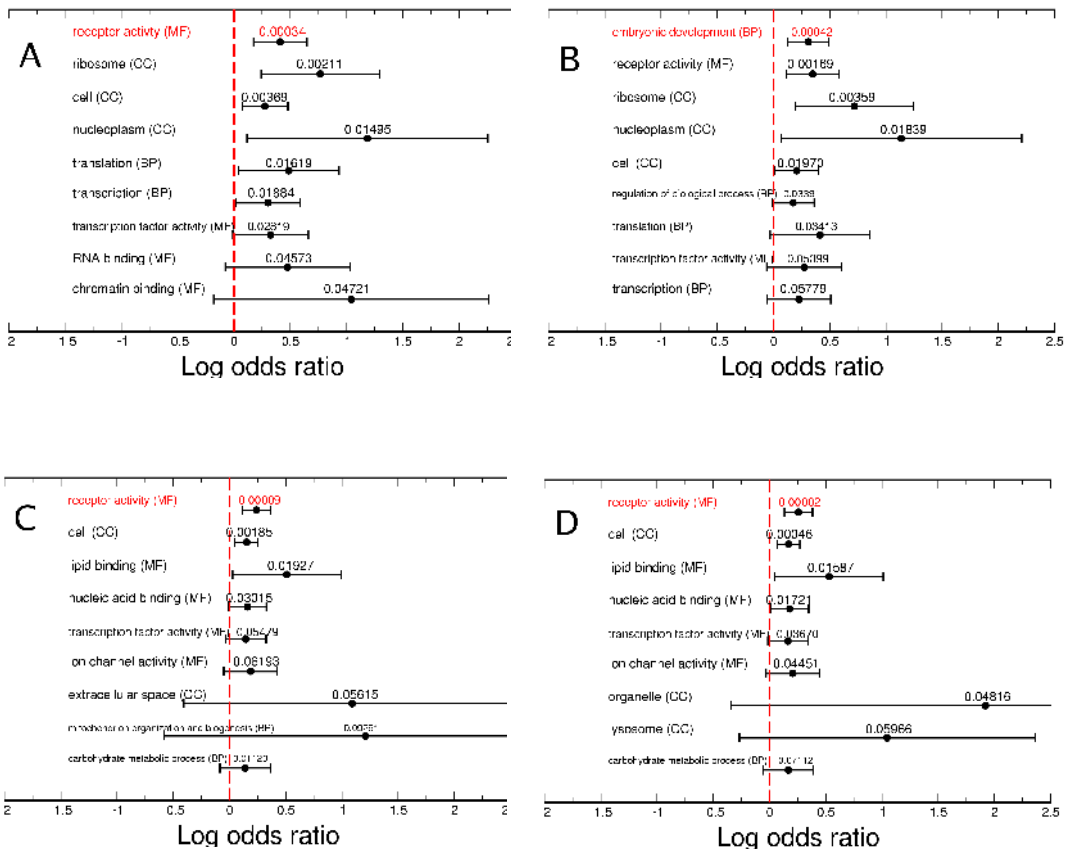


Figure 3.13 GO term enrichment in MRS-genes

The log odds ratios and 95% confidence intervals (two-tailed test) for the top most over-represented GO slim terms for 4 MRS-gene and annotation set combinations: (A) 'close stop' IEA-only, (B) 'close stop' full, (C) '1k start or stop' IEA-only, (D) '1k start or stop' full. The GO terms are split into three ontologies; cellular component (CC), biological process (BP) and molecular function (MF). The number above the bar represents the *p* value. Significant p-values are shown in red.

MRS genes - Position in operons

The following analysis set out to investigate any relationship between MRS-genes and operons. Previously discussed results have shown MRS to be particularly abundant in the regions immediately flanking genes. This has lead to the suggestion that MRS may be involved in transcription of the nearby genes. Operons consist of multiple genes that are initially transcribed into one polycistronic pre-mRNA. If the increase in MRS frequency is related to transcription then it is possible that operons are flanked by a similar pattern of increased MRS frequency, while internal operon genes lack this pattern.

One way of testing this hypothesis is to look at the relative occurrence of MRS-genes in different operon positions. Two specific questions about the relationship between genes-genes and operons were formulated:

i)Are MRS-genes more (or less) likely than non-MRS-genes to be found in operon start, stop or internal positions?

ii) Are MRS-genes in many-gene operons (i.e. operons with > 2 genes) more likely to be found in internal or external (i.e. start or stop) positions?

In answering both questions three types of MRS-genes were considered: ‘close start’, ‘close stop’, and ‘1k start or stop’.

The degree of association between the MRS-gene sets and their location in one of three operon positions (start, stop, internal) was measured by calculating the log odds ratio (Table 3.5). A positive log odds ratio would indicate a positive association: MRS-genes are relatively enriched in that position. Conversely, a negative log odds ratio would indicate MRS-gene depletion in that position and a ratio of zero indicates there is no association between MRS-genes and that operon position. As two-gene operons have no internal genes, MRS-gene incidence in two-gene and many-gene operons was analysed separately, with a further analysis of start, stop and total external positions for all operons.

		`1k start or stop'				`close start'				`close stop'			
	number of MRS-genes	6124	log OR	95% conf limit +/-	prob	1078	log OR	95% conf limit +/-	prob	1057	log OR	95% conf limit +/-	prob
2 gene operons	Number of 2 gene operons	675				675				675			
	MRS-genes in start pos	208	0.012	0.17	0.442	39	0.078	0.33	0.320	40	0.128	0.33	0.221
	MRS-genes in stop pos	204	-0.016	0.17	0.425	43	0.186	0.32	0.123	38	0.071	0.33	0.337
>2 gene operons	Number of >2 operons	378				378				378			
	Total genes in internal pos	612				612				612			
	MRS-genes in start pos	109	-0.084	0.22	0.231	17	-0.192	0.49	0.222	25	0.246	0.41	0.120
	MRS-genes in stop pos	119	0.044	0.22	0.347	24	0.180	0.42	0.200	23	0.015	0.43	0.239
	MRS-genes in internal pos	199	0.093	0.17	0.143	41	0.242	0.32	0.071	37	0.149	0.34	0.193
All operons	Number of all operons	1053				1053				1053			
	MRS-genes in start pos	317	-0.023	0.13	0.371	56	-0.013	0.28	0.464	65	0.177	0.26	0.090
	MRS-genes in stop pos	323	0.006	0.13	0.467	67	0.189	0.26	0.073	61	0.105	0.27	0.220
	MRS-genes in external pos	640	-0.009	0.10	0.428	123	0.098	0.19	0.161	126	0.150	0.19	0.062

Table 3.5 Association between MRS-genes and position in operon

In general there were no consistent patterns between the MRS-gene and operon combinations studied. Most of the log odds scores were quite low (indicating no strong association) and none were statistically significant at the 5% level. Even in the case of the strongest association scores (log odds of 0.246 and 0.242), the 95% confidence interval spanned a negative association (confidence limits of 0.41 and 0.32, respectively). Overall, these data offered no suggestion that MRS-genes were more likely to occur in any particular position in an operon than non MRS-genes.

To answer the second question, the log odds ratio of MRS-genes occurring in internal (positive log odds) or external (negative log odds) positions was calculated for each of the three MRS-gene sets (Table 3.6). Here, a positive log odds score would indicate MRS-genes were more likely to be found in external positions in operons, and a negative score that they were more likely to be found in internal positions. Again the log odds scores were close to zero and none of them were considered statistically significant at the 5% level. 'Close start' MRS-genes had an appreciable negative log odds score (-0.22), suggesting MRS-genes were more likely in internal than external positions but the confidence interval was large (+/- 0.45). The log odds ratio for 'close stop' MRS-genes was very small and offered no evidence of MRS-genes being more prevalent in internal or external positions.

Considering the results of both sets of analyses as a whole, there is little or no evidence to support the hypothesis of increased MRS frequency at operon boundaries. The MRS-gene and operon position pairings generally showed only small associations that were not statistically significant.

MRS-gene set	‘1k start or stop’	‘close start’	‘close stop’
>2 gene operons	378	378	378
Total genes in external positions	756	756	756
Total genes in internal positions	612	612	612
MRS-genes in external positions	228	41	48
MRS-genes in internal positions	199	41	37
log odds ratio	-0.11	-0.22	0.05
95% confidence interval +/-	0.23	0.45	0.44
p-value for log OR	0.175	0.162	0.409

Table 3.6 Likelihood of MRS-genes in internal and external positions of many-gene operons

MRS genes - correlation with spliced leader acceptor sites

Spliced leader (SL) acceptor sites tend to occur just upstream of translation start sites, in a similar location to the upstream spike in average MRS frequency. The following analysis was carried out to determine if there was a relationship between genes with an MRS and genes with a SL acceptor site. About half of the annotated SL acceptor sites have been assigned to a gene, as summarised in Table 3.3. Some genes had multiple SL acceptor sites assigned to them but this was very rare. The proportion of genes in five MRS-gene sets with an SL acceptor site was compared to the proportion of genes with an SL acceptor site in the remaining non-MRS genes (Table 3.7). The degree of association between MRS-gene sets and SL acceptor sites is measured by the log odds ratio, summarised for each set in Table 3.8. Strong association between MRS-genes and SL acceptor site status is marked by a large and positive log odds ratio.

MRS-gene set	Genes with SL acceptor site / genes without
'close stop'	255 / 802
(non 'close stop')	4557 / 14438
'close start'	274 / 804
(non 'close start')	4538 / 14436
'1k stop'	822 / 2585
(non '1k stop')	3990 / 12655
'1k start'	807 / 2533
(non '1k start')	4005 / 12707
'1k start or stop'	1495 / 4629
(non '1k start or stop')	3317 / 10611

Table 3.7 Number of MRS-genes and remaining 'non' MRS-genes with and without SL acceptor site.

MRS-gene set	Log odds ratio	95% confidence +/-	p-value
'close stop'	0.007	0.143	0.181
'close start'	0.081	0.141	0.404
'1k stop'	0.008	0.086	0.423
'1k start'	0.011	0.087	0.131
'1k start or stop'	0.033	0.07	0.460

Table 3.8 Log odds ratios for association between MRS-genes and SL acceptor sites.

All the MRS-gene sets had a positive log odds ratio, meaning a higher proportion of MRS-genes had a SL acceptor site compared to the corresponding non-MRS gene sets. However, the log odds ratios were very low for all MRS-gene sets, indicating very weak association with SL acceptor sites. Furthermore, the p-values obtained indicated that none of the associations calculated were significant. These data show there is no strong association between MRS-genes and SL acceptor sites, although the possibility of a very weak positive correlation cannot be excluded.

As mentioned in the methods section, it appears that the annotation of SL acceptor site to genes is far from complete. However, we may assume that the available SL acceptor site annotation is a random sample of the complete set of SL acceptor site locations. In this case, the proportions of MRS- and non MRS-genes with an SL acceptor site would not change and the conclusion of no strong positive relationship would hold. On the other hand, it is possible that there is a bias in the genes that are annotated with a SL acceptor site, in such a way that MRS-genes are over or under-represented in the current SL acceptor annotations. Therefore, due to the incompleteness of the SL acceptor site annotation, we cannot rule out a significant relationship between MRS-genes and SL acceptor sites.

MRS-genes - expression levels

The prospect that MRS may be involved in the transcription and hence regulation of expression of genes has been raised above. The impact of MRS on gene expression was investigated by comparing Serial Analysis of Gene Expression (SAGE) data for MRS-genes with non-MRS-genes. The average expression level of *C. elegans* genes, as determined by levels of SAGE tags, was calculated for 43 SAGE libraries. The mean expression levels (calculated as the number of SAGE tags normalised to a sample size of 100,000 tags per library) of all the genes in two MRS-gene sets, 'close stop' and '1k start or stop', was compared to their respective non-MRS genes using a t-test (Tables 3.9 and 3.10).

After correction for multiple testing, neither set of MRS-genes had a mean

expression statistically significantly different to non-MRS-genes for any individual SAGE library. However, when mean expression levels across all libraries were considered, a significant pattern emerged. In 'close stop' MRS-genes, the expression level was higher than non MRS-genes in 30 of 43 SAGE libraries (chi squared = 6.721, p = 0.0095) and the overall mean expression level across all SAGE libraries was 8.06 for ‘close stop’ MRS-genes compared to 7.04 for non MRS-genes (two sample t-test p= 0.0499). Therefore, despite 'close stop' MRS-genes not having a statistically significant higher expression level for any single library, the expression level across all libraries is higher than for non MRS-genes. This expression pattern is not repeated in '1k start or stop' MRS-genes where expression levels were higher for just 16 of 43 libraries (chi squared = 2.814, p = 0.0934) and the mean expression level across all libraries was slightly lower (7.05 / 7.40 two sample t-test p= 0.4098).

SAGE library	Number of scoring tags assigned		mean expression level		t-test p-val
	'close stop' MRS-genes (total 1057)	Non MRS-genes (total 18995)	'close stop' MRS-genes	Non MRS-genes	
2week_dauer	417	6937	10.74	6.58	0.261
afd_nuerons	572	9746	4.27	4.54	0.638
aser_nuerons	502	8594	3.83	4.55	**0.080**
ciliated_nuerons	413	6998	5.17	6.61	0.232
develpomental_SWL12	392	6630	10.38	7.86	0.435
develpomental_SWL21	475	7928	6.84	6.15	0.661
develpomental_SWL32	497	7530	8.27	6.66	0.352
develpomental_SWL41	525	9009	5.64	5.32	0.759
develpomental_SWN21	487	8382	6.35	5.54	0.382
develpomental_SWYA1	419	6771	8.28	7.33	0.595
dissected_gonad	351	5673	9.65	8.46	0.565
embryonic_SW023	343	5924	5.99	7.60	0.258
embryonic_SW028	363	5993	6.67	8.61	**0.020**
embryonic_SW030	420	7096	9.19	7.39	0.246
embryonic_SW031	358	6537	9.60	7.76	0.313
embryonic_SW032	325	5558	7.86	7.82	0.973
embryonic_SW033	463	7747	7.75	6.50	0.363
embryonic_SW034	437	7816	5.25	5.39	0.851
embryonic_SW035	461	7905	5.21	5.34	0.836
embryonic_SW037	438	7393	4.24	5.13	**0.058**
embryonic_SW038	473	8061	4.05	4.92	0.295
embryonic_SW039	235	4877	7.97	7.22	0.64
embryonic_SWEG1	304	5160	13.90	9.02	0.147
embryonic_SWEM1	272	4583	13.07	10.78	0.402
embryonic_SWN22	480	8169	7.00	6.01	0.343
fer-15_1day	407	6688	8.88	6.86	0.407
fer-15_6day	416	6979	8.88	7.33	0.343
fer-15_daf2_10day	381	5994	9.49	8.16	0.520
fer-15_daf2_1day	330	5383	9.08	8.10	0.636
fer-15_daf2_6day	287	4300	12.69	10.05	0.440
glp4_adult	403	6708	11.00	7.83	0.197
glp4_gut	324	4692	13.31	11.42	0.621
gut_cells	325	5621	13.07	8.37	0.136
hypodermal_cells	471	8147	8.31	6.57	0.210
mixed_stage_wQual	604	10486	5.73	4.51	0.233
muscle_cells	290	4871	12.38	10.26	0.409
pannueral_cells	393	6502	6.24	8.02	**0.022**
pharyngealGland_cells	285	5818	6.75	6.21	0.638
pharyngealMarginal_cells	548	9511	4.54	4.61	0.900
pharynx_cells	527	9162	6.93	5.65	0.290
punc4_cells	633	10736	3.28	3.94	**0.035**
purified_oocyes	388	5711	7.59	7.63	0.970
young_dauer	367	5755	11.35	8.02	0.220
mean	413.98	6978.63	8.06	7.04	
standard deviation			2.87	1.75	

Table 3.9 Gene expression levels for 'close stop' MRS-genes

SAGE library	Number of scoring tags assigned		mean expression level		t-test p-val
	'1k start or stop' MRS-genes (total 6124)	Non MRS-genes (total 13928)	'1k start or stop' MRS-genes	Non MRS-genes	
2week_dauer	2319	5035	6.54	6.94	0.712
afd_nuerons	3283	7035	4.53	4.52	0.963
aser_nuerons	2838	6258	13.10	12.93	0.631
ciliated_nuerons	2290	5121	5.45	7.01	0.182
develpomental_SWL12	2225	4797	7.30	8.33	0.304
develpomental_SWL21	2626	5777	6.08	6.23	0.814
develpomental_SWL32	2625	5402	6.72	6.78	0.924
develpomental_SWL41	2936	6598	5.09	5.45	0.503
develpomental_SWN21	2788	6081	5.46	5.63	0.704
develpomental_SWYA1	2232	4958	6.79	7.65	0.234
dissected_gonad	1857	4167	7.15	9.14	**0.036**
embryonic_SW023	1920	4347	6.30	8.05	0.199
embryonic_SW028	1989	4367	7.88	8.78	0.244
embryonic_SW030	2329	5187	7.55	7.46	0.876
embryonic_SW031	2110	4785	7.99	7.80	0.759
embryonic_SW032	1860	4023	7.91	7.78	0.768
embryonic_SW033	2607	5603	6.52	6.59	0.881
embryonic_SW034	2590	5663	5.46	5.35	0.770
embryonic_SW035	2626	5740	5.21	5.38	0.579
embryonic_SW037	2457	5374	5.17	5.04	0.699
embryonic_SW038	2729	5805	4.62	4.99	0.291
embryonic_SW039	1504	3608	6.38	7.62	**0.043**
embryonic_SWEG1	1752	3712	9.47	9.20	0.754
embryonic_SWEM1	1493	3362	10.93	10.90	0.972
embryonic_SWN22	2693	5956	6.40	5.91	0.159
fer-15_1day	2187	4908	6.87	7.03	0.842
fer-15_6day	2343	5052	7.30	7.47	0.786
fer-15_daf2_10day	1987	4388	8.08	8.31	0.823
fer-15_daf2_1day	1775	3938	7.44	8.48	0.319
fer-15_daf2_6day	1486	3101	10.87	9.90	0.654
glp4_adult	2299	4812	8.28	7.88	0.629
glp4_gut	1663	3353	10.25	12.18	0.187
gut_cells	1898	4048	8.83	8.53	0.71
hypodermal_cells	2683	5935	6.69	6.65	0.934
mixed_stage_wQual	3493	7597	4.13	4.78	**0.041**
muscle_cells	1588	3573	10.39	10.37	0.981
pannueral_cells	2164	4731	7.32	8.18	0.228
pharyngealGland_cells	1800	4303	5.50	6.55	**0.041**
pharyngealMarginal_cells	3194	6865	4.46	4.67	0.420
pharynx_cells	3092	6597	5.66	5.75	0.813
punc4_cells	3626	7743	3.73	3.99	0.315
purified_oocyes	1916	4133	7.71	7.60	0.799
young_dauer	1949	4173	7.50	8.56	0.174
mean	2321.42	5070.02	7.047	7.403	
standard deviation			1.996	1.989	

Table 3.10 Gene expression levels for '1k start or stop' MRS-genes

3.5 Discussion

3.5.1 Analysis of the MRS

The degeneracy, variable size and bipartite nature of the MRS combine to make it a complex structure. The analyses carried out in this chapter have sought to identify peculiar characteristics that may influence the study of the frequency and distribution of the MRS. Despite the large potential for overlap between the two constituent motifs of the MRS, 70% of MRS in *C. elegans* do not overlap. Therefore, in most MRS the two motifs are distinct and both may be involved in functional interactions. The MRS size profile shows a large number of MRS to be longer than the 50 bp window length used to describe the frequency of MRS near genes (Chapters 2 and 4). This means that for about half of the MRS, both motifs will lie in a different 50 bp window to the one containing the MRS mid-point. However, it is inevitable that a certain number of MRS will overlap the window boundaries, with the mid-point in one window and a motif in another, irrespective of window size. Therefore, a window size of 50 bp is a good compromise between accurate inclusion of MRS and resolution of MRS frequency variation.

The main finding of the analysis of the distribution of MRS motif 1 and 2 was that motif 1 was far more abundant in the *C. elegans* genome than motif 2. However, the frequency of motif 2 was also significantly greater than that of the complete MRS. Additionally, the distribution of both motifs is similar to the complete MRS. This suggests that no single motif is responsible for the occurrence of MRS. Furthermore, as both motifs appear to be required to generate the pattern of MRS distribution, both motifs could be potentially functional. The effect of randomising the genomic sequence was found to be the same for motif 1 and 2 as for MRS; a small increase in frequency under the 2 Mb randomisation protocol and a greater increase in frequency under the 10 bp randomisation protocol. However, motif 1 varied more in terms of frequency and distribution compared to real sequence. This indicates that if the complex MRS motif is under selection, this may act through selection against motif 1

to cause a reduction in MRS frequency. In summary, although motif 1 is by far the most abundant both motifs are likely to be important to the occurrence and potential function of the MRS.

The complexity of MRS made prediction of the frequency of randomised MRS in the *C. elegans* genome difficult. Empirical investigation revealed that the number and distribution of randomised MRS is similar to that observed for (actual) MRS in sequence randomised in 10 bp sections. Furthermore, using the information in Figure 2.7 (MRS in sequence randomised in different section lengths) we can estimate that the number of MRS in sequence randomised in sections equal in length to the number of specified nucleotides in the MRS would match the number of randomised MRS even more closely. The spacing between MRS in the *C. elegans* chromosomes was also shown to differ from a random simulation of the spacing between the same number of elements. Together with the findings made in the previous chapter regarding MRS frequency in randomised sequence, these data provide the consistent observation that the incidence of MRS in *C. elegans* differs from a random pattern.

3.5.2 MRS-MAR

MRS-MAR were inferred as a means of using the MRS to more accurately represent MAR. Closely apposed MRS were merged with the intention that the number and boundaries of the MRS-MAR thus created would be as close as possible to actual MAR. The rationale behind this procedure was that MRS that were situated very close to each other were likely, biologically, to be in single MAR. However, MAR have been reported as having a wide range of sizes, from several hundred base pairs to several kilobase pairs and the spacing between MAR can be highly variable. The MRS spacing parameter used for MRS-MAR is therefore a best estimate compromise. Furthermore, even when they are combined, the boundaries of (multiple) MRS will not necessarily match the boundaries of the MAR they may signify. Although these factors may limit the usefulness of MRS-MAR as predictions of actual MAR, they are an improvement on MRS alone.

One of the most useful aspects of MRS-MAR is that, except for the limitations described above, they allow for more meaningful comparison with alternative MAR prediction methods. When compared with SMART-MAR, a significant degree of discrepancy between the two sets of inferred elements was observed: There were many more SMART-MAR, and SMART-MAR were evenly distributed, both along chromosomes and in the experimentally investigated cosmid M88. It is not surprising therefore that the two features overlapped only slightly more than would be expected of two sets of random annotations of that size. Therefore, it would appear that one or both of the methods are not predicting MAR effectively. The published specificity (68%) and sensitivity (38%) of SMARTest suggests that SMARTest actually under-predicts the number of MAR. However, the values for SMARTest specificity and sensitivity are based largely on estimates from mammalian genomes, and it may perform differently in *C. elegans.* For example, it has been estimated that in the human genome there are 100,000 MAR [26]. If the *C.elegans* genome has the same density of elements then we would expect about 3300 MAR. This estimate is far lower than the numbers identified by SMARTest or MRS incidence and the number of experimentally defined MAR found in cosmid M88 also suggests a higher overall genome incidence. Without more extensive data for experimentally predicted MAR it is difficult to determine which method gives the more accurate prediction of MAR number in *C. elegans*.

One possibility is that the MRS (and MRS-MAR) and SMARTest identify different types of element, which may or may not be true biologically active MAR. In their original description of the MRS, van Drunen *et al.* suggest that, as the MRS did not faithfully predict all experimentally defined MAR, it may be representative of only a subset of MAR [42]. The evenly spaced distribution of SMARTest predicted MAR (see chromosome distribution Figure 3.11 and cosmid M88, Figure 3.12) could be indicative of a structural element that provides constitutive stability to the genome. Conversely, MAR represented by MRS could be the sites of more transient attachment to the matrix and therefore play a more direct role in gene expression.

3.5.3 MRS relationship with genes

If the MRS signature defines a functional DNA element, then one possible function might be modulation of expression of the genes that lie in close proximity to them. This possibility was tested by analysing MRS-genes to determine if they had a common quality that distinguished them from other genes. Several features were studied: functional classification of protein product (GO terms), position in operons, presence of a spliced leader acceptor site, and the expression patterns of the genes (as determined by SAGE). If the presence of an MRS is associated with a specific function or characteristic then this would be good evidence that the MRS is functional.

Both the ‘close-stop’ and ‘1k start or stop’ MRS-gene sets were significantly enriched for the molecular function GO term ‘receptor activity’. Although this term was common in the MRS-gene sets in comparison to all remaining genes, only ~15% of MRS-genes were annotated with this term. But as only about half of the genes in the MRS-gene sets have a GO annotation, the true number of MRS-genes involved in receptor activity could be higher. Closer inspection of the MRS-genes annotated with receptor activity could reveal that they have a more specific function in common. For example, they may belong to one class of receptors or form an equivalent component of different receptors. Closer definition of the function of these gene products would also help clarify the role of the MRS in relation to these genes. If the genes all act in one part of the cell, then the MRS may be involved post-transcriptionally in directing the mRNA to specific location for translation. It could be that MRS are involved in expression of genes that form a protein component most commonly found in receptors, but also found elsewhere. This would explain why not all MRS-genes have receptor activity annotation. Alternatively, MRS could be *cis*-regulatory sites that control when the receptor activity gene is expressed. Studying the individual MRS-genes with 'receptor activity' annotation could show if they have any peculiar characteristics, such as a high number of MRS or MRS in a precise location in relation to the gene. If the MRS-genes could be narrowed down in this way it would

explain why not all MRS-genes in the current set have the 'receptor activity' annotation.

There was a bias to annotations resulting from RNAi lethal phenotypes shown in ‘close stop’ MRS-genes, a situation that has been reported by other researchers [13]. This issue was avoided by restricting the analysis to IEA annotations but it highlights the susceptibility of GO annotation analysis to the quality and completeness of the annotation data. This susceptibility was also evident after preliminary research showed that use of GO gene annotation files produced on different dates generated different patterns of over-representations. Furthermore, while numerous on-line GO analysis tools were tested during the course of this work, the results were far from consistent.

The rationale for studying MRS genes in relation to operons was that if MRS were transcriptionally active, then this should be reflected in their position in operons. This issue was approached from two angles. Firstly, the likelihood of MRS-genes appearing in certain operon positions compared to non-MRS-genes was studied. Secondly, the likelihood of MRS-genes appearing in external or internal operon positions was studied. In general, most of the MRS-gene operon position comparisons produced non-significant results. The data therefore offered no evidence in support of the hypothesis that MRS-genes have a non-random distribution in operons. It is possible that if the positioning of actual MRS around operons was studied (instead of MRS-genes) then there may be an observable effect. However, based on the data presented here it appears MRS have no specific relationship with operons.

The similar positioning of the spike in MRS frequency reported in chapter 2 and SL acceptor sites motivated analysis of genes with both these characteristics. Across most of the MRS-gene sets studied, a positive association between MRS-genes and the presence of an SL acceptor site was observed. However, the magnitude of the effect was quite small and they were not statistically significant. It is possible that the MRS acts post-transcriptionally as a signal to direct the pre-mRNA to a spliced

leader processing site. If the MRS does act post-transcriptionally, this should be reflected by a strand bias for the MRS motifs. Alternatively, the MRS may encourage the spliced leader spliceosome to form in the correct location by preventing it assembling at the terminal end of the RNA strand.

Many of the possible functions of the MRS that are raised above will ultimately affect the expression of the associated MRS-genes. Examination of the valuable SAGE resource available for *C. elegans* was used to investigate differences in gene expression between MRS-genes and non-MRS-genes. No statistically significant differences in expression level between MRS-genes and non-MRS-genes were observed for individual SAGE libraries. However, 'close stop' MRS-genes had higher expression for most libraries and the overall expression level was higher than that of non-MRS-genes. This provides an indication that the presence of MRS near to CDS stop sites is related to an increase in gene expression. However, there are several complex aspects to the SAGE data that make it difficult for firm conclusions to be drawn. Firstly, the mean tag counts for the gene sets studied here are low (~10/100,000) compared to some individual genes (e.g. genes encoding ribosomal proteins) that may have tag counts of several thousand. This means that individual genes can have a large influence on the mean count for a library. Secondly, in most libraries the SAGE tags represented only 35-40 % of the genes (from either set) and of the genes that were represented most had a tag count of one. Tag counts of zero were not included in the calculation of mean expression which may create too great a distinction between very low expression (a tag count of one) and no (detectable) expression (a tag count of zero). Finally, although the overall trend is for 'close stop' MRS-genes to have higher expression than non-MRS-genes, for the top five libraries with the greatest difference in expression levels MRS-genes actually have lower mean expression levels. This conflict emphasises the complexity of the SAGE expression data.

The analysis of MRS-genes revealed that the enrichment of the GOslim term 'receptor activity' was common to all types of MRS-genes studied. In contrast, no

significant relationship between any of the MRS-gene sets and position in operons, or presence of SL acceptor sites was found. One explanation for this is that the MRS is involved in multiple functional pathways, dependant on its position relative to a gene. Alternatively, these results may reflect the need to refine the definition of MRS-genes. In doing so, more robust relationships for MRS-genes may be identified.

Chapter 4 MRS in Animals and Plants

4.1 Abstract

This final results chapter comprises two main sections. The first section concerns the precise positioning of the previously observed peaks of MRS frequency that flank *C. elegans* genes. The downstream MRS frequency peak was found to be clearly located at the transcript stop site. The location of the upstream MRS frequency peak was less well defined but likely occurs in relation to the CDS start site, rather than transcript start site. In the second section, the MRS frequency around genes of six diverse species was considered. A peak of MRS frequency in the region of the transcript stop site was observed in most species. The AT content around genes was also shown to vary at transcript boundaries. However, overall there was only a partial correlation between AT content and MRS frequency.

4.2 Introduction

The work described in this chapter falls into two broad sections. The first section describes efforts to precisely locate the previously described peaks in MRS frequency surrounding *C. elegans* genes with reference to transcript and coding sequence boundaries. In the second section, these characteristics of MRS frequency surrounding *C. elegans* genes are compared to those found in a diverse set of taxa spanning the animal and plant kingdoms.

4.2.1 MRS around *C. elegans* genes

In the previous chapters, the frequency of MRS around the coding regions of genes was studied. This work established that there is a relationship between MRS occurrence and genes, namely that peaks of average MRS frequency occur just upstream and downstream of CDS in *C. elegans,* and that *C. briggsae* only has a peak just downstream of CDS. Coding regions were selected as the point of reference as

the universal start and stop codons mean they can be precisely and accurately positioned in the genome with little ambiguity. However, the previous work did not establish if the peaks of average MRS frequency were located relative to the CDS start and stop positions themselves or to other, related genomic features, such as transcript start and stop positions. The first part of this chapter makes use of transcript (or 'gene') annotation with the aim of resolving this uncertainty. In comparing the frequency of MRS around both transcript and CDS sequences the question of whether the peaks in MRS frequency are located with respect to CDS boundaries or transcript boundaries is answered.

As a consequence of extending the analysis to investigate gene transcript data, it was also appropriate to consider non-protein coding genes. The emphasis on (protein-) coding sequences in the work described up to this point has meant that the various RNA genes, non-coding RNAs and pseudogenes have been neglected. The annotation of these types of features in *C. elegans* is comparatively good. Therefore the presence or absence of MRS frequency variation in the vicinity of non-protein coding genes in *C. elegans* was used as the basis for determining the scope of MRS analysis in the other species studied here. Another factor explored in *C. elegans* to focus analysis of the other species is the effect of multiple gene models. As a result of alternative splicing, many genes in *C. elegans* are associated with multiple transcripts, some of which encode different proteins while others vary only in their UTR. Therefore, a single protein may be represented by multiple gene models (transcripts). In this chapter, *C. elegans* is used as a model to test the validity of reducing this one to many relationship to one to one by randomly selecting a representative gene model for each protein.

A main aim of the first part of this chapter is to identify where, in respect to genes, the increase in MRS frequency occurs. This attempt to reconcile the positioning of the peaks relative to transcript and CDS boundaries necessitated an investigation of the relative positioning between transcript and CDS start and stop sites. Again, the principles established through study of *C. elegans* were extended to the other species

studied in the second part of this chapter.

4.2.2 Comparing MRS around genes of six diverse species

So far this thesis has been concerned solely with the species in which the MRS was first described, *C. elegans.* However, the original motivation for the MRS was the identification of matrix attachment regions and these have been identified in a large number of species across the animal and plant kingdoms. Furthermore, regardless of the efficacy of the MRS as a predictor of matrix attachment regions, MRS in *C. elegans* have been shown to have a close relationship with genes, many characteristics of which are shared across virtually all forms of life. It is appropriate then that in this final part of this thesis the study of MRS frequency around genes is extended to another five species: *Caenorhabditis briggsae, Arabidopsis thaliana, Danio rerio, Drosophila melanogaster* and *Homo sapiens.*

These five species have been chosen based partly on the availability of good quality annotation and their wide use as model organisms. But they also provide a wide range of genome sizes and gene densities over which comparisons with *C. elegans* and each other can be made. *C. briggsae* is a close relative of *C. elegans,* and they have much in common, such as the same number of chromosomes and similar genome size and gene density. The main model plant, *A. thaliana,* has a slightly larger genome than the nematodes comprising five chromosomes and over 27,000 genes. By contrast, although the genome the fruit fly *D. melanogaster* is larger still (~180 Mb) it only has just over 14,000 genes. The zebrafish, *D. rerio,* has a much larger genome, approximately 1.5 Gb over 25 chromosomes, containing about 19,000 genes. However, the largest genome (and lowest gene density) studied here is the 3 Gb human genome with approximately 22,000 genes. This phylogenetically diverse group of organisms allows inferences to be drawn about the origins of MRS peaks at genes and perhaps insight into their function.

Several studies have previously reported on distinctive nucleotide changes around genes primarily focused on transcription boundaries [74-77]. Another study by

Mizuno and Kanehisa investigated GC content around the translation initiation sites across a range of species [99]. The window size employed by Mizuno and Kanehisa was too large to reveal sharp changes in nucleotide content, however a transition at the translation initiation site from low to high GC content was observed [99]. With regards to nucleotide content at transcription boundaries, *H. sapiens* have been reported to have a gradual drop in AT at transcription start and a peak in AT content at the transcription stop [74-77]. A similar pattern has also been observed in other mammals and chicken [74, 77]. The pattern of nucleotide frequencies around genes in fish and invertebrates is more complex. Zhang *et al.* report the AT content around *D. melanogaster t*ranscription boundaries as a rise then fall at the transcription start and the reverse pattern at the transcription stop [77]. In a study of individual nucleotide frequencies at transcription start sites, Aerts *et al.* observed various combination of peaks and troughs of all four nucleotides in *C. elegans, C. briggsae, D. rerio* and *D. melanogaster* among other species [74]. Aerts *et al.* established that the changes in nucleotide composition at transcription start sites are correlated to CpG nucleotide frequencies and gene expression in humans but not *D. melanogaster* [74]. The sharp changes in nucleotide content at transcription boundaries have also been described as genomic punctuation and as a consequence of the presence of regulatory elements [75, 77].

4.3 Methods

Some of the methods used in the work in this chapter have already been described in the earlier chapters. The following sections largely describe methods not previously covered.

4.3.1 Data collection

C. elegans

Previous analyses in this thesis used data from the WS150 release of WormBase. However, this version of the data was not available through either WormMart or BioMart through Ensembl, so data from the most recent available data freeze, version

WS180 was used for analyses in this chapter. The synchronous raw genome sequence files for version WS180 were downloaded from the WormBase ftp server [98]. Annotations were collected for six transcript and CDS sets:

(i) CDS single - The CDS (coding sequence) starts with the translation initiation codon “ATG” and ends in a termination codon. In this set each gene is represented by a single coding sequence that stretches from the start of the most upstream protein coding exon to the most downstream protein coding exon of all annotated transcripts.

(ii) CDS multiple - Coding sequence as above. All transcripts of a gene are included in this set, providing they encode a protein sequence not encoded by another transcript. Multiple transcripts that differ only in untranslated sequence are represented by a single (randomly chosen) member.

(iii) Transcript single coding - The transcript is defined as starting and ending at the annotated transcription start and end. In this set each gene is represented by the transcript start and stop that span the largest region. It includes protein coding genes only.

(iv) Transcript single all - As above, but all gene types (e.g RNA genes, pseudogenes) included.

(v) Transcript multiple (coding) - Transcript defined as transcript start and end as for Transcript single coding. In this set each annotated transcript coding for a unique protein from each gene is represented.

(vi) Non-coding genes - All non-protein coding genes.

Each member of these data-sets consisted of an identifier and genomic coordinates (chromosome or contig, strand, start position and end position).

To unify the process of gathering data across multiple species as far as possible, all annotations were collected from BioMart through Ensembl. Table 4.1 shows the procedure used to extract information from the Ensembl “genes” database of BioMart. For both CDS sets, the procedure described in Table 4.1 retrieves annotations for coding exons. These data were then processed by using the highest

and lowest exons positions for each gene/transcript identifier to create the CDS coordinates. The Non-coding gene set was generated by taking all the members from the Transcript single all set that were not members of the Transcript single coding set.

Other species

The transcript and CDS annotations collected for *C. briggsae, A. thaliana, D. rerio, D. melanogaster* and *H. sapiens* were limited to CDS single and Transcript single data sets. Where possible, data was gathered from Ensembl release 48 using the same procedure as for *C. elegans*. Data for the CB3 release of *C. briggsae* was obtained from WormMart, using essentially the same procedure as for data collection from BioMart. However, it was discovered that the "gene" annotation for *C. briggsae* was limited to describing the CDS. Therefore the analysis in this chapter is limited to CDS data for *C. briggsae.* Data for *A. thaliana* was obtained from The Arabidopsis Information Resource website. The transcript and CDS data sets were extracted from the TAIR 7 release GFF file using Perl scripts.

Gene set type	BioMart	
	Filters - Genetype	**Attributes**
CDS single	protein_coding	Ensembl Gene ID Strand Chromosome Coding start Coding end
CDS multiple	protein_coding	Ensembl Transcript ID Strand Chromosome Coding start Coding end
Transcript single - coding	protein_coding	Ensembl Gene ID Strand Chromosome Gene start Gene end
Transcript single - all	none	Ensembl Gene ID Strand Chromosome Gene start Gene end
Transcript multiple (coding)	protein_coding	Ensembl Transcript ID Strand Chromosome Transcript start Transcript end

Table 4.1 Description of how the various gene set types relate to BioMart filters.

Data issues

A number of data quality control procedures were implemented as part of the data processing described below. Through these a number of issues with the raw data came to light, over and above the expected number of data inaccuracies. These issues were discovered primarily in relation to *C. elegans* as it was the most intensively studied, although they may also be relevant to other species.

A review of *C. elegans* genes that had failed checks for a valid stop codon at the end of the CDS showed that, in some cases, the stop codon was annotated as being split by an intervening intron. As it was not clear whether this was a biological reality or an artefact of automated gene prediction, and as it affected only a relatively small number of genes (12 in *C. elegans)*, all affected genes were discarded from subsequent analysis.

Close inspection of the difference between CDS and transcript start positions revealed that in some genes an intron was annotated between the 5' UTR and the first coding exon. While this does not directly affect the results presented here, it is worth noting that the difference between 5' UTR start and CDS start may not equate to the length of the UTR.

For 3,111 genes on the Y chromosome of *H. sapiens* annotated as protein coding, no protein coding exons were found. These genes were therefore discounted from the protein coding genes set.

Finally, some problems were encountered when trying to match transcripts to genes. At least two pairs of *C. elegans* transcripts share the same promoter and 5' UTR, but have no protein coding exons in common. In the case of ZC416.8a and ZC416.8b, all the exons of ZC416.8a are contained in an intron of ZC416.8b. In another example, the mature transcripts of B0564.1b and B0546.1a are very similar, but the protein coding exons of each transcript are untranslated in the other transcript. Usually, all transcripts with the same root identifier (everything up to and including the first number after the period) are grouped under a single gene identifier. However, in both

these cases the transcripts have been assigned different gene identifiers, as there is no overlap in their protein sequences.

4.3.2 MRS and AT surrounding genes and CDS

A three stage process was used to determine the AT content and frequency of MRS in regions surrounding genes. First, the start and stop coordinates for the transcript or CDS set to be investigated and the corresponding raw genome sequence was obtained using the methods described above. Then, two FASTA format files, one containing 'start-region' and the other 'stop-region' sequence records for each transcript or CDS in the set, were created from the coordinate data using a Perl script. The 'start-region' represented sequence from 1200 bp upstream to 600 bp downstream of the start site and the 'stop-region' represented sequence from 600 bp upstream to 1200 bp downstream of the stop site. With the aid of MRSfinder, the mid-points of all the MRS in each of the sequence records were calculated. To account for variation in MRS size, only those MRS mid-points found after the first 200 bp and before the last 200 bp of each sequence record were used for analysis. Therefore the region over which MRS mid-points were analysed included 1000 bp outside the CDS/transcript and 400 bp inside it.

Various quality checking procedures were implemented in the Perl script used to carry out this sequence collection step. Sequences were rejected if they were too short (for example if the gene lay very close to the chromosome start), contained Ns, other ambiguous characters or the start/stop codon was not valid (for CDS only). In the final stage, the MRS frequency and AT content of the sequences in the two FASTA format files were determined. MRSfinder was used to calculate the number of MRS mid-points at each position from all the sequence records in the file. The frequency of MRS in non-overlapping 50 bp windows was then calculated. The AT content was expressed as a percentage of A and T bases in non-overlapping 10 bp windows.

4.3.3 Difference between gene and CDS start and stop annotations

The transcript start is defined here as being the annotated transcription start site and similarly the stop site is the annotated transcription stop site. The CDS start and stop sites are defined by the annotated translation start and stop positions. A simple Perl script was used to calculate the difference between the annotated CDS and transcript start and stop positions for each record in the single and multiple CDS/transcript sets described above.

Using these data the CDS and transcript sets were divided into sub-categories. Where there was no difference between the CDS and transcript start or stop sites the transcript and the associated CDS were assigned to the 'diff 0' category. The remaining transcripts and CDS were assigned to the 'non-diff 0' category. Further categories were devised to include transcripts and CDS with between 1 and 50 bp difference between start or stop sites ('diff 1-50'), 51 to 100 bp difference ('diff 51-100') and 101 to 300 bp difference ('diff 101-300'). The number of genes assigned to each category for the six species investigated is shown in Table 4.2.

	Start		Stop	
	'diff 0'	**'non-diff 0'**	**'diff 0'**	**'non-diff 0'**
C. elegans	11416	8713	9750	10379
C. briggsae	No data available			
A. thaliana	8503	18316	7550	19269
D. melanogaster	3519	10521	3972	10068
D. rerio	13195	8128	13186	8137
H. sapiens	3813	18949	3500	19262

Table 4.2 Number of genes in the 'diff 0' and non-diff 0' categories.

4.3.4 Distance between genes

The distance between genes was broken down into the distance from a transcript start to the next transcript and the distance from a transcript stop to the next transcript. A Perl script was used to scan upstream from each transcript start, and downstream from each transcript stop, until either the start or stop of the next transcript on the chromosome (depending on gene orientation) was encountered. In this way two sets of figures were obtained; the distances to the next transcript upstream of the transcript starts and the distances to the next transcript from the transcript stops. Where overlaps between the start or stop of one transcript with another transcript occurred, the distance was indicated by recording a distance of -1.

These data were used to create sub-categories of the main datasets, based on genes with at least 1200 bp and 2000 bp of non-genic DNA upstream from their start or downstream from their stop site respectively. These gene distance sets were combined with the sets describing the difference between CDS and transcript start and stop positions, to create further gene set sub-categories.

4.3.5 Nucleotide frequency around genes and CDS

The frequency of each of the standard nucleotides, adenine (A), cytosine (C), guanine (G), thymine (T) in the region surrounding transcript and CDS start and stop positions was calculated for several gene sets for each of the six species under investigation in this chapter. This analysis used the same start and stop FASTA format sequence files described above for calculation of MRS frequency. As previously discussed, the sequence records in these files spanned 1200 bp upstream of the start position (downstream of the stop position) and 600 bp towards the centre of the transcript or CDS. The nucleotide frequencies were expressed as the percentage of all sequence records in the set in which a particular nucleotide was found at each position in the 1800 bp start or stop region. For most analyses the frequencies were 'binned' into non-overlapping windows of 10 bp as this gave the best compromise between detail and reduction of noise.

4.4 Results and discussion

4.4.1 MRS around various CDS and gene sets in *C. elegans* – Are the peaks in MRS frequency related to transcription or translation?

Are the peaks of MRS frequency described previously located with respect to gene (transcript) or CDS boundaries? The initial approach was to compare MRS frequency plotted in relation to transcript start and stop sites with MRS frequency plotted in relation to CDS start and stop sites. Additionally, further insight into whether the peaks in MRS frequency are specific to coding sequences or related to more general transcription was sought by studying MRS frequency in the regions surrounding non protein-coding genes.

4.4.1.1 Transcripts differ from CDS in MRS frequency in the stop region

When all genes are considered, clear peaks in MRS frequency are evident at both the start and stop regions (Figure 4.1a). The upstream peak appears to occur at approximately the same distance upstream of the start of the gene transcript as it does from the CDS start. However, the downstream peak is much closer to the transcript stop position than the CDS stop position.

Figure 4.1a also shows that there is no evidence of peaks in MRS frequency in the region surrounding both the start and stop of non protein-coding genes. The MRS frequency remains at a constant level up- and downstream of these genes and in the genes themselves. This suggests that the MRS frequencies are not related to universal transcription processes, only those specific to protein coding genes. For example, they may be related to the use of RNA polymerase II, rather that RNA polymerase I which is used to transcribe rRNA genes. However, these data leave open the possibility that the MRS frequency peaks are in some way related to translation of protein coding sequence. In either case, as the MRS frequency peaks appear to be a phenomenon restricted to only protein coding genes, subsequent analyses excluded non-protein coding genes.

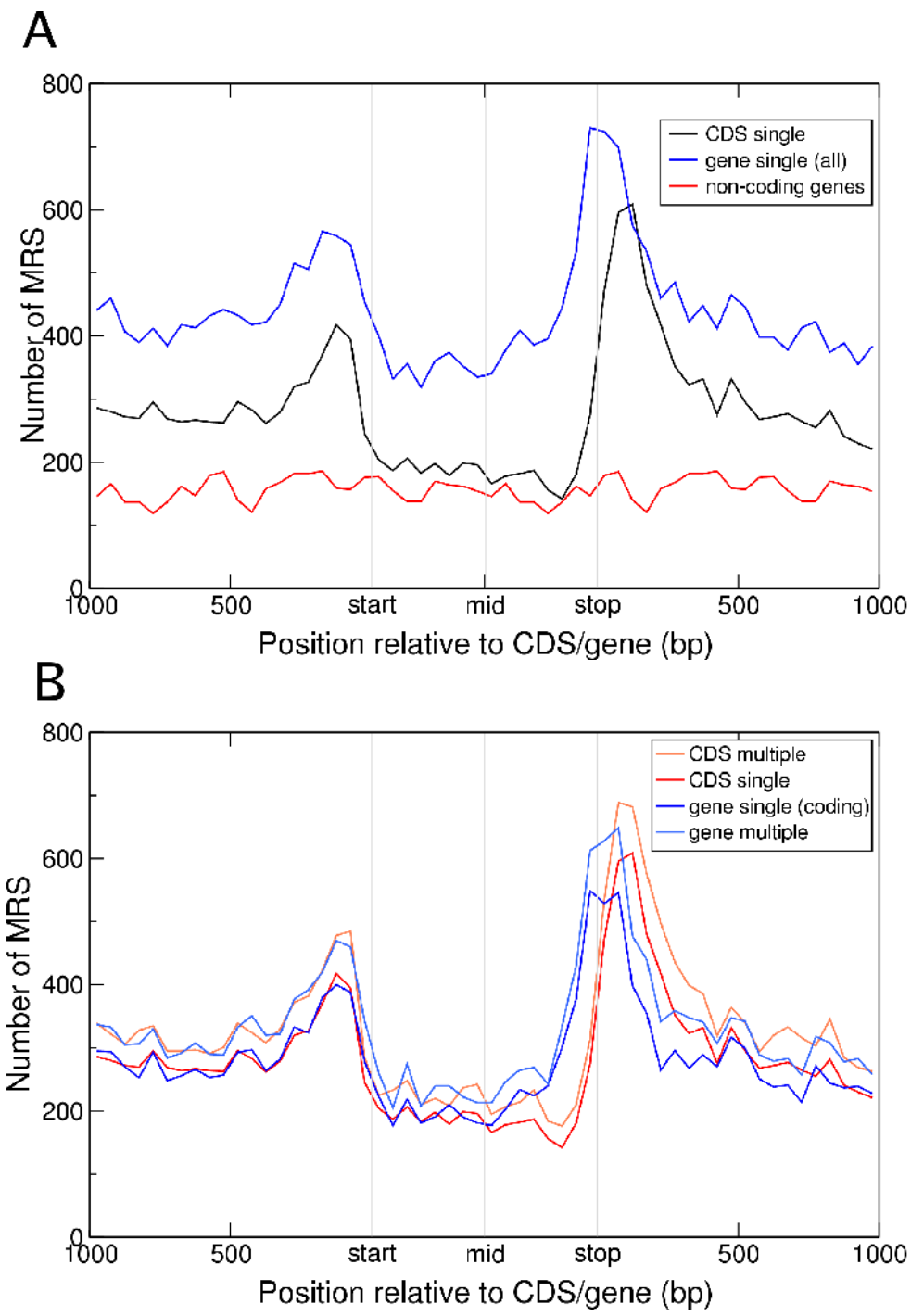


Figure 4.1 Frequency of MRS relative to C. elegans genes and CDS.

[A] Frequency of MRS in 50 bp bins relative to CDS single (black), transcript (gene) single all (blue) and non-coding genes (red). **[B]** Comparison of MRS frequency relative to CDS multiple (light red), CDS single (red) transcript (gene) single (blue) and transcript (gene) multiple (light blue). (continued overleaf)

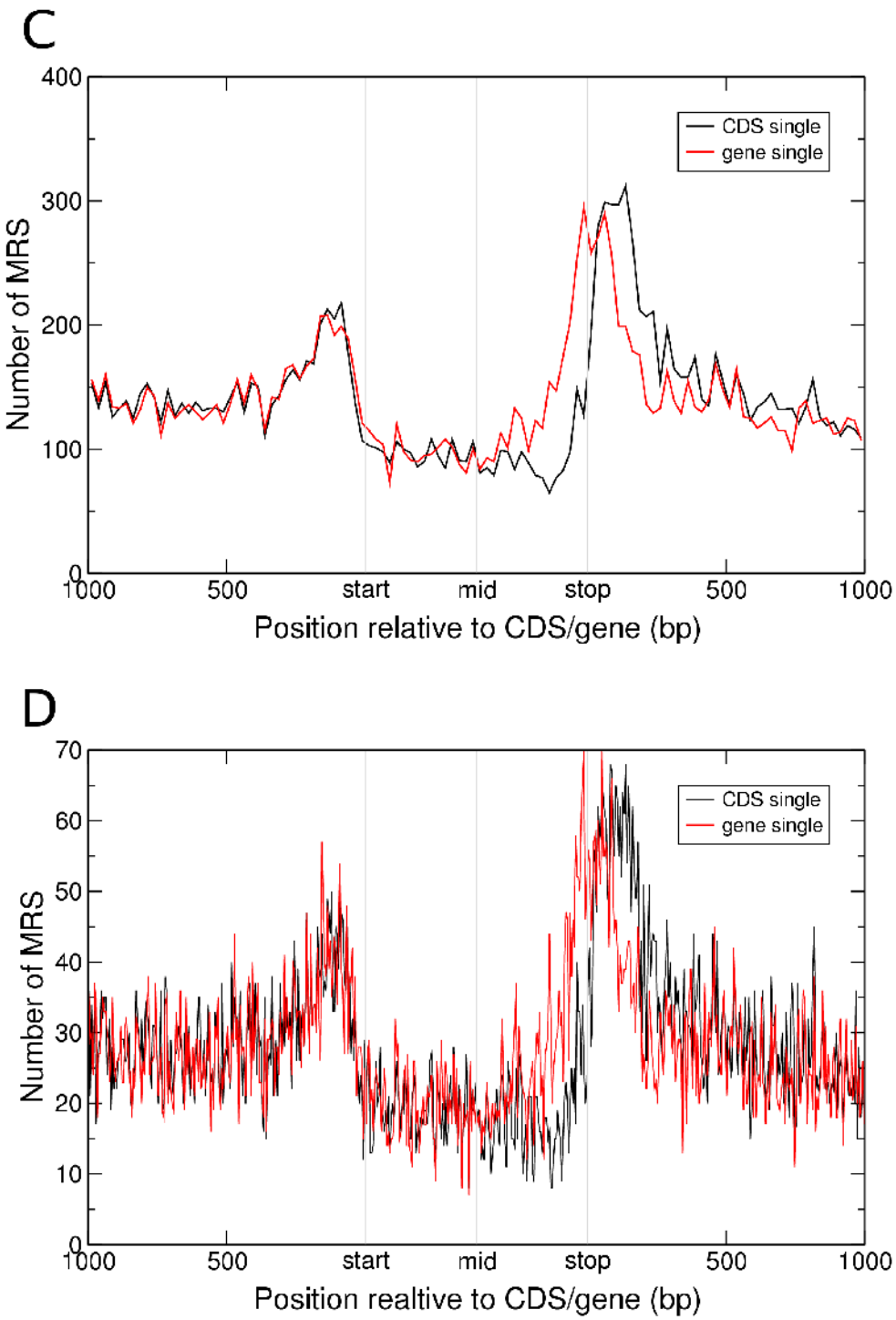


Figure 4.1 Frequency of MRS relative to C. elegans genes and CDS
[C] MRS frequency relative to CDS single (black) and transcript (gene) single (red) in 25 bp bins. **[D]** MRS frequency relative to CDS single (black) and transcript (gene) single (red) in 5 bp bins.

Exclusion of non protein-coding genes means that CDS and transcript (protein coding only) sets have the same number of objects. In addition, transcript and CDS sets can be expanded to include all gene models in which a distinct amino acid sequence is encoded (i.e. the CDS and transcript (coding) multiple sets). These data are shown in Figure 4.1b. For both transcript and CDS there is a close correlation between the MRS frequency for single and multiple transcripts per gene. The only difference is that the multiple transcripts per gene sets have a higher number of MRS, a reflection of the greater number of objects in the multiple sets. As the single and multiple sets have very similar patterns of MRS frequency, most subsequent analyses concentrate on the single transcript and single CDS sets.

It is also worth noting from this graph the apparently high degree of similarity in MRS frequency between equivalent transcript and CDS sets at the 5' end. For example, both transcript and CDS sets have a 5' MRS peak centred about 100 bp upstream of their respective start sites. In contrast, although transcripts and CDS display a similar shape of MRS frequency around the stop sites, the MRS peak for CDS lags about 100 bp behind that of transcripts. Another peculiar feature of the MRS frequency profile is the 'double peak' observed in the downstream MRS frequencies of the gene data. Both these characteristics could be a result of the relatively large bin size; this may mask subtle differences between transcript and CDS sets and deform the top of the downstream MRS frequency peak for transcripts.

The work carried out in the previous chapter indicated that a bin size smaller than 50 bp would not be generally appropriate due to the size characteristics of the MRS. However, in this instance, alternative bin sizes were considered to ensure that bin size was not having an unexpected effect. Figures 4.1c and 4.1d show the MRS frequency around the start and stop regions of the single transcript and CDS sets, in bin sizes of 25 and 5 bp respectively. As predicted by earlier analysis of bin sizes, the 5 bp bin data is very noisy and of limited use. The 25 bp bin is much clearer and it clearly reinforces the features noted above for the 50 bp bin Figure 4.1b. Therefore, bin size does not appear to be responsible for the similarity between transcript and CDS MRS

frequency nor the 'double peak' in the downstream MRS frequency for transcripts.

4.4.1.2 Many C. elegans transcripts have little or no UTR sequence

This means the only explanation for the close similarity in MRS frequency between transcripts and CDS around the start regions is that there is very little difference in the genomic location of the transcript and CDS start positions, or more specifically, that MRS predominantly occur in genes where there is little difference between the transcript and CDS start positions. Figure 4.2, shows the genomic distance between transcript and CDS start and stop positions and goes some way to confirming the first part of this explanation. A large number of genes have little or no difference between the transcript start and CDS start (and to a lesser extent between the stop sites). Despite the apparent proximity of transcript and CDS start/stop in many genes, the differences between some transcript and CDS start/stop positions are vary large. The large differences are not shown on the graph but are included in the summary statistics in Table 4.3.

Over two thirds of genes have only a 10 bp or less difference between the start of the transcript and the start of the CDS. Therefore, in two thirds of genes the distance of MRS from the CDS start will be within 10 bp of the distance of the same MRS from the transcript start. The same is true for MRS near the stop site of about half of genes. These data go some way to explaining the similarity between transcript and CDS MRS frequency in the start regions. Additionally, the transcript and CDS stop positions have a mean difference between them of about 100 bp, this fits with the observed lag in CDS MRS frequency compared to transcripts. However, these data cannot fully explain the similarity of MRS frequency peak location relative to the two sets of start regions as a small but significant proportion of genes have a difference between transcript and CDS start sites greater than 50 bp and the mean difference is 82 bp. The implication is therefore that MRS are only present in genes where transcript and CDS start positions are located in close proximity.

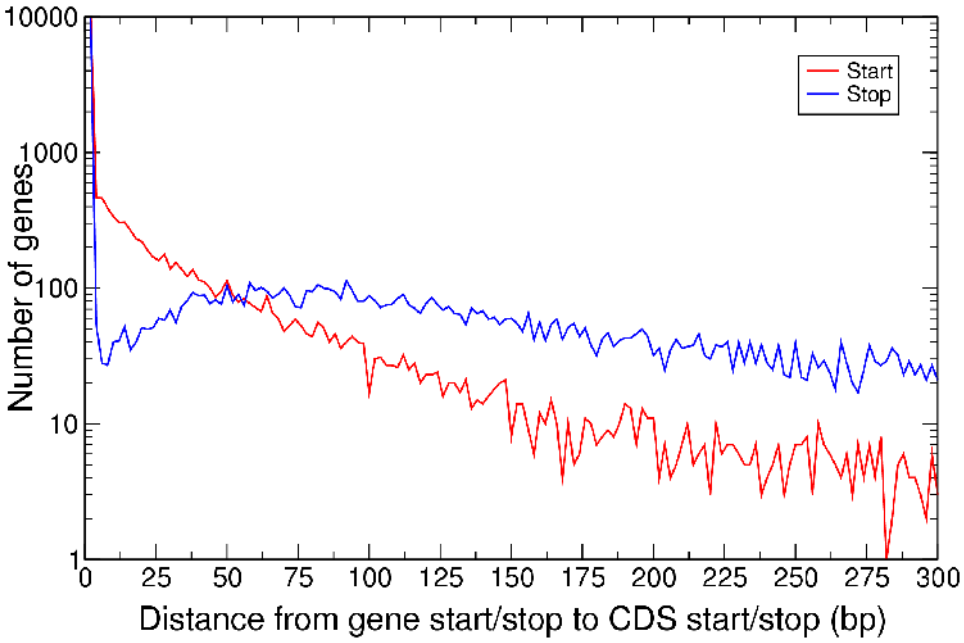


Figure 4.2 Distance between gene (transcript) and CDS boundaries

Distance in bp from transcript (gene) single start positions and the corresponding CDS start position shown in red, distances between stop positions shown in blue.

	0 bp difference (%)	≤10 bp difference (%)	≤50 bp difference (%)	Mean (bp)	Median (bp)	Maximum (bp)*
start	57	68	84	82	0	20166
stop	48	49	56	107	0	8721

Table 4.3 Difference between transcript and CDS start and stop positions.

(*some longer distances were omitted as the annotations were found to be inaccurate and corrected in subsequent versions of WormBase.)

4.4.1.3 The 3' MRS frequency peak is associated with transcription stop sites

The existence of genes in which the transcript start site and the CDS start site are at the same position may be an artefact of the data. It is highly unlikely that so many, if any, genes produce transcripts *in vivo* that contain solely the CDS with no 5' or 3' untranslated regions. Instead, the annotations for these genes are likely to be incomplete, and do not describe the true transcription start and stop positions. On this basis, the transcript and CDS sets were sub-divided into gene objects where the start/stop positions differed between CDS and transcripts ('non-diff 0') and where they were identical ('diff 0'). The MRS frequency in the start and stop region of these newly defined gene sets is shown in Figure 4.3. One consequence of segregating the gene objects in this way is that the CDS 'diff 0' set and the transcript 'diff 0' set contain objects with identical coordinates. Therefore, as shown in Figure 4.3, the transcript 'diff 0' and CDS 'diff 0' sets have an identical MRS frequency. Segregating the data in this way makes the MRS frequency pattern at the 3' end of genes much clearer. The double-top noted in Figure 4.1c for the complete coding genes set is absent. This was apparently caused by the 'diff 0' transcripts - they show a broad MRS peak ~100 bp downstream of the stop site. However, the most striking consequence of segregating 'diff 0' genes is the distinct association of MRS with transcription stop sites. In addition, the tight MRS frequency peak at the 'non-diff 0' transcript stop site correlates with a relatively broad MRS frequency peak, centred approximately 100 bp downstream of the 'non-diff 0' CDS stop site.

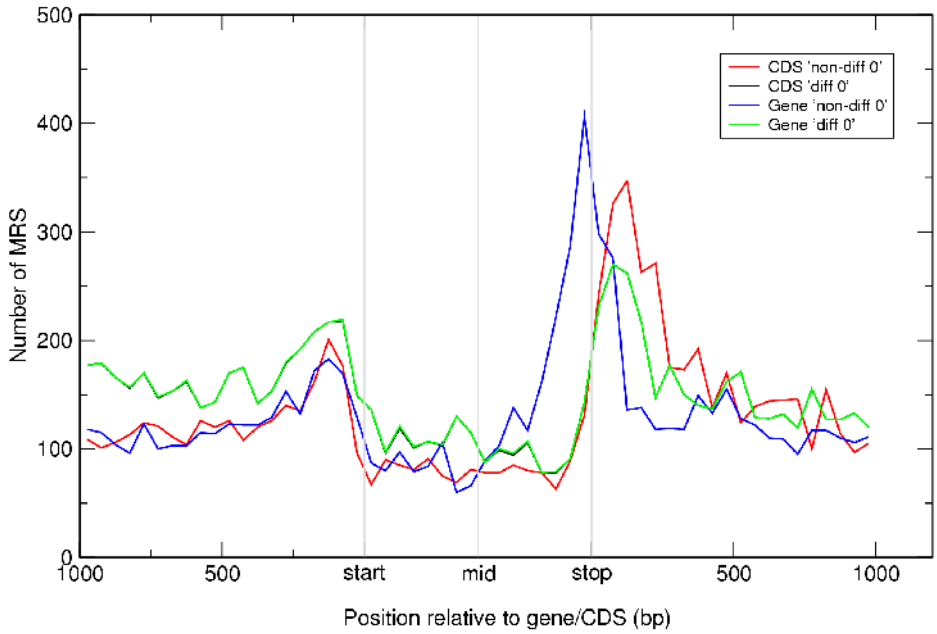


Figure 4.3 Comparison of MRS in 'diff 0' and non-diff 0' transcripts (gene) and CDS

MRS relative to 'diff 0' transcripts (genes) (green), 'non-diff 0' transcripts (genes) (blue), 'diff 0' CDS (black) and 'non-diff 0' CDS (red). Transcript (gene) 'diff 0' and CDS 'diff 0' overlap completely.

4.4.1.4 The 5' MRS peak dissected

In contrast, the relationships between transcript and CDS MRS frequency at the start region of genes remains unclear. In Figure 4.3 the 5' MRS peak appears to occur at the same location relative to both transcript and CDS start positions, even for gene objects that have a difference between their transcript and CDS start sites. The MRS peak for both 'non-diff 0' and 'diff 0' CDS (and therefore 'diff 0' transcripts) occurs in the same place, about 125 bp upstream of the CDS start site. Based on this and following the pattern observed at the stop region, we would expect the MRS frequency peak for 'non-diff 0' transcripts to occur close to the transcript start site. Its actual location, about 125 bp upstream of the transcript start site, seems incompatible with the CDS MRS frequency peak. To dissect the MRS frequency pattern further, two additional analyses were performed. Firstly, the MRS frequency was calculated as the number of MRS per gene object, rather than the total number of MRS as in the previous analyses. Secondly, the MRS frequency was calculated for a bin size that was not only smaller but also not divisible by 5. This provided an alternative viewing window compared to the previous analyses. This new representation of the data, specifically for 'non-diff 0' gene objects, is shown in Figure 4.4. The MRS per gene object for CDS and transcript sets in bins of 50 showed that, in general, the MRS frequency per gene was the same across all the sets, something not immediately apparent from the data in Figure 4.3. One exception to this was for the MRS peak. This peak is higher and tighter for 'non-diff 0' CDS compared to the 'non-diff 0' transcript and complete CDS sets. This is further emphasised by the 17 bp bin presentation of the data that shows the 'non-diff 0' transcript set MRS peak overlapping the tighter 'non-diff 0' CDS set MRS peak by at least 17 bp both upstream and downstream. These data therefore hint that the transcript and CDS 5' MRS peaks may be slightly off-set, as was observed in the 3' MRS peaks. However, despite this, these data alone are insufficient to fully explain the location of the 5' MRS peaks. Based on the average difference between the transcript and CDS start positions, we would expect either the CDS set MRS peak to occur further upstream

of the CDS start, or the 'non-diff 0' transcript set MRS peak to occur closer to the transcript start, depending on whether the MRS peak is orientated about the transcript or CDS start positions.

One remaining explanation for the data is that within the 'non-diff 0' set the positioning of the MRS peak differs for increasing transcript and CDS start position differences. Three sub categories of the 'non-diff 0' transcript set were created (Table 4.4) where the CDS-transcript start position difference is from 1 to 50 bp, from 51 to 100 bp and from 101 to 300 bp.

In the 'diff 1-50' category, there is a high incidence of MRS forming a peak in the same location (about 125 bp upstream of the start site) as observed for all 'non-diff 0' transcripts (Figure 4.5). However, the 'diff 51-100' category has no MRS peak at this position, and instead displays an MRS peak about 25 bp upstream of the start site. The 'diff 101-300' category also does not have an MRS peak at 125 bp upstream of the transcript start site. This set displays a more complex pattern with several sharp MRS peaks just downstream of the start site and a wider peak centred about 500 bp downstream of the transcript start site.

Although this sub categorisation of the 'non-diff 0' transcript set may have raised additional questions, it has gone some way to resolving the issue surrounding the apparently incompatible placing of the transcript and CDS set 5' MRS peaks. The 'non-diff 0' transcript set 5' MRS peak is dominated by a large number of genes for which the transcription start site is within 50 bp of the CDS start site. Additionally, these genes have a relatively high frequency of MRS at the peak, which enhances their influence over the apparent positioning of the 5' MRS peak in the complete transcript dataset. The lower number of genes with greater distance between the transcript and CDS start and the differing pattern of MRS frequency for these genes means that their influence on the complete transcript set MRS frequency is limited. Therefore the 'non-diff 0' transcript set behaves almost as if there is no difference between its start sites and the corresponding CDS start sites, so the MRS pattern appears very similar.

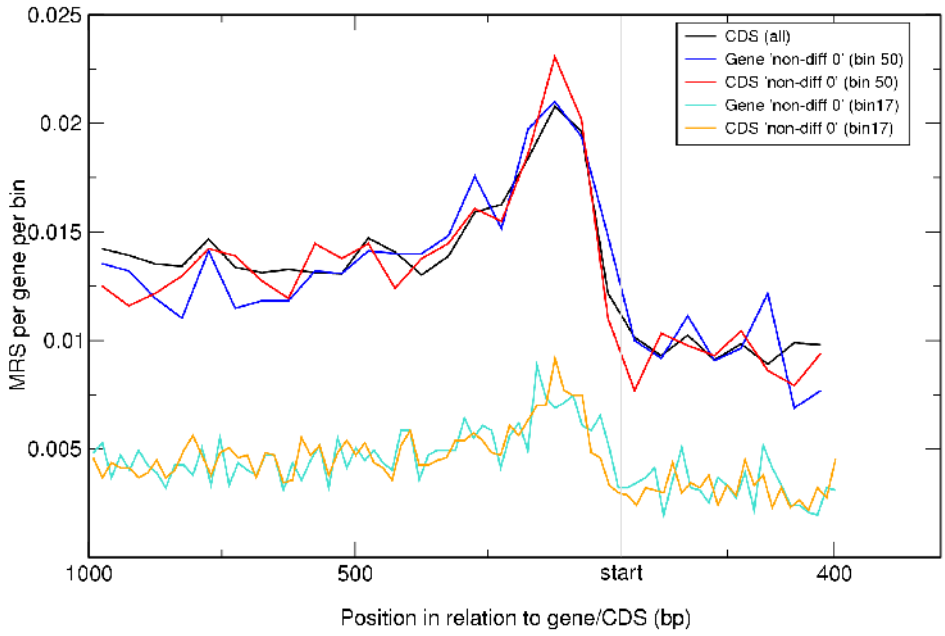


Figure 4.4 MRS frequency at the start region of 'non-diff 0' transcripts (genes) and CDS

	'diff 1-50'	'diff 51-100'	'diff 101-300'
Number of genes	5537	1394	1052

Table 4.4 Number of genes in 'non-diff 0' sub-categories

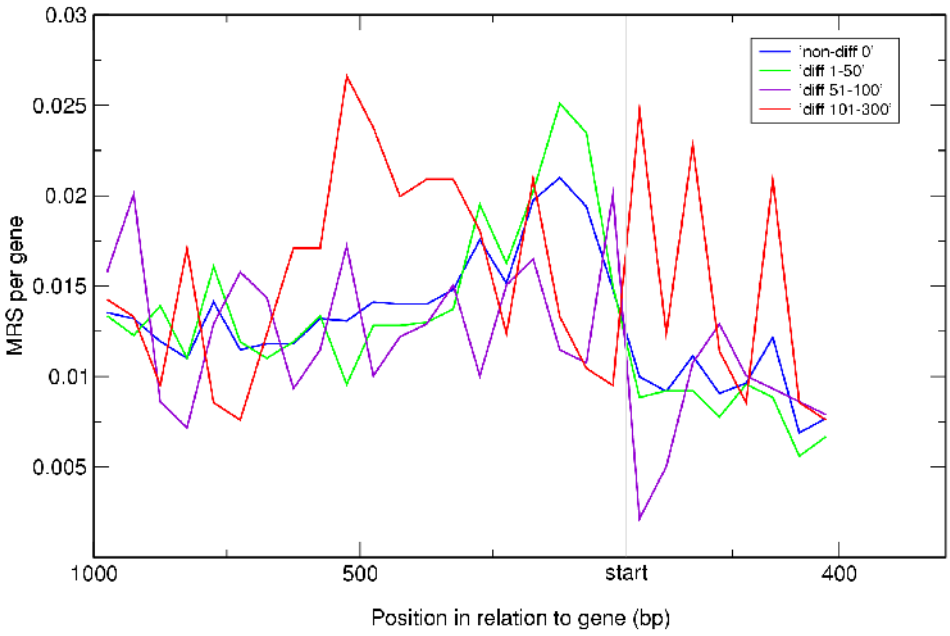


Figure 4.5 Frequency of MRS in start region of for various 'non-diff 0' gene categories

Two lines of evidence suggest that, in contrast to the 3' MRS frequency peak, the 5' MRS frequency peak is CDS orientated. Firstly, the MRS frequency peak for the 'non-diff 0' CDS set shown in Figure 4.4 is tighter and higher than for the equivalent transcript peak. Secondly, if the MRS frequency peak was transcript orientated then we would expect it to occur in the same position relative to the start site for all genes. In fact, as shown in Figure 4.5, the MRS frequency peak occurs in a different location, depending on the distance between the CDS and transcript start.

4.4.2 MRS around genes in other animals and plants

The analyses carried out above provided evidence that the peaks in MRS frequencies in *C. elegans* may be orientated about both CDS and transcripts. Therefore, in analysing the MRS frequency in other species both CDS and transcript sequences were used. Table 4.5 shows the number of genes included in each sequence set for the six species investigated here.

4.4.2.1 A survey of MRS in six eukaryotes

Figure 4.6 shows the total number of MRS surrounding transcripts (black line) and CDS (red line) for *Arabidopsis thaliana, Caenorhabditis briggsae, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans* and *Homo sapiens.* There is a peak in MRS incidence at the transcript stop of *A. thaliana.* The broader peak of MRS downstream of the *A. thaliana* CDS stop indicates that this rise in MRS frequency is orientated about the transcript stop site. There is a broad peak in MRS incidence about 500 bp upstream of the CDS start but no clear peak near the transcript start.

	Transcript annotations	**Transcripts after filtering**	**CDS annotations**	**CDS after filtering**
A. thaliana	26,819	26,804	26,819	26,784
C. briggsae	*	*	19,531	13,640
D. rerio	21,322	18,880	21,322	9,296
D. melanogaster	14,039	13,981	14,039	13,895
C. elegans	20,130	20,128	20,130	20.116
H. sapiens	22,762	22,532	22,740	19,629

Table 4.5 The number of transcript and CDS sequences used for the six species analysed here.

(* transcript annotations not available for the version of *C. briggase* used.)

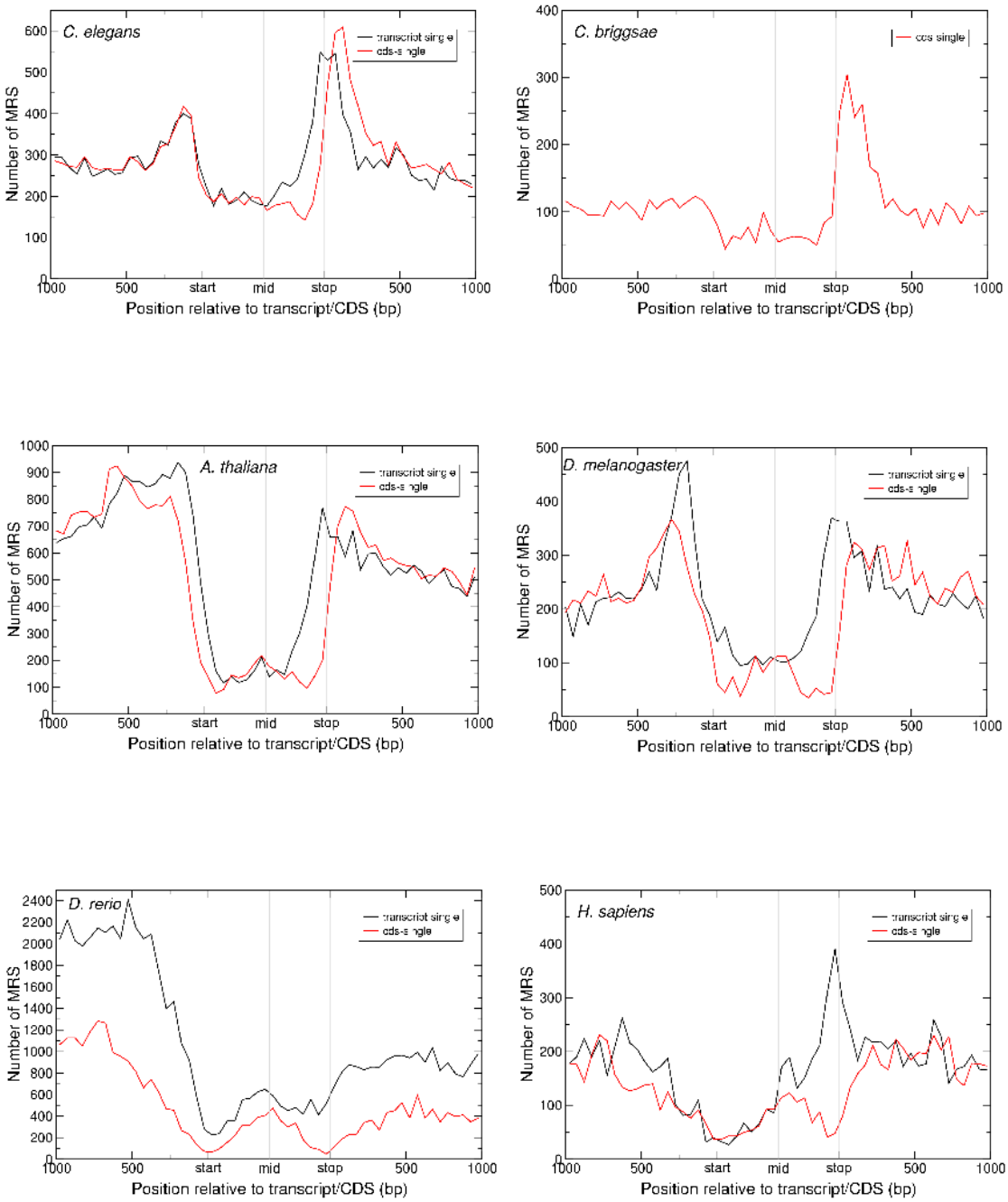


Figure 4.6 MRS relative to transcripts and CDS in six eukaryotes

In *H. sapiens* there is a marked difference in the incidence of MRS for CDS and transcripts over the regions studied. This is likely a reflection of the comparatively large UTRs found in *H. sapiens.* There is a strong 3' MRS peak at the transcript stop, although no corresponding peak is evident downstream of the CDS. There is some evidence for an indistinct MRS peak several hundred bp upstream of the transcript and CDS start sites but this may be due to stochastic fluctuations.

In *D. melanogaster* there are MRS peaks on the transcript stop and 200 bp upstream of the transcript start, although only the start peak has a clear equivalent in the CDS sequences. For *D. rerio* there is a large discrepancy between the number of MRS surrounding transcripts and the number surrounding CDS. This is caused by the much lower number of CDS sequences used in the analysis, due to the failure of over half of the CDS sequences to pass quality filters. Regardless of this, the pattern of MRS incidence around *D. rerio* is the most dissimilar to the other species studied. There are many more MRS upstream of transcripts and CDS than downstream and there are no obvious peaks of MRS. Unfortunately, no transcript data were available for *C. briggsae.* The pattern of MRS frequency around the CDS was described in chapter 2, the main feature being the peak in MRS about 100 bp downstream of the CDS stop site.

In summary, all species, with the exception of *D. rerio,* have a peak in MRS incidence close to the transcript stop, although a corresponding peak in MRS incidence downstream of CDS stop sites is not apparent in all species. Only *D. melanogaster* and *C. elegans* have a clear peak in MRS frequency at or near the transcript and/or CDS start, although weak peaks may exist in other species.

4.4.2.2 MRS and AT% in six eukaryotes

Work previously described in this thesis showed that the MRS, to a certain extent, had a positive correlation with AT content in *C. elegans*. An initial assessment of the relationship between AT and the incidence of MRS in other species was made using the data in Figures 4.7 (describing CDS) and 4.8 (describing transcripts).

Additionally, instead of displaying the total number of MRS from all transcripts at each locus, in these figures the frequency of MRS, calculated as the number MRS per CDS or transcript sequences in the set, is shown. This allows direct comparisons to be made between transcript and CDS and different species. In *A. thaliana* there is a very broad positive correlation between AT content and MRS frequency. However, there are some increases in MRS frequency not matched by a change in AT content, such as the peak of MRS around 500 bp upstream of the CDS start. The MRS peaks around the transcript start and stop are less clear. In contrast, there are sharp changes in AT content at the *A. thaliana* start and stop in comparison to the more gradual changes around the CDS.

The frequency of MRS between the transcript and CDS sequences of *D. rerio* is consistent, showing that the discrepancy in the number of MRS between transcripts and CDS shown in Figure 4.6 was indeed due to the lower number of CDS sequences used. The MRS frequency in *D. rerio* is consistently much higher than that observed for all other species, particularly upstream of the transcript and CDS start sites. Despite scaling for the number of transcripts/CDS used, there is still a significant discrepancy between the level of MRS 1000 bp upstream of the transcript and CDS starts and 1000 bp downstream of the transcript and CDS stop. This is in contrast to the AT content of *D. rerio*, and indeed the MRS frequency of all other species, all of which are approximately equal at the ends of the 1000 bp regions studied. The only clear MRS peak occurs about 750 bp upstream of the CDS start. There may be a corresponding MRS peak 500 bp upstream of the transcript start, although it is not as distinct.

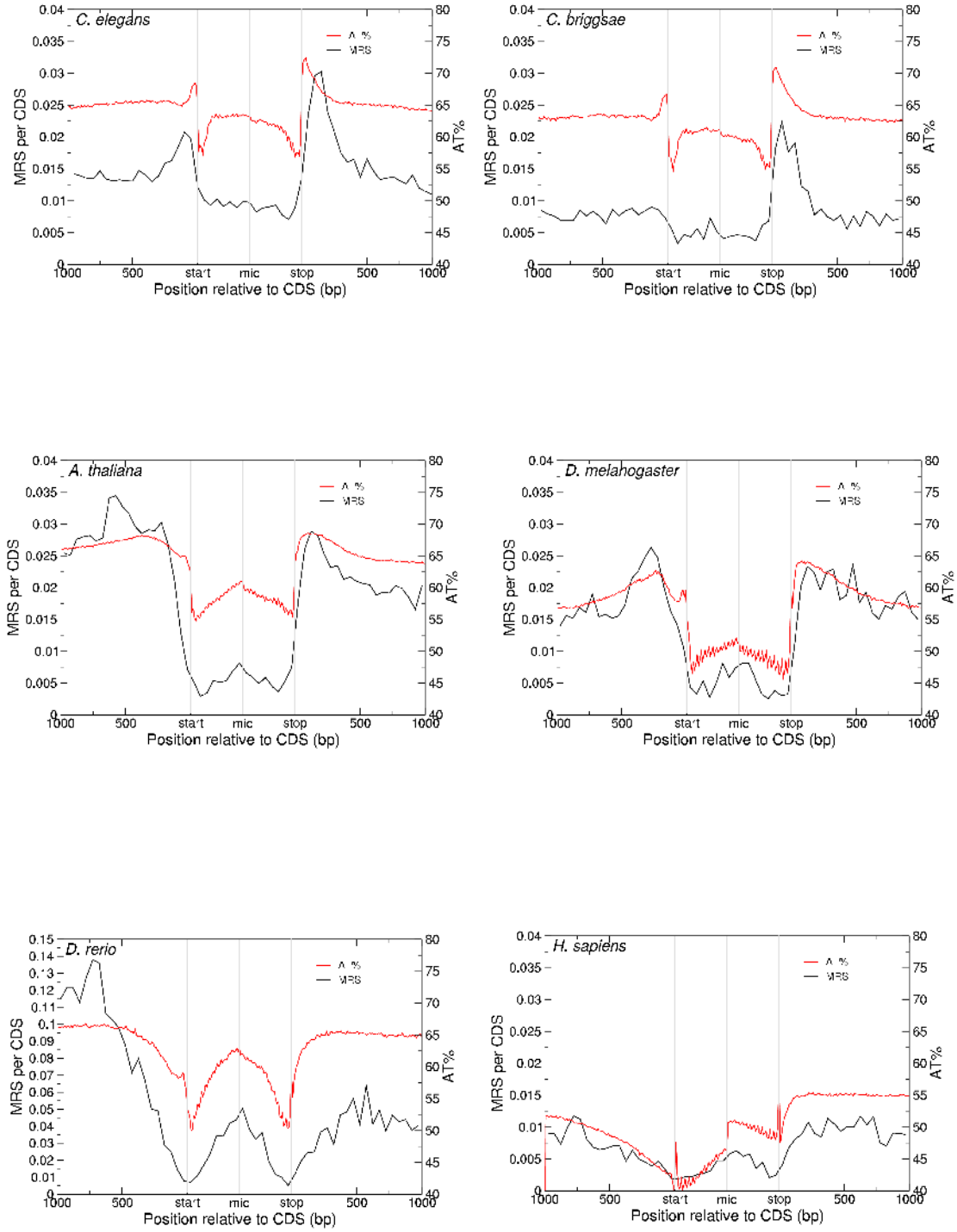


Figure 4.7 Frequency of MRS and AT content relative to CDS from six eukaryotes

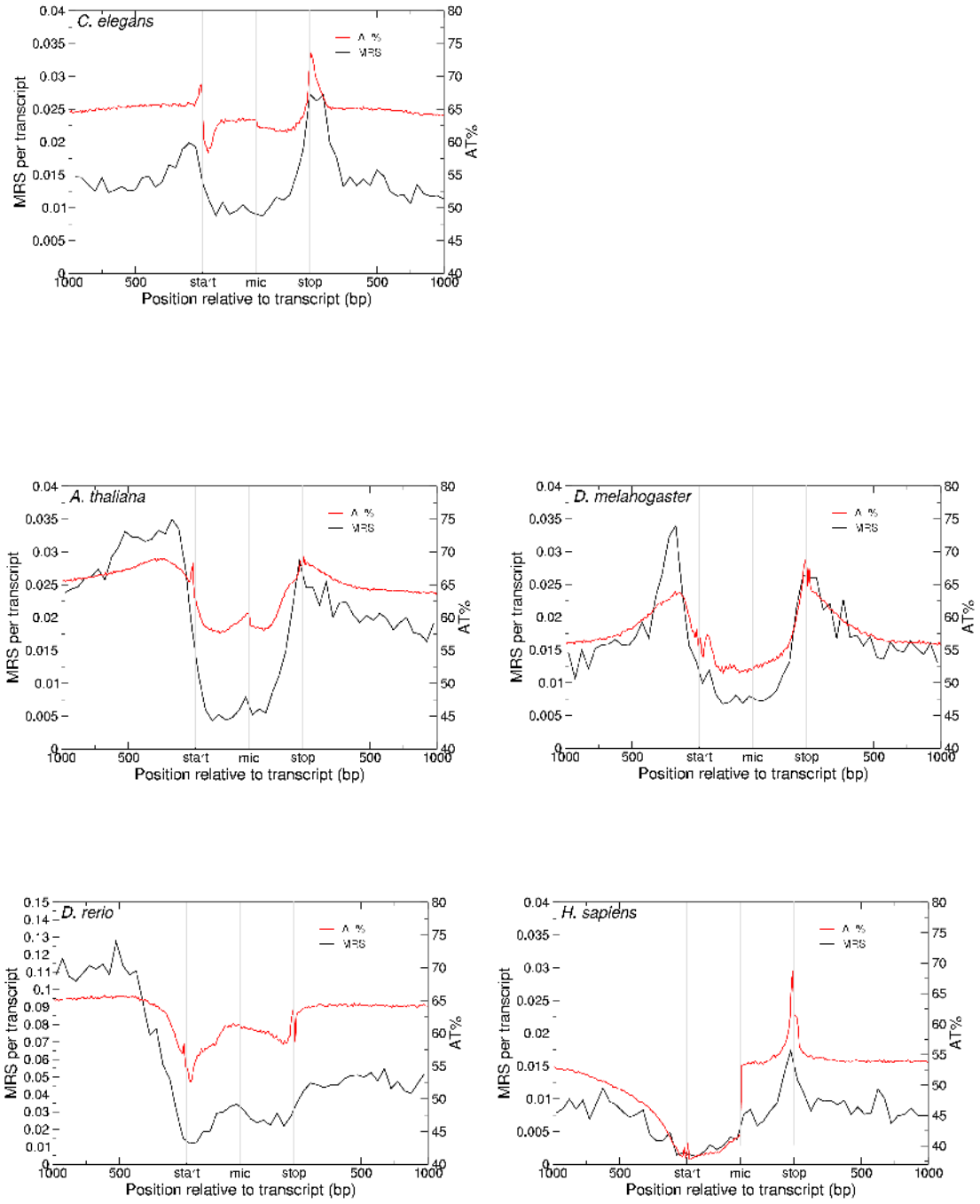


Figure 4.8 Frequency of MRS and AT content relative to transcripts from five eukaryotes

No data were available for *C. briggsae*

In general, the MRS frequency of *D. melanogaster* has a clear positive correlation with AT content. The MRS peak about 300 bp upstream of the CDS start is represented as a higher and narrower peak about 200 bp upstream of the transcript start. Most of the data presented in Figures 4.7 and 4.8 for *C. elegans* and *C. briggsae* has been discussed previously. The AT profile around the CDS of both species is very similar, with peaks and troughs at the start and stop sites. However, the MRS frequency upstream of the start site differs markedly due to the absence of an MRS peak in this location for *C. briggsae.* While the AT content in the *C. elegans* transcript start sequences is similar to the equivalent CDS sequences, there is no dip in AT preceding the transcript stop site and the peak in AT is narrower.

In *H. sapiens,* MRS frequency and AT content are quite closely correlated. However, while the peak in MRS frequency at the transcript stop is matched by a peak in AT content, similar but smaller peaks in AT content at the start and stop of CDS sequences are not matched by peaks in MRS frequency. In comparison to the other species studied here, both the AT content and MRS frequency of *H. sapiens* is low. Another distinguishing feature of *H. sapiens* is the asymmetry of AT content about both the CDS and transcript start and stop regions. The other five species have approximately symmetrical AT content in the transcript and CDS start and stop regions. However, none of the species have symmetry in MRS frequency about the start and stop regions. Further distinction between AT content and MRS frequency is also found in the differing levels of MRS observed in sequence of similar AT content. The most striking example if this is the exceptionally high MRS frequency, 0.11 MRS per transcript, upstream of *D. rerio* transcript start sites where the AT content is about 65%. In comparison, in the equivalent region of *C. elegans* and *A. thaliana* where the AT content very similar, the MRS frequency is just 0.015 and 0.025 MRS per transcript respectively. The sharpest changes in AT content tend to occur at transcript rather than CDS start and stop positions. However, the location of sharp peaks in MRS frequency is much less consistent and from these data it is difficult to determine if the peaks flanking transcripts (where they exist) are orientated about the transcript or CDS start and stop positions. This may be due to a re-occurrence of the

effect generated by genes in which the annotated transcript and CDS start and/or stop locations do not differ. Further evidence for this comes from the MRS peak at the stop site of transcripts in *D. melanogaster* which displays the characteristic double-top that was indicative of the presence of 'diff 0' transcripts in *C. elegans.*

4.4.2.3 The effect of genes with little or no UTR annotation

To gauge the influence of 'diff 0' transcripts on the MRS frequency of the five new species, 'non-diff 0' transcript sets were created in the same way as previously done for *C. elegans.* The breakdown of the difference between transcript and CDS start and stop annotations is shown in Table 4.6. As there were no transcript annotations for *C. briggsae* it was not possible to include it in this part of the analysis.

The MRS frequency and AT content for the sequences surrounding the 'non-diff 0' transcripts for the five species shown in Figure 4.9 can be compared with the equivalent data for all-transcripts shown in Figure 4.8. Most (~85%) of all *H. sapiens* transcripts were classed as 'non-diff 0' transcripts, so it is not surprising that the the 'non-diff 0' transcript data is very similar to the all-transcript data. In contrast, only 38% of *D. rerio* transcripts were classed as 'non-diff 0' and as a result there are some significant differences between the all- and 'non-diff 0' transcript datasets. The small peak and trough in AT content at the stop site of the all-transcript set is greatly amplified in the 'non-diff 0' dataset, although there is very little change in the MRS frequency about the stop region. However, there is a difference in MRS frequency in the start region, where the 'non-diff 0' set shows a large peak about 500 bp upstream of the start site that was not observed in the all-transcript set. Unlike some MRS frequency peaks in other species, this large peak is not accompanied by a change in AT content. The sharpness of the peak is an indicator that the increase in MRS is orientated a specific distance (i.e. 500 bp) upstream of the transcript start sites, rather than with reference to the CDS start sites. This is confirmed by the MRS frequency at the *D. rerio* CDS start region (Figure 4.7), which shows a broad peak starting about 700 bp upstream of the CDS start, created due to the MRS peak 500 bp upstream of transcript start plus ~200 bp median difference between transcript and

CDS start sites. The mean difference between *D. rerio* transcript and CDS sites is in excess of 2.5 kb, indicating the presence of some transcripts with huge transcript-CDS start site differences. Under the assumption that the MRS peak is orientated at the transcript start, these genes have the effect of extending the CDS start region MRS peak further upstream.

The exclusion of 'diff 0' transcripts elicited a similar effect in *A. thaliana, D. melanogaster* and *C. elegans.* For each of these species the MRS frequency and AT content patterns in the 'non-diff 0' transcript start regions remained unchanged from that observed in the all-transcript start regions. However, the amplitude of the MRS frequency and AT content peaks at the 'non-diff 0' transcript stop site is greatly increased when compared to the smaller and, in the case of *C. elegans* and *D. melanogaster,* 'double peaked' MRS frequencies observed for the all-transcript sets. In both *C. elegans* and *A. thaliana* the MRS peak correlates with a broader peak several hundred bp downstream of the CDS stop. This indicates the MRS frequency peak is orientated about the transcript stop site in these species. This pattern is not as clear in *D. melanogaster*, as no obvious MRS frequency peak was observed downstream of *D. melanogaster* CDS stop sites. However, this could be due to a more complex profile of differences between the CDS and transcript stop coordinates.

		Gene objects (post-filtering)	Number 'diff 0'	Number 'non-diff 0'	Mean difference all genes (bp)	Median difference all genes (bp)	Mean in 'non-diff 0' (bp)	Median in 'non-diff 0' (bp)
A. thaliana	start	26,804	8,494	18,310	135	62	199	109
	stop		7,542	19,262	167	164	233	204
D. rerio	start	18,880	11,365	7,515	955	0	2,506	170
	stop		11,370	7,510	864	0	2,265	461
D. melanogaster	start	13,981	3,481	10,500	1,018	94	1,359	153
	stop		3,925	10,056	294	123	410	213
C. elegans	start	20,128	11,415	8,713	81	0	189	30
	stop		9,748	10,380	120	16	233	139
H. sapiens	start	22,532	3,686	18,846	6,246	166	7,504	273
	stop		3,386	19,146	1,990	584	2,352	798

Table 4.6 Summary statistics for the difference between transcript and CDS annotations

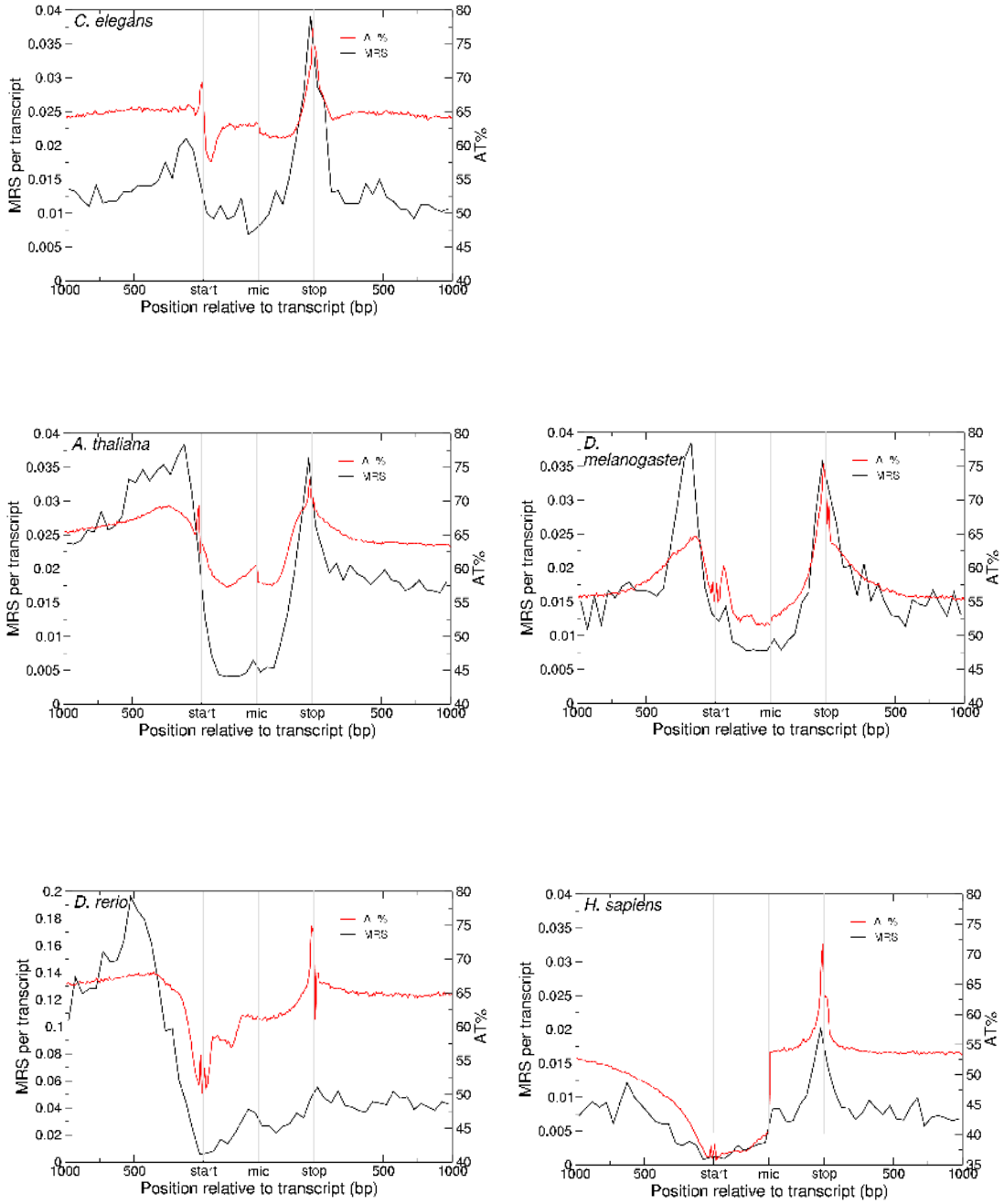


Figure 4.9 MRS frequency and AT content in 'non-diff 0' transcripts for five eukaryotes

No data were available for *C. briggsae*

These data show that the presence of 'diff 0' transcripts in the all-transcript set masked distinctive MRS frequency and AT content patterns that are present in 'non-diff 0' transcripts. Additionally, the effect of the 'diff 0' transcripts was to skew the patterns towards those observed in the CDS orientated sequences. This is further evidence that the annotations of the 'diff 0' transcripts are indeed incorrect. Following the assumption that the 'non-diff 0' transcript set contains a more accurate representation of the true sequences surrounding genes, it is appropriate to re-visit the comparison MRS frequency and AT content patterns between CDS and ('non-diff 0') transcript orientated sequences. The 'non-diff 0' data provides good evidence for an MRS frequency peak that is located 500 bp upstream of the *D. rerio* transcript start site. There is no peak near *D. rerio* stop sites, although there is a peak and trough in AT content at the stop site. In contrast, *H. sapiens, A. thaliana, C. elegans* and *D. melanogaster* all exhibit a peak in both MRS frequency and AT content at, or close to, the transcript stop site. With the exception of *H. sapiens,* these species also show evidence of MRS frequency peaks in the start region but it is difficult to discern if they are orientated around the transcript or CDS start site.

4.4.2.4 The effects of neighbouring genes on MRS patterns

The cryptic nature of the 5' MRS frequency peaks observed in *A. thaliana, C. elegans* and *D. melanogaster* could be a result of interference caused by juxtaposition of other genes. For example, if the stop site of transcript A was just upstream of the start site of transcript B, the increased likelihood of MRS at the stop site of transcript A would influence the likelihood of MRS near the start site of transcript B. The potential for this situation to occur was initially investigated by calculating the distance from the start and stop position of each transcript to the nearest start or stop of another transcript. Figure 4.10 shows the distances between transcripts up to 1000 bp, which is the length of sequence upstream and downstream of the transcript start and stop used in Figures 4.6-9.

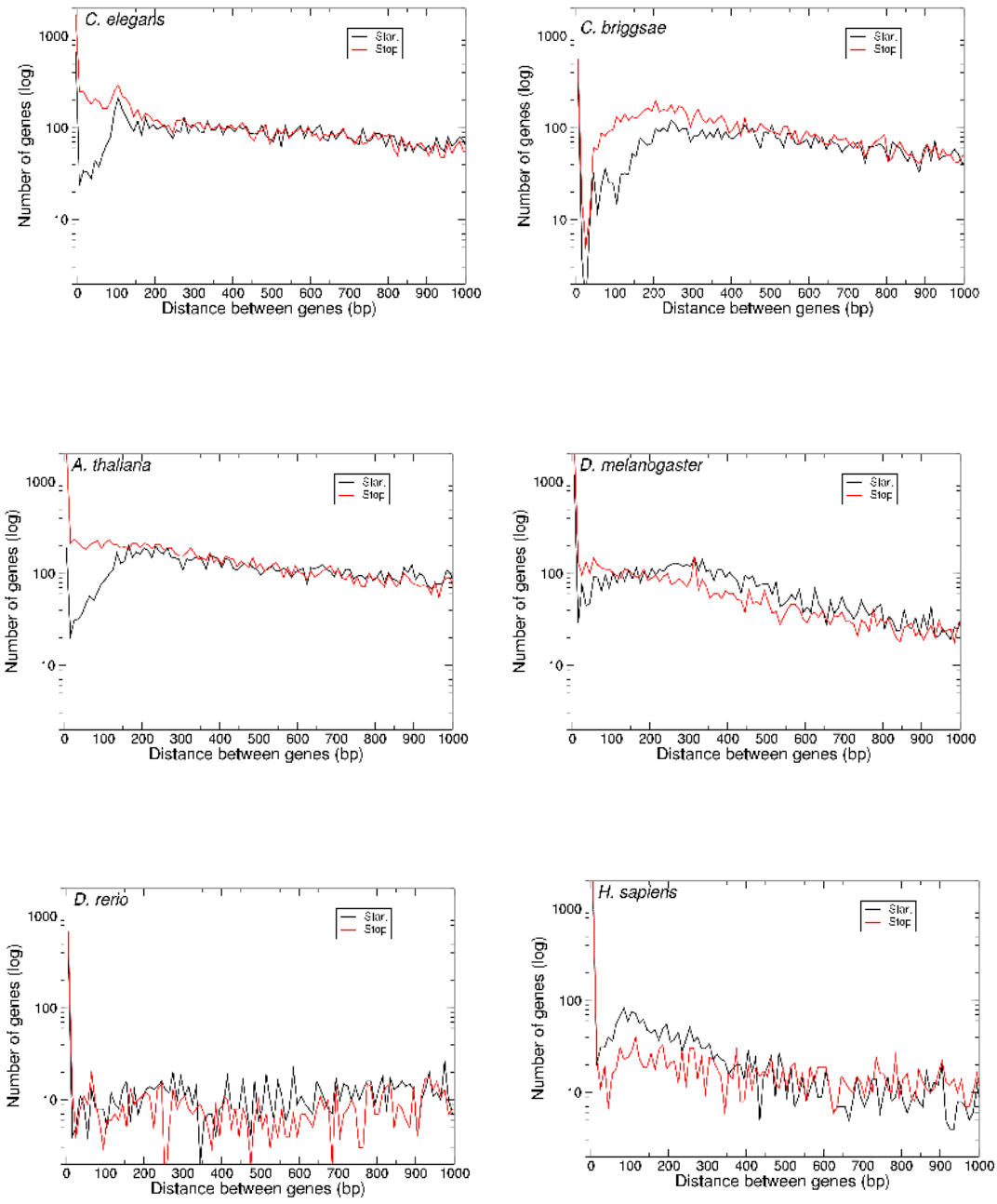


Figure 4.10 Difference between genes from start (black) and stop (red) positions

Several interesting features are illustrated in Figure 4.10. Of direct interest to this study are, firstly, that in all six species there is a large fraction of transcripts (CDS for *C. briggsae*) that are less than 0 bp from another transcript, in other words the genes overlap. Secondly, many genes start or end within 1000 bp of another gene. Therefore, a significant number of genes share the DNA flanking their start and/or stop sites with another gene. This in turn supports the hypothesis raised above whereby the MRS frequency surrounding genes is complicated by the presence of other genes nearby.

At this stage it is also worth commenting on some of the other features revealed in Figure 4.10. For instance, in *A. thaliana, C. briggsae, C. elegans* and *D. melanogaster* there is more likely to be a very short distance (less than ~150bp) to the next transcript from the stop position than from the start position. This situation is reversed in *H. sapiens* and in *D. rerio,* where there is very little difference in distance to the transcript between the start and stop positions. The general trend is for the number of transcripts to decrease as the distance to the next transcript from start and stop sites increases. However, in five of the species (all bar *D. rerio*) the number of transcripts increases as the distance from the transcript start to the next transcript increases from zero bp to 100-400 bp. In other words, fewer genes than expected occur close to the start position of another gene. This pattern is much less evident, or absent, in distances from the stop positions to the next transcript. This could be explained by the involvement of the DNA immediately upstream of the start site as promoter elements. Another interesting feature is the peak of *C. elegans* transcripts with a distance of about 75-125bp from their start and stop positions to the next transcripts. Nearly all the transcripts in this peak are found in operons, as shown in Figure 4.11.

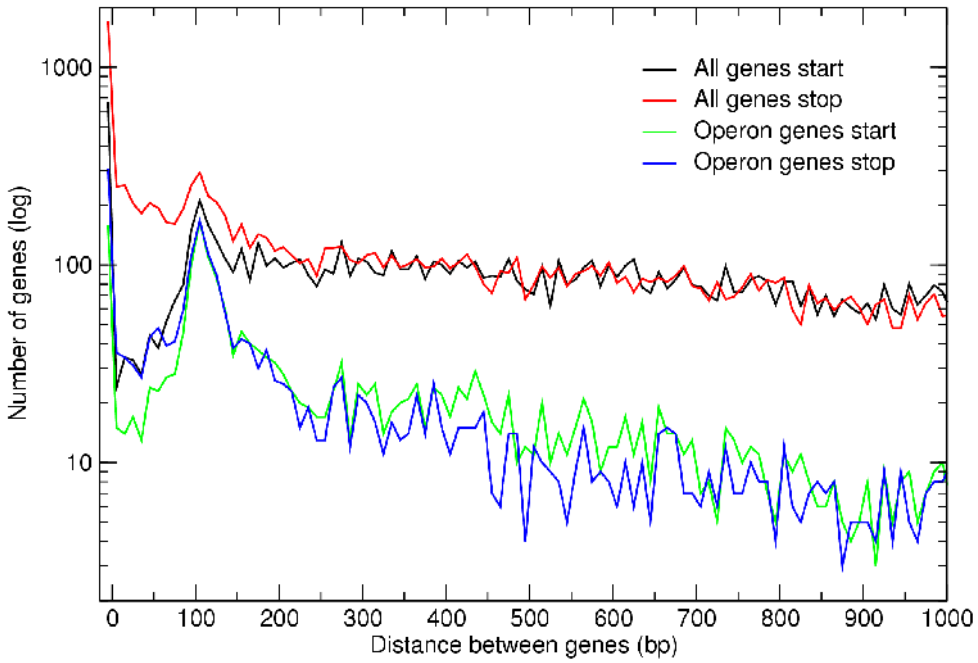


Figure 4.11 Difference between C. elegans *genes from start (black) and stop (red) positions*

The data described in Figure 4.10 establish that the proportion of genes that lie close to each other is significant enough to warrant further investigation as to their affect on the pattern of average MRS frequency surrounding genes. To this end a further subset of genes was created containing transcripts with at least 1000 bp of sequence upstream of the start site or downstream of the stop site that was not occupied by any other transcripts. This set was then combined with the previously calculated 'non-diff 0' transcript subset. This gave a set of transcripts for analysis that had transcript annotations distinct from their CDS annotation and had flanking DNA free from other genes. The number of transcripts contained in each of these sets is shown in Table 4.7.

The effect of removing closely spaced genes from the analysis of MRS frequency and AT content surrounding transcript start and stops was analysed and the results are shown in Figure 4.12. Comparison with the 'non-diff 0' transcripts shown in Figure 4.9 reveals that the MRS frequency and AT content around the transcripts of *A. thaliana* and *D. melanogaster* were the most affected. In both species instead of a gradual decay from in AT content stretching outward from the peaks near the start and stop, the AT content remains constant at about 70% for *A. thaliana* and 60% for *D. melanogaster.* In tandem with the higher AT content outside transcripts, the MRS frequency is also higher. The result of this is that upstream of the transcript start site there are no peaks in MRS frequency that stand out from the background variation. However, both species do still exhibit a clear peak at the transcript stop site.

		All transcripts (post-filtering)	'non-diff 0' transcripts	Transcripts with 1000 bp space	'non-diff 0' transcripts with 1000 bp space	% of all transcripts in 'non-diff 0' with 1000 bp space
A. thaliana	start	26,804	18,316	15,499	10,086	38
	stop		19,269	10,738	6,804	25
C. briggsae*	start	19,531	-	10,410	10,410	53
	stop		-	8,023	8,023	41
D. rerio	start	18,880	7,515	16,764	6,592	35
	stop		7,510	16,882	6,692	35
D. melanogaster	start	13,981	10,500	6,085	3,988	29
	stop		10,056	5,482	3,421	24
C. elegans	start	20,128	8,713	10,914	4,424	22
	stop		10,380	7,816	3,405	17
H. sapiens	start	22,532	18,846	18,222	15,102	67
	stop		19,146	18,230	15,478	69

Table 4.7. Transcripts with at least 1000 bp of flanking DNA free from other transcripts

(* figures for *C. briggsae* refer to CDS data)

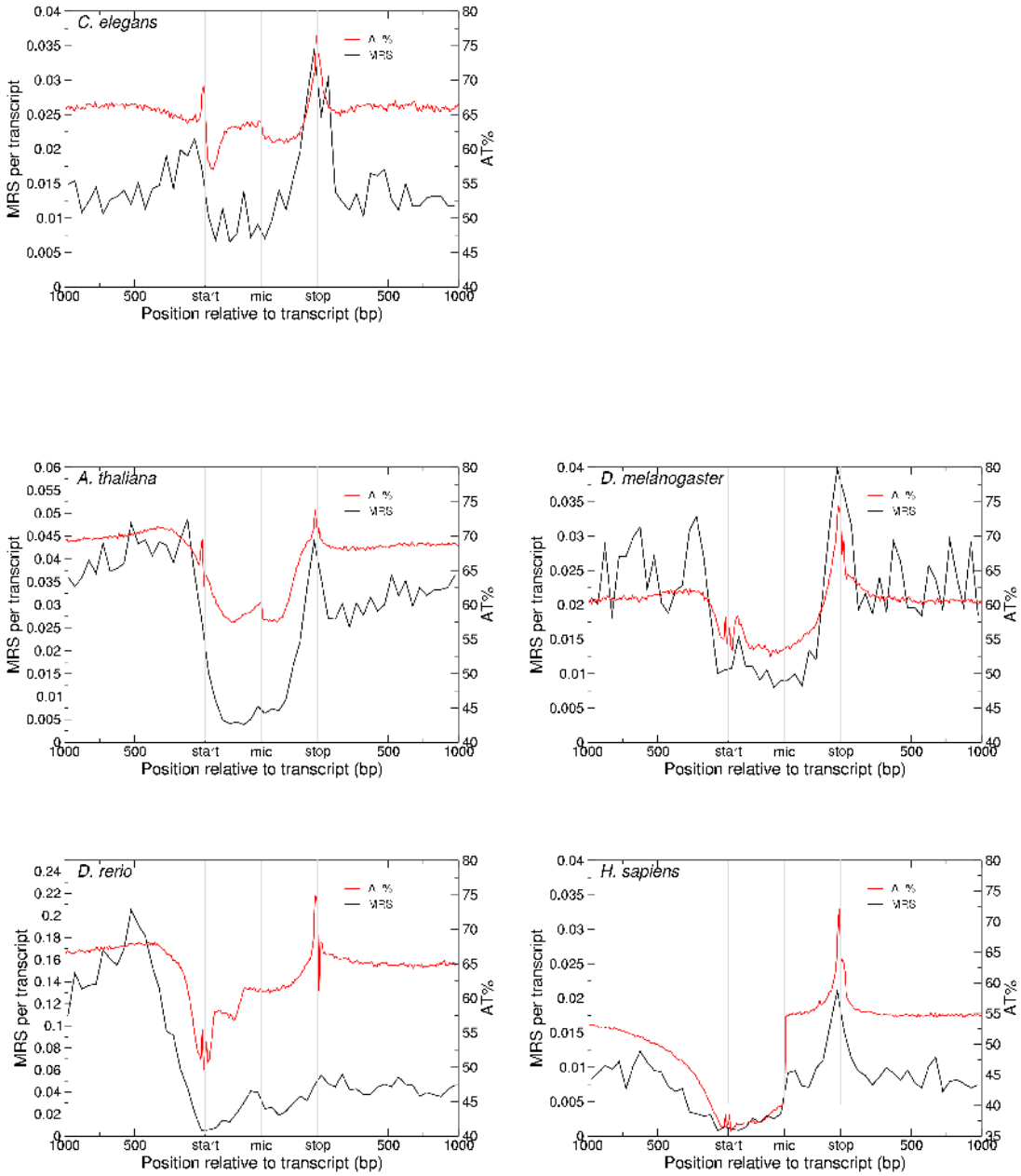


Figure 4.12 MRS frequency and AT content around 'non-diff 0' transcripts with 1000 bp of space from five eukaryotes

No data were available for *C. briggsae*

One further difference that particularly affects *D. melanogaster* is the increase in variability of MRS frequency across the start and stop regions. This is likely to be an effect of the 'non-diff 0' with 1000 bp of space set comprising only 29% and 24% of the total number of genes. *C. elegans* was also affected by increased variability of MRS frequency, however the overall pattern of AT content and MRS frequency remained similar to that observed for the 'non-diff 0' transcript set. The main difference for *C. elegans* is that the MRS frequency at the transcript stop site has a double peak in the 'non-diff 0' with 1000 bp of space transcript set. This could be due to an increase in the proportion of genes with very short differences between CDS and transcript stops in this set compared to the 'non-diff 0' set. If so, this would lead to a similar effect as created by the inclusion of 'diff 0' transcripts. The MRS frequency and AT content of the other two species analysed, *D. rerio* and *H. sapiens* showed very little difference between 'non-diff 0' and 'non-diff 0' with 1000 bp of space transcript sets. This is not surprising as in both these species very few transcripts were less than 1000 bp from the adjacent transcript.

As described above, the exclusion of transcripts with less than 1000 bp to the next transcript can elicit a significant change on the pattern of MRS frequency and AT content around transcripts, at least for some of the species under analysis. To investigate if the close spacing of genes also had an effect on the MRS frequency and AT content around CDS start and stop positions, the CDS with 1000 bp of space sub set was created. Figure 4.13 shows the MRS frequency and AT content surrounding the start and stop positions of the CDS in this set. Comparing these data with Figure 4.7 (which shows MRS frequency and AT content around all CDS, there is no 'non-diff 0' CDS set), it is apparent that the effect of closely spaced genes was the same for CDS related sequences as for transcripts. In *A. thaliana* and *D. melanogaster* the main effects are the same: the AT content remains constant instead of dipping after the peaks near the start and stop, the MRS frequency is elevated upstream and downstream of the CDS and the MRS frequency is more variable, particularly for *D. melanogaster.* The general pattern of MRS frequency and AT content for the two nematode species remains the same as for all-CDS but in both species the MRS

frequency peak near the CDS stop is lower and broader. The large degree of similarity between the *D. rerio* and *H. sapiens* CDS-all and CDS with 1000 bp of space data sets meant that there was little change in their MRS frequency or AT content, as was the case for the equivalent transcript sets.

In summary, analysing only those transcripts and CDS with at least 1000 bp to the next transcript or CDS resulted in a number of differences from the previously analysed data sets. Four of the six species analysed showed bigger variations in MRS frequency, although there was very little change in MRS frequency or AT content for *D. rerio* and *H. sapiens.* The biggest changes were observed for *A. thaliana* and *D. melanogaster* where the AT content was level in sequence flanking the transcripts and CDS, other than sharp fluctuations at the start and stop sites. A rise in MRS frequency upstream and downstream of transcript and CDS start and stop sites, coupled with an increase in variability obscured some, but not all, of the previously identified peaks in *A. thaliana* and *D. melanogaster.*

An outstanding issue with the *D. rerio* data is the notable difference in MRS frequency between 1000 bp upstream of transcripts and CDS and 1000 bp downstream of transcripts and CDS. In contrast, the AT content of sequence surrounding *D. rerio* transcripts is approximately the same on both side of transcripts and from 500 bp of CDS start and stop. Furthermore, this phenomenon is not observed for the MRS frequency of any of the other species. To gain a full impression of the MRS frequency surrounding *D. rerio* transcripts the region of sequence upstream and downstream of the transcripts was extended from 1000 to 2000 bp. In line with the findings described above, only transcripts with a different annotation to their CDS and with at least 2000 bp separating them from the next transcript were included in the analysis. Of the 21,322 *D. rerio* genes, 5,676 start and 5,758 stop region sequences passed all the filters and were used for the analysis shown in Figure 4.14.

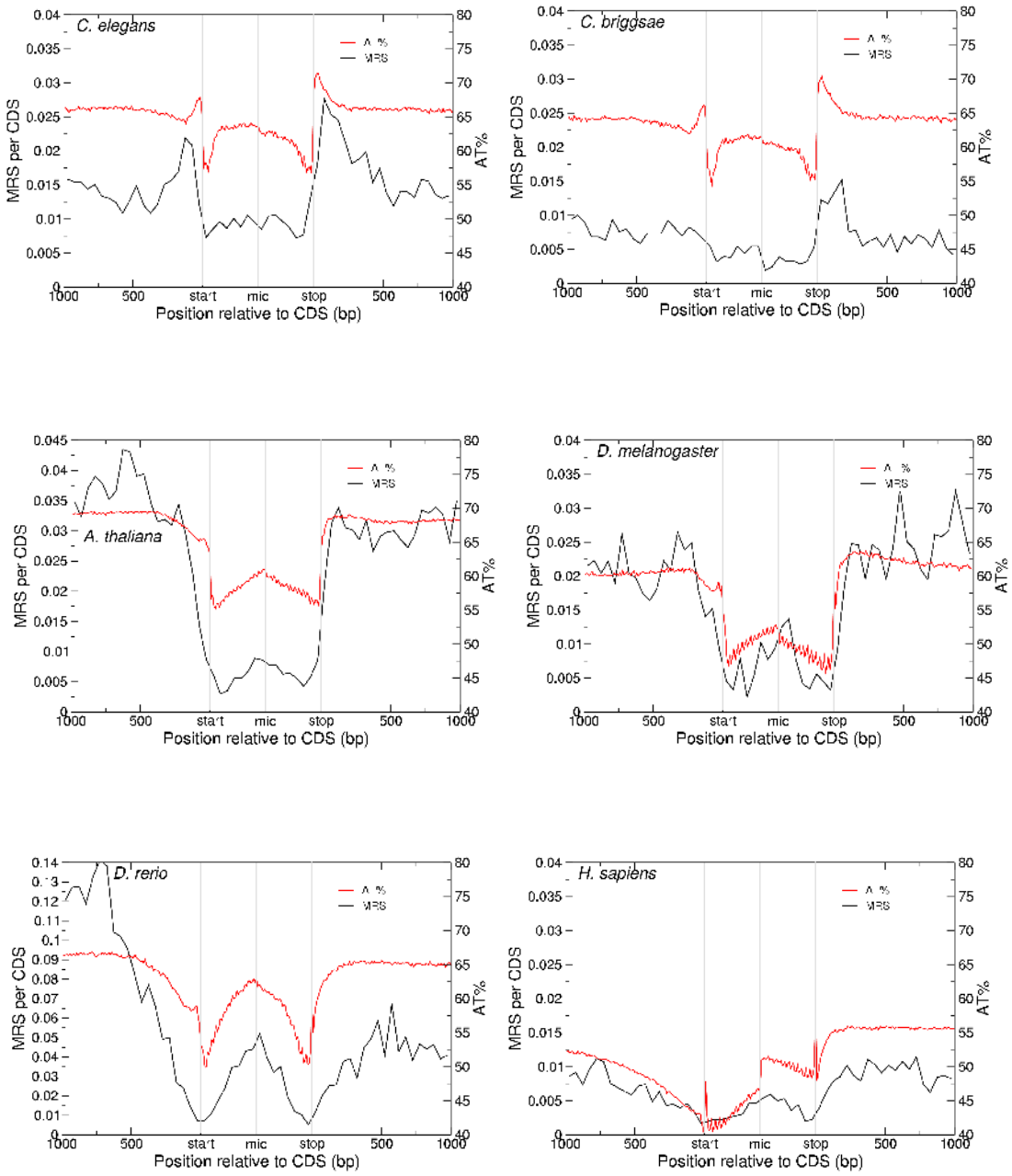


Figure 4.13 MRS frequency and AT content around 'non-diff 0' CDS with 1000 bp of space from six eukaryotes

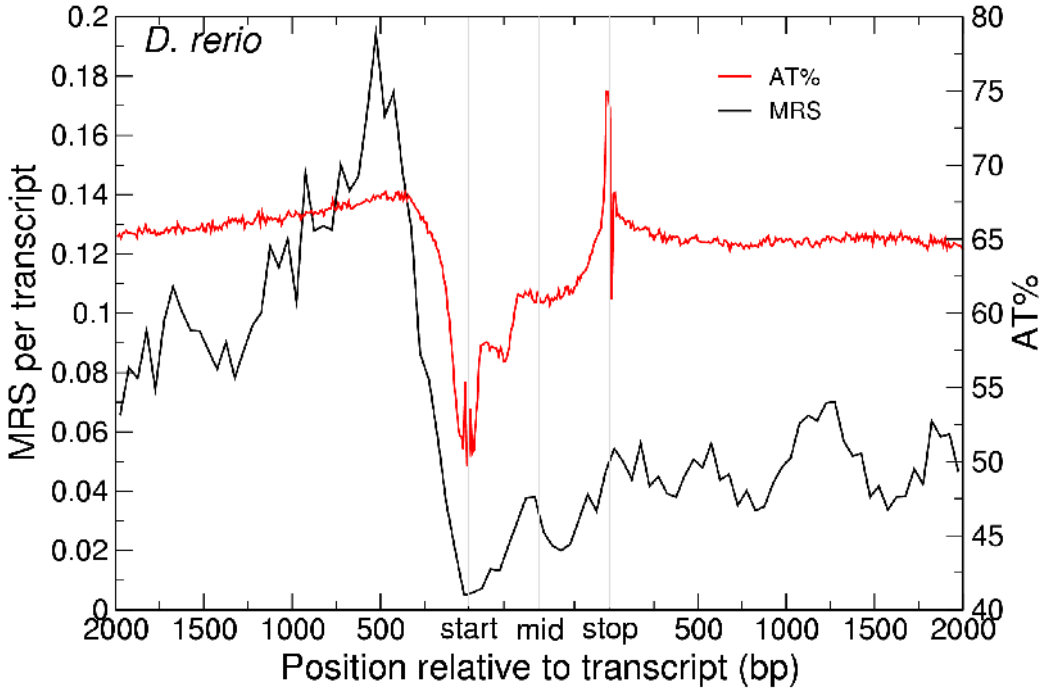


Figure 4.14 MRS frequency and AT content around D. rerio 'non-diff 0' transcripts with 2,000 bp of space

This new dataset maintains the MRS frequency and AT content characteristics previously noted in the 1000 bp nearest the transcript start and stop sites, including the peak in MRS frequency 500 bp upstream of the transcript start site. The downward trend in MRS frequency from this peak continues beyond 1000 bp and by 2000 bp upstream of the transcript start site it is almost within the range of MRS frequency that exists downstream of genes. In contrast to the other species then, the increase in MRS frequency upstream of *D. rerio* genes extends for several kb from the transcript start sites.

4.5 Further Discussion and summary

4.5.1 MRS around various CDS and transcript sets in *C. elegans*

In comparing the MRS frequency in *C. elegans* transcript sequences with CDS sequences the aim was to establish the genomic feature about which the MRS peaks were orientated. It was hoped that this information could then be used to narrow the focus of analysis in the other species. The MRS frequency peak in the stop region is located at the transcript stop site, rather than a certain distance downstream of the CDS stop site. The precise location of the MRS frequency peak in the start region is apparently less strictly defined but it is most likely orientated about the CDS start site and situated about 100 bp upstream of it. The fact that the MRS frequency peak in the stop region is more precisely located and of greater magnitude indicates that it is of greater functional importance. Further, the positioning of this peak at the transcript stop site is suggestive of the MRS being a signature that acts in DNA to affect transcription, rather than acting in RNA to affect translation, though it could have a role in primary transcript maturation.

The apparent importance of both transcript and CDS boundaries in controlling the positioning of the MRS frequency peaks prompted analysis of MRS related to both transcript and CDS sequence in other species. There were no MRS frequency peaks present in the sequence surrounding non-protein coding genes of *C. elegans* and it

was thus assumed that the same would be true for all species. Removing non-protein coding genes from the analyses cut down background noise and allowed the variation of MRS frequency around protein coding genes to be more readily perceived. The *C. elegans* study also revealed that the inclusion of all known transcripts of a gene, versus a single representative transcript made no difference to the MRS frequency pattern.

Surprisingly, a significant number of annotated genes in which the transcript and CDS had the same coordinates was identified. It was also established that the number of these genes was great enough to alter the MRS frequency pattern. In addition to the biological improbability of such annotations being correct, the analysis conducted in this chapter provided further evidence to support the discarding of such genes. The MRS frequency peak in 'non-diff 0' transcripts (i.e. those in which the CDS and transcript coordinates differ) lay on the transcript stop site, while for 'diff 0' transcripts (i.e. those in which the CDS and transcript coordinates are the same) the MRS frequency peak lay downstream of the annotated transcript stop site. This suggests that for 'diff 0' transcripts the true stop site is actually downstream of the annotated position. In excluding 'diff 0' transcripts, the MRS frequency pattern around the remaining transcripts was defined more sharply, and so this procedure was also extended to the analysis of the other species.

The complex nature of the MRS frequency pattern around genes was underlined by the discovery that the MRS frequency peak in transcript start regions occurs in different places depending on the genomic distance between transcript and CDS start. It was not possible to undertake similar detailed analysis of the precise location of the MRS frequency peak for all the species described in this chapter. However, this analysis did serve to highlight the potential complexity that may be observed in the MRS frequency patterns of other species. It may be that the *trans* splicing mechanism employed in *C. elegans* is related to the variable positioning of the MRS frequency peak in start regions, though no association with spliced leader addition and the presence of MRS upstream of genes was found when analysed in chapter 3.

However, this does not rule out the existence of a more complex relationship, perhaps also involving distance from the transcript start to the translation initiation site.

In summary, the detailed analysis of the MRS frequency around *C. elegans* genes identified how the start and stop region MRS frequency peaks were orientated in relation to genes. Furthermore, using this *C. elegans* data as a model allowed a more focussed approach to be taken in subsequent analyses when the other species were considered.

4.5.2 MRS and AT in other species

Various refinements to the transcript and CDS datasets have been described, the aim of which was to improve the accuracy of the average MRS frequency and AT content pattern around the genes. There now follows a summary of the final results obtained, with particular reference to peaks in MRS frequency surrounding genes and their relationship with AT content, for each of the six species studied.

4.5.2.1 A. thaliana

The MRS frequency patterns around the transcripts and CDS of this organism are difficult to reconcile. The MRS frequency reaches a peak at around 600 bp upstream of the CDS start. We would therefore expect an MRS frequency peak to occur within 600 bp of the transcript start, however no clear peak was observed anywhere in the transcript start region sequence. The reverse is true of the stop region sequences, where a clear peak in MRS frequency at the transcript stop is not matched by a peak near the CDS stop. These two discrepancies could be explained by the wide range of different distances between the transcript and CDS start and stop sites.

The AT content in *A. thaliana* CDS and the surrounding sequence follows a pattern common to several of the species studied here. From the centre of the CDS, the AT level steadily drops to the start and stop sites. From here is rises sharply to a consistent level of the surrounding non-coding DNA. The AT profile of genes largely reflect these changes, although there are small fluctuations over just a few base pairs at the transcript start and stop sites. While the MRS frequency peak at the transcript

stop coincides with the peak in AT content at this location, the MRS frequency peak upstream of the CDS start occurs in a region of constant AT.

4.5.2.2 C. elegans

The MRS frequency peak about 100 bp upstream of the transcript start is broader and less distinct in the 'non-diff 0' with 1000 bp space dataset than in the less refined datasets, but it is clear nonetheless. The corresponding peak, which occurs slightly further upstream of the CDS start, is narrower suggesting that this peak may be orientated about the CDS start. A larger MRS frequency peak is situated on the transcript stop site, in this case the corresponding peak centred 50 bp downstream from the CDS stop is lower and broader.

The symmetrical AT pattern of the CDS is characterised by a peak and trough at the CDS start and stop, with the peak at the stop slightly higher. In transcript sequences the pattern at the start region is similar to that in CDS, although the peak and trough are tighter at the transcript start. Unlike the CDS, there is no trough preceding the AT content peak at the transcript stop site. As the AT pattern around the *C. elegans* and *C. briggsae* CDS is nearly identical, it is reasonable to predict that the AT pattern in the transcript sequences would also be very similar.

4.5.2.3 C. briggsae

A single MRS frequency peak was observed in the CDS based sequences, situated about 200 bp downstream of the CDS stop site. The broadness of this peak suggests that it is orientated about the transcript stop site but this cannot be confirmed as transcript annotation for *C. briggsae* was not available. The start of the CDS sequence was marked by a peak, immediately followed by a trough, in AT content. This pattern was mirrored at the CDS stop site where a trough in AT content was followed by a peak. In contrast to the non-nematode species studied here, the AT content reaches a constant level within the 200 bp windows at the beginning and end of the CDS.

4.5.2.4 D. rerio

The MRS frequency in both transcript and CDS sequences was generally very high, particularly upstream of transcripts and CDS. The peak in MRS frequency upstream of transcripts occurred at 500 bp from the start site, with a corresponding peak 700 bp upstream of the CDS start site. As the peak was higher and narrower in the analysis of the transcript sequences, it is likely to be orientated in relation to transcript starts, rather than the CDS start sites. No peaks in MRS frequency in the stop regions were observed, despite a high but narrow peak in AT content at the transcript stop sites.

In general, the AT content of the CDS sequences was similar to that of *A. thaliana;* high in the centre, lowering towards the CDS boundaries and then rising fairly sharply to a constant level in the non-protein coding DNA. In contrast, the AT profile of the gene sequences is not symmetrical, the AT level drops sharply near the transcript start, while it rises more gradually approaching the stop sites. There are small but sharp fluctuations in AT content on both the transcript start and stop sites.

4.5.2.5 D. melanogaster

The MRS frequency, in particular for the transcript sequences, was notably variable, an apparent consequence of the lower number of sequences used for analysis due to the refinement of the sequence data sets. This variation made identification of distinct MRS frequency peaks difficult, although the AT peak at the transcript stop site stands out above the background variation. However, a corresponding peak downstream of the CDS stop site is not clearly identifiable. Given the degree of variation in MRS frequency in the sequences and the effect of variable distance between transcript and CDS stop sites, this is perhaps not surprising.

The AT content of the CDS sequences follows the same, symmetrical pattern previously described for *A. thaliana* and *D. rerio.* The AT content of the transcript sequences also bears similarity with these species. It is characterised by a relatively low AT at with small, sharp fluctuations at the start site and rise in AT towards the stop site, culminating in a sharp peak that coincides with the MRS frequency peak.

4.5.2.6 H. sapiens

The MRS frequency is generally very low. However, a clear peak in MRS frequency was observed at the transcript stop, coincident with a high, narrow peak in AT content. No correlating peak was observed in the CDS sequences, probably because of the large range of distances between transcript and CDS stop sites.

The AT content is more complex than that observed in the other five species. Within both the CDS and transcript sequences the AT level 200 bp from the start site was much lower than 200 bp from the stop site. Another feature unique to *H. sapiens* among the species studied here is that the decline in AT towards the transcript and CDS start sites stretches back over at least 1000 bp.

4.5.2.7 Summary of MRS frequency patterns

When considering all six species there are several characteristics that they all share. The start and stop of CDS sequences are marked by troughs in AT content. All the transcript stop sites are AT rich. With the exception of the nematodes the transcript start sites are marked by a series of small fluctuations in AT content, indicative of consensus sequences for transcription initiation. Table 4.8 summarises the presence or absence of a peak in MRS frequency and its position relative to transcript and CDS start and stop positions.

	Relative position of MRS frequency peak (bp)			
	Transcript start	**CDS start**	**Transcript stop**	**CDS stop**
A. thaliana	not clear	**600**	**0**	not clear
C. briggsae	no data	none	no data	200
D. rerio	**500**	700	none	none
D. melanogaster	not clear	not clear	**0**	not clear
C. elegans	100	**125**	**0**	50
H. sapiens	none	none	**0**	not clear

Table 4.8 Relative position of MRS frequency peaks in six eukaryotes

Figures in bold show that the peak is orientated about this position.

In terms of MRS frequency, the most consistent pattern is for an MRS peak at the transcript stop. Only *D. rerio* lacks a peak at the transcript stop site and a peak would almost certainly have been observed in *C. briggsae* had the data been available. Furthermore, in each case where they were observed, these peaks were judged to be orientated about the transcript stop site. The transcript stop sites were also characterised by a peak in AT content. The co-occurrence of high AT and high MRS frequency may be related, however there are several reasons that mean that the relationship, if present, is not a simple one. Firstly, the peaks in AT and MRS are not located in precisely the same location, as was shown for *C. elegans* in Chapter two. Secondly, the peak in AT content observed at the stop site of *D. rerio* transcripts is not matched by a peak in MRS frequency. Furthermore, there is no close relationship between AT content and MRS frequency at the start region of transcripts. Whatever its relationship with AT content, the consistency of an MRS frequency peak on the transcript stop site across such a wide evolutionary range is striking and is further evidence of a functional role.

In contrast, an MRS frequency peak in the start region of transcripts is a much less definitive feature. A clear peak, consistent between both transcript and CDS sequences, was observed in the start region of just two species, in each case a considerable distance from both the transcript and CDS start sites. It is clear, therefore, that whatever role the MRS plays at the transcript stop site, it does not play at the transcript start site. In *C. elegans* and *D. rerio,* the start region MRS peak may be involved in a functionally related role to the stop site MRS, or it may have an entirely different function.

The concurrent study of AT content around genes proved to be a useful tool in helping to refine the datasets. For example, the broad rise in AT content, particularly obvious upstream of the *A. thaliana* and *D. melanogaster* transcript start sites, was a clear characteristic of the distortion to the data caused by overlapping and closely apposed genes. The novel solution of removing from the analysis genes in which there was another gene in close proximity revealed an AT content that was much

more stable and consistent with the other species (for which overlapping genes were a far rarer occurrence). However, this phenomenon appears to have gone uncorrected in two previous reports of nucleotide frequency around genes. In the analysis of nucleotide frequency around the *D. melanogaster* transcript start by Aerts *et al.*, the A and T nucleotide frequencies are observed to rise in a broad hump several hundred bp upstream of the transcript start [74]. Similarly, a dip in GC content (equivalent to a rise in AT content) upstream of the *D. melanogaster* start site is visible in the data presented by Zhang *et al*. [77]. In both these cases the data they present matches the data generated for *D. melanogaster* gene sets not corrected for overlapping genes that is presented in this chapter. It should be noted that the data presented by Aerts *et al*. and Zhang *et al*. is not incorrect, it is a true representation of the average nucleotide content of all genes. However, by not taking into account overlapping and closely apposed genes, they have missed an opportunity to present a more instructive picture of the nucleotide content around the genes of *D. melanogaster.*

Comparative analysis of the six species studied here can provide further insight into the relationship between MRS frequency and AT content. If AT content and MRS frequency were linked in a close, direct relationship then we would expect that the number of MRS for a given level of AT would be constant between species. In the event of a less close relationship we would expect at least the relative change in MRS frequency for change in AT content to be similar across species. Figure 4.15 shows the MRS frequency observed for specific AT levels. Each of the six species is represented by two data points taken from areas of relatively stable AT content, one inside the genes and one outside the genes. The wide spread of points shows that there is no consistent relationship between number of MRS and sequence AT content. In addition, there is a wide variation in the gradient of lines connecting the two data points from each species. This shows that for a given change in AT content, the change in MRS frequency is different for each species. However, the species do fall into three groups based on the gradient of the change in MRS frequency for AT change. The two nematodes have the steepest gradient, followed by *H. sapiens* and *A. thaliana,* then *D. rerio* and *D. melanogaster.*

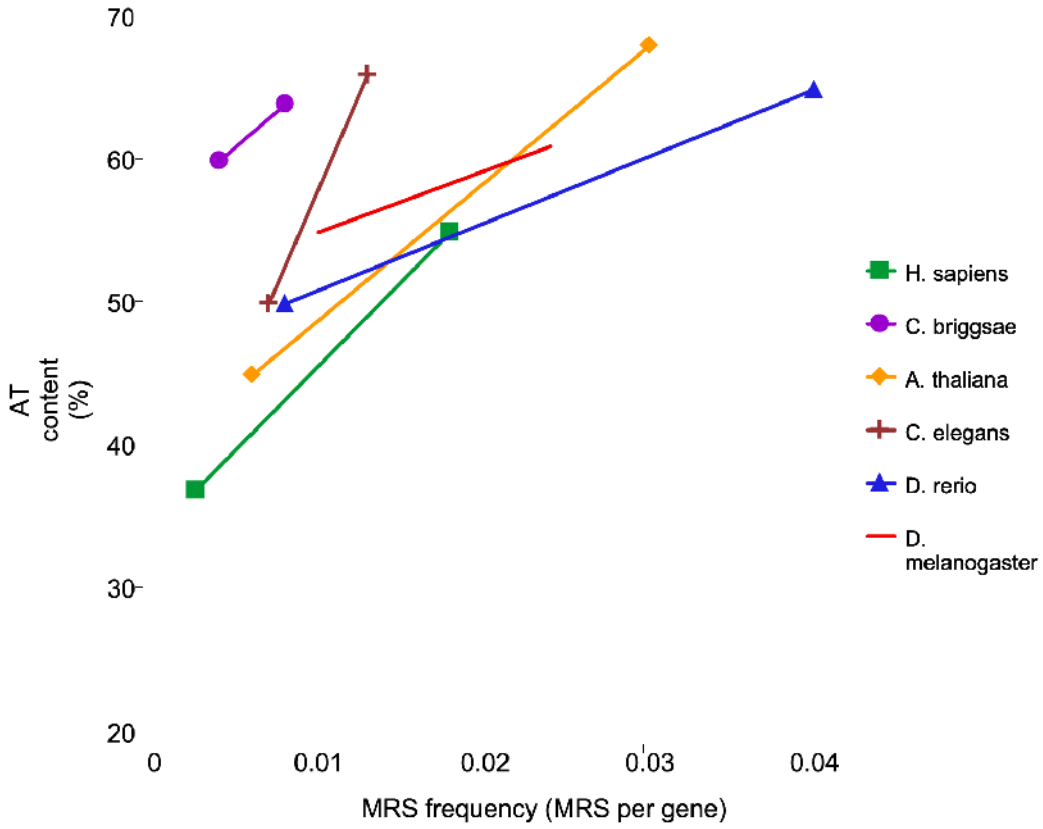


Figure 4.15 MRS frequency and AT content for six eukaryotes

Chapter 5 – General Discussion

5.1 The MRS is a functional element

In the preceding chapters the MRS has been intensively studied, primarily in the genome of *C. elegans* but also in the genomes of five other species. The main aims were to establish if the MRS could be regarded as a faithful predictor of MAR and to ascertain what function, if any, the MRS may have. The progress towards achieving these aims is discussed below but first a summary of the main findings is presented.

Although the configuration of the MRS allows its two constituent motifs to overlap, in most incidences of the MRS in the *C. elegans* genome this does not occur. Both motifs occur in excess of the MRS and influence its frequency. The incidence of the MRS in the genome of *C. elegans* is different to that expected of a randomly occurring motif. Its frequency in random sequence, however defined, is different to that found in actual genomic sequence, so its pattern of occurrence is not dictated purely by nucleotide content. The spacing between MRS is different to that of a randomly occurring element and MRS display a chromosomal distribution similar to that of genes. The MRS was not found to be over-abundant in intergenic and intronic regions, as is expected of MAR. The MRS is specifically enriched in the UTR of genes and there are striking peaks of MRS frequency in the regions flanking *C. elegans* CDS. The precise location of the upstream peak was difficult to discern, partly due to incomplete transcription start site annotation because of *trans* splicing to some of *C. elegans* mRNAs. However, the downstream peak in MRS frequency was found to lie on the gene transcription stop site. This peak in MRS frequency at transcript stop sites is common to a wide spectrum of species, absent only in *D. rerio* of the species studied here. The presence of an MRS in the CDS stop region is conserved in a significant number of *C. briggsae* orthologs to *C. elegans* genes. However, the MRS frequency peak close to *C. elegans* CDS start sites is not found in

C. briggsae, nor any of the other species studied here. *C. elegans* genes harbouring a MRS near the CDS stop site were found to be significantly enriched for the GO term 'receptor activity'. They also have higher expression levels than other genes. There is no correlation between the MRS-genes and operon position or mRNA *trans* splicing.

5.2 Does the MRS predict MAR?

One of the methods used in this thesis to assess the MRS as a predictor of MAR was the comparison with predictions generated using the SMARTest tool. An obstacle to the direct comparison of MRS and SMARTest was that the MRS does not represent MAR. Although a method for creating MAR predictions from MRS was developed, it was not able to satisfactorily define the boundaries of MAR. A further problem with comparison against SMARTest is that it is a far from perfect MAR prediction tool. Nonetheless, the rationale for comparison with SMARTest was that it was used for the only genome-wide study of MAR [58]. In that study, Rudd *et al.* reported that SMART-MAR were excluded from genes and exons in *A.* thaliana [58]. This is what would be expected of MAR and the same was found to be true of MRS. After correcting for AT content, the genes and exons of *C. elegans* were no longer depleted in MRS. But as SMARTest is also largely based on AT rich motifs, it is likely that should an AT content correction to be applied to the *A. thaliana* SMARTest data, a similar trend would emerge. So it appears that MRS in *C. elegans* and SMART-MAR in *A. thaliana* both exhibit the same pattern of reduced relative abundance in genes and exons that is expected of MAR. However, no spatial association was found between *C. elegans* genes and SMART-MAR. Another aspect in which MRS and SMART-MAR differ is in the expression levels of the genes they are associated with. *A. thaliana* genes containing a MAR were found to have relatively low expression levels [58, 73]. In contrast, *C. elegans* genes with an MRS near the CDS stop were found to have relatively high expression levels. Therefore, it seems that the presence of MRS and SMART-MAR near or on genes have different consequences for gene expression. However, MAR are known to act as both gene expression enhancers and repressors, so differing effects on gene expression levels are not necessarily

incompatible with both SMART-MAR and MRS-MAR representing actual MAR. The concept of multiple types of MAR has also been introduced [46] and it is possible that MRS and SMARTest predict different types of MAR.

Another method of assessing the likelihood that MRS predict MAR is to determine if MRS distribution is compatible with the proposed functions of MAR. MAR have been implicated in the control of gene expression through mediation of chromatin loop positioning. It is plausible that in gene-rich regions finer control of loop positioning is required due to the greater density of genes. In this scenario, we might expect these regions to be correspondingly MAR-rich, in order to facilitate the finer loop control. Therefore, the coincidence of a high frequency of MRS in gene-rich regions of *C. elegans* chromosomes is consistent with MRS representing MAR. However, many of the MRS in the gene-rich regions are located very close to, or within, genes. In seems unlikely that points of attachment to the nuclear matrix would occur so close to coding sequence. Indeed, where MAR are used to stabilise and enhance the expression of transgenes, they are generally positioned several kilobases from the transcribed regions.

There are several areas in which the work presented in this thesis could be extended to help resolve the question of whether MRS predict MAR. For example, the MRS could be studied in mitochondrial DNA which resides in a different environment to that of nuclear DNA. Mitochondrial DNA is positioned by attachment to the mitochondrial inner membrane [100, 101]. As the nature and structure of the nuclear matrix and the mitochondrial inner membrane differ, it is likely that mitochondrial DNA does not contain sequences that would be experimentally defined as MAR. Indeed, mitochondrial DNA has been used as a control when measuring the MAR affinity to the nuclear matrix [102]. In their analysis of the *A. thaliana* genome using SMARTest, Rudd *et al.* reported that no MAR predictions were found in the 620 kb mitochondrial DNA insertion on chromosome 2 [58]. Therefore, extending the description of MRS incidence to mitochondrial genomes, where MAR are not expected, would be useful in assessing MAR prediction by MRS.

In summary, the analyses in this thesis do not provide a definitive demonstration that the MRS does, or does not predict MAR. Ultimately, the only truly accurate way of testing the MRS, or any other MAR prediction method, is to compare against a large number of experimentally defined MAR. Surprisingly few MAR have been experimentally defined so far, but new sequencing technologies may make genome-wide determination of MAR a realistic proposition. Until then, conclusive evidence for the ability of the MRS to predict MAR is likely to remain elusive.

5.3 What functional roles may the MRS play?

Although the ability of the MRS to predict MAR remains in doubt, much of the work presented here does not exclude some form of functional role for the MRS. The effects of purifying selection mean that functional sequence is conserved between species. Although there is no evidence the MRS are conserved to the extreme levels reported recently for some non-coding sequences, the presence or absence of an MRS was conserved to a certain extent in orthologs of *C. elegans* and *C. briggsae*. One factor that may have limited greater conservation of MRS between these orthologs is the relatively high rate of divergence between *C. elegans* and *C. briggsae* [3, 67]. It is possible that a comparison of orthologs from less diverged species, such as within orders of mammals, may identify a higher degree of MRS conservation. Additionally, recent evidence has shown that many experimentally defined functional elements are evolutionarily unconstrained [14-16] and the MRS may fall into this category of elements.

The MRS therefore, has the potential to be a functional element. The most distinctive characteristic of the MRS identified in this thesis is the high incidence of MRS at the gene transcript stop site. The most obvious role for the MRS is therefore some kind of involvement in the termination of transcription. The MRS may represent a binding site for factors promoting the disassociation of the RNA polymerase complex. Alternatively, the sequence of the MRS may cause a secondary structure in DNA to form that acts as a barrier to further progression of transcription machinery. If the MRS is involved in transcription then it is surprising that no correlation between *C.*

elegans operons and MRS-genes was observed. There may be a number of reasons for this. Firstly, it is possible that transcription termination in operons follows a different mechanism to singly transcribed genes, although this seems unlikely. Alternatively, by looking only at MRS-genes in operons but not specifically measuring MRS frequency around operons, it is possible that a relationship between MRS and operons exists but failed to be detected. Another possibility is that the MRS is only involved in transcription termination in a certain class of genes, for example those with 'receptor activity' GO annotation, and that those genes do not occur in operons.

The association of the MRS with 'receptor activity' annotated genes means that it could be specifically involved with transcriptional regulation of these genes. The MRS may be a binding site for a transcriptional regulator that controls the expression of 'receptor activity' genes, or it may be involved in the post-transcriptional control of these genes. Due to the proximity of MRS peaks to the transcription stop site, it is likely that many MRS in this region are transcribed. This opens the possibility that MRS may play a role in mRNA, involved in for example, stability, processing or transport of the nascent mRNA strand.

Several routes may be taken to further investigation of MRS with the specific aim of defining its potential function. One area which may benefit from improved methodology is the definition of MRS-genes. The classification of MRS-genes based on the presence of an MRS in a relatively wide region surrounding genes was perhaps too loose. Tightening the definition of MRS-genes to include only those genes in which an MRS was found on the transcript stop site may allow further functional correlations to be identified. The initial intention of defining MRS-genes was to capture genes that contributed to the peaks of MRS frequency flanking *C. elegans* CDS. However, analysis of other MRS frequency in other parts of the gene structure, such as first introns, may also prove worthwhile. One of the most convincing pieces of evidence that the MRS is functional, is the common occurrence of a peak in MRS frequency at the gene transcription stop site in wide range of

species. Thus, comparison of appropriately defined MRS-genes from multiple species may prove valuable. For example, are MRS-genes from other species significantly enriched for 'receptor activity'?

Another intriguing potential function for MRS comes from work conducted by Donev *et al.*[61]. They found that upon activation of adjacent genes, MRS-containing MAR actively recruited heterogeneous nuclear ribonucleoprotein A1 (hnRNP-A1), which is involved in mRNA transport and alternative splicing. Associating the presence of MRS with genes that are alternatively spliced would provide evidence for the involvement of the MRS in this process. Donev *et al.* identified a 35 bp binding site for hnRNP-A1 in the MAR but it is GC rich and not related to the MRS [61]. However, it is possible that the MRS is involved indirectly in, for example, exposing the binding site to hnRNP-A1.

There are also two areas in which the MRS could be functional that do not depend on direct association with genes. Firstly, regions of high MRS frequency in *C. elegans* chromosomes coincide with regions of low recombination, as well as gene rich regions. Therefore, there is potential for the MRS to act as a mediator of recombination, protecting certain domains from possibly deleterious break-points. This could be investigated by identifying if the MRS has a high frequency in chromosomal regions with a high recombination rate in other species. This would not currently be possible in all species due to the need for full chromosome assemblies. A good candidate for initial study would be *A. thaliana,* as virtually all its genome is assembled into chromosomes. It has the added advantage that gene-rich regions and regions with high recombination rates exist in different parts of the chromosomes, making it possible to analyse association of MRS with recombination rates.

Secondly, in their original description of the MRS, van Drunen *et al.* suggest that the two parts of the MRS are brought together when wrapped round a nucleosome [42]. They offer some evidence that the two motifs of the MRS are found at the dyad centre or the entry/exit point of the nucleosome. However, their analysis was conducted using just a few MRS. With the advantage of a complete genomic set of

MRS and the recent availability of improved nucleosome position maps it should be possible to clarify the position of the component parts of the MRS with respect to nucleosomes. If the two parts of the MRS are found to be adjacent when wrapped around the nucleosome, then this may suggest that the MRS is a protein binding site, the activation of which is intimately related to nucleosome positioning.

At this stage the possible functions of the MRS remain open to speculation. However, given the non-random distribution of the MRS, the significant enrichment of MRS-genes for 'receptor activity' and the striking peaks of MRS frequency at the stop sites of genes from a wide cross section of species, there is evidence that the MRS is functional in some way.

References

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

2. Saccharomyces Genome Database - Available at: http://yeastgenome.org/.

3. The C. elegans Sequencing Consortium (1998). Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science *282*, 2012-2018.

4. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A.

P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

5. GOLD: Genomes OnLine Database Homepage - Available at: http://www.genomesonline.org/.

6. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K.,

Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S., Zdobnov, E. M., Zody, M. C., and Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-62.

7. Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K. C., Cooney, A. J., D'Souza, L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., Tuzun, E.,

Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B. J., Young, J. M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Albà, M., Abril, J. F., Guigo, R., Smit, A., Dubchak, I., Rubin, E. M., Couronne, O., Poliakov, A., Hübner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H. J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M. I., Kwitek, A. E., Lazar, J., Pasko, D., Tonellato, P. J., Twigger, S., Ponting, C. P., Duarte, J. M., Rice, S., Goodstadt, L., Beatson, S. A., Emes, R. D., Winter, E. E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R. C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T. D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S. M., Cooper, G. M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E. A., Venter, J. C., Payseur, B. A., Bourque, G., López-Otín, C., Puente, X. S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V. B., Caspi, A., Tesler, G., Pevzner, P. A., Haussler, D., Roskin, K. M., Baertsch, R., Clawson, H., Furey, T. S., Hinrichs, A. S., Karolchik, D., Kent, W. J., Rosenbloom, K. R., Trumbower, H., Weirauch, M., Cooper, D. N., Stenson, P. D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R. R., Taylor, M. S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., and Collins, F. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493-521.

8. Cooper, G. M., Brudno, M., Stone, E. A., Dubchak, I., Batzoglou, S., and Sidow, A. (2004). Characterization of evolutionary rates and constraints in three Mammalian genomes. Genome Res *14*, 539-48.

9. Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., and Elgar, G. (2005). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. PLoS Biology *3*, e7 EP -.

10. Bejerano, G. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

11. Dermitzakis, E. T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., and Antonarakis, S. E. (2003). Evolutionary Discrimination of Mammalian

Conserved Non-Genic Sequences (CNGs). Science *302*, 1033-1035.

12. Glazov, E. A., Pheasant, M., McGraw, E. A., Bejerano, G., and Mattick, J. S. (2005). Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res. *15*, 800-808.

13. Vavouri, T., Walter, K., Gilks, W., Lehner, B., and Elgar, G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biology *8*, R15.

14. Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi,

F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

15. Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X., Biggin, M. D., and Eisen, M. B. (2006). Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. PLoS Comput Biol *2*, e130.

16. McGaughey, D. M., Vinton, R. M., Huynh, J., Al-Saif, A., Beer, M. A., and McCallion, A. S. (2008). Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. *18*, 252-260.

17. Pennisi, E. (2007). GENOMICS: DNA Study Forces Rethink of What It Means to Be a Gene. Science *316*, 1556-1557.

18. Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet *2*, 919-29.

19. Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat. Rev. Genet *6*, 151-7.

20. Dermitzakis, E. T., Kirkness, E., Schwarz, S., Birney, E., Reymond, A., and Antonarakis, S. E. (2004). Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. Genome Res *14*, 852-9.

21. Glazko, G. V., Koonin, E. V., Rogozin, I. B., and Shabalina, S. A. (2003). A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet *19*, 119-24.

22. Robinson, P. J. J., Fairall, L., Huynh, V. A. T., and Rhodes, D. (2006). EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. Proc. Natl. Acad. Sci. U.S.A *103*, 6506-11.

23. Berezney, R., Mortillaro, M. J., Ma, H., Wei, X., and Samarabandu, J. (1995). The nuclear matrix: a structural milieu for genomic function. Int. Rev. Cytol *162A*, 1-65.

24. Platts, A. E., Quayle, A. K., and Krawetz, S. A. (2006). In-silico prediction and observations of nuclear matrix attachment. Cell Mol Biol Lett *11*, 191-213.

25. Nickerson, J. (2001). Experimental observations of a nuclear matrix. J Cell Sci *114*, 463-474.

26. Frisch, M., Frech, K., Klingenhoff, A., Cartharius, K., Liebich, I., and Werner, T. (2002). In Silico Prediction of Scaffold/Matrix Attachment Regions in Large Genomic Sequences. Genome Res. *12*, 349-354.

27. Mirkovitch, J., Mirault, M. E., and Laemmli, U. K. (1984). Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. Cell *39*, 223-32.

28. Cockerill, P. N., and Garrard, W. T. (1986). Chromosomal loop anchorage of the kappa immunoglobulin gene occurs next to the enhancer in a region containing topoisomerase II sites. Cell *44*, 273-82.

29. Michalowski, S. M., Allen, G. C., Hall, G. E., Thompson, W. F., and Spiker, S. (1999). Characterization of randomly-obtained matrix attachment regions (MARs) from higher plants. Biochemistry *38*, 12795-804.

30. Hancock, R. (2004). Internal organisation of the nucleus: assembly of compartments by macromolecular crowding and the nuclear matrix model. Biol. Cell *96*, 595-601.

31. Hancock, R. (2000). A new look at the nuclear matrix. Chromosoma *109*, 219-25.

32. Broers, J., Machiels, B., van Eys, G., Kuijpers, H., Manders, E., van Driel, R., and Ramaekers, F. (1999). Dynamics of the nuclear lamina as monitored by GFP-tagged A-type lamins. J Cell Sci *112*, 3463-3475.

33. Razin, S. V., Kekelidze, M. G., Lukanidin, E. M., Scherrer, K., and Georgiev, G. P. (1986). Replication origins are attached to the nuclear skeleton. Nucleic Acids Res *14*, 8189-207.

34. Ludérus, M. E., van Steensel, B., Chong, L., Sibon, O. C., Cremers, F. F., and de Lange, T. (1996). Structure, subnuclear distribution, and nuclear matrix association of the mammalian telomeric complex. J. Cell Biol *135*, 867-81.

35. Kumar, P. P., Bischof, O., Purbey, P. K., Notani, D., Urlaub, H., Dejean, A., and Galande, S. (2007). Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. Nat Cell Biol *9*, 45-56.

36. Gerasimova, T. I., and Corces, V. G. (1996). Boundary and insulator elements in chromosomes. Curr Opin Genet Dev *6*, 185-92.

37. Mlynarova, L., Jansen, R. C., Conner, A. J., Stiekema, W. J., and Nap, J. P. (1995). The MAR-Mediated Reduction in Position Effect Can Be Uncoupled from Copy Number-Dependent Expression in Transgenic Plants. Plant Cell *7*, 599-609.

38. Allen, G. C., Spiker, S., and Thompson, W. F. (2000). Use of matrix attachment regions (MARs) to minimize transgene silencing. Plant Mol Biol *43*, 361-76.

39. Mlynárová, L., Hricová, A., Loonen, A., and Nap, J. (2003). The presence of a chromatin boundary appears to shield a transgene in tobacco from RNA silencing. Plant Cell *15*, 2203-17.

40. Halweg, C., Thompson, W. F., and Spiker, S. (2005). The rb7 matrix attachment region increases the likelihood and magnitude of transgene expression in tobacco cells: a flow cytometric study. Plant Cell *17*, 418-29.

41. Heng, H. H. Q., Goetze, S., Ye, C. J., Liu, G., Stevens, J. B., Bremer, S. W., Wykes, S. M., Bode, J., and Krawetz, S. A. (2004). Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. J Cell Sci *117*, 999-1008.

42. van Drunen, C. M., Sewalt, R. G., Oosterling, R. W., Weisbeek, P. J., Smeekens, S. C., and van Driel, R. (1999). A bipartite sequence element associated with matrix/ scaffold attachment regions. Nucleic Acids Res *27*, 2924-30.

43. Goetze, S., Baer, A., Winkelmann, S., Nehlsen, K., Seibler, J., Maass, K., and Bode, J. (2005). Performance of Genomic Bordering Elements at Predefined Genomic Loci. Mol. Cell. Biol. *25*, 2260-2272.

44. Ma, H., Siegel, A. J., and Berezney, R. (1999). Association of chromosome territories with the nuclear matrix. Disruption of human chromosome territories correlates with the release of a subset of nuclear matrix proteins. J Cell Biol *146*, 531-42.

45. Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J. A., Lopes, S., Reik, W., and Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet *36*, 1065-71.

46. Razin, S. V. (2001). The nuclear matrix and chromosomal DNA loops: is their any correlation between partitioning of the genome into loops and functional domains? Cell. Mol. Biol. Lett *6*, 59-69.

47. Chambeyron, S., and Bickmore, W. A. (2004). Does looping and clustering in the nucleus regulate gene expression? Curr Opin Cell Biol *16*, 256-62.

48. Linnemann, A. K., Platts, A. E., Doggett, N., Gluch, A., Bode, J., and Krawetz, S. A. (2007). Genomewide identification of nuclear matrix attachment regions: an analysis of methods. Biochem. Soc. Trans *35*, 612-7.

49. Liebich, I., Bode, J., Frisch, M., and Wingender, E. (2002). S/MARt DB: a database on sca old/matrix attached regions. Nucleic Acids Res *30*, 372-4.

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

51. Benham, C., Kohwi-Shigematsu, T., and Bode, J. (1997). Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions . Journal of Molecular Biology *274*, 181-196.

52. Glazko, G. V., Rogozin, I. B., and Glazkov, M. V. (2001). Comparative study and prediction of DNA fragments associated with various elements of the nuclear matrix. Biochim Biophys Acta *1517*, 351-64.

53. Morgenstern, B., Dress, A., and Werner, T. (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Acad. Sci. U.S.A *93*, 12098-103.

54. Purbowasito, W., Suda, C., Yokomine, T., Zubair, M., Sado, T., Tsutsui, K., and Sasaki, H. (2004). Large-scale identification and mapping of nuclear matrix-attachment regions in the distal imprinted domain of mouse chromosome 7. DNA Res *11*, 391-407.

55. Liebich, I., Bode, J., Reuter, I., and Wingender, E. (2002). Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs). Nucl. Acids Res. *30*, 3433-3442.

56. Ostermeier, G. C., Liu, Z., Martins, R. P., Bharadwaj, R. R., Ellis, J., Draghici, S., and Krawetz, S. A. (2003). Nuclear matrix association of the human {beta}-globin locus utilizing a novel approach to quantitative real-time PCR. Nucl. Acids Res. *31*,

3257-3266.

57. Evans, K., Ott, S., Hansen, A., Koentges, G., and Wernisch, L. (2007). A comparative study of S/MAR prediction tools. BMC Bioinformatics *8*, 71-71.

58. Rudd, S., Frisch, M., Grote, K., Meyers, B. C., Mayer, K., and Werner, T. (2004). Genome-wide in silico mapping of scaffold/matrix attachment regions in Arabidopsis suggests correlation of intragenic scaffold/matrix attachment regions with gene expression. Plant Physiol *135*, 715-22.

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

60. Banerjee, S., and Lohia, A. (2003). Molecular analysis of repetitive DNA elements from Entamoeba histolytica, which encode small RNAs and contain matrix/ sca old attachment recognition sequences. Mol Biochem Parasitol *126*, 35-42.

61. Donev, R., Horton, R., Beck, S., Doneva, T., Vatcheva, R., Bowen, W. R., and Sheer, D. (2003). Recruitment of heterogeneous nuclear ribonucleoprotein A1 in vivo to the LMP/TAP region of the major histocompatibility complex. J Biol Chem *278*, 5214-26.

62. Shaposhnikov, S. A., Akopov, S. B., Chernov, I. P., Thomsen, P. D., Joergensen, C., Collins, A. R., Frengen, E., and Nikolaev, L. G. (2007). A map of nuclear matrix attachment regions within the breast cancer loss-of-heterozygosity region on human chromosome 16q22.1. Genomics *89*, 354-61.

63. Hukriede, N., Fisher, D., Epstein, J., Joly, L., Tellis, P., Zhou, Y., Barbazuk, B., Cox, K., Fenton-Noriega, L., Hersey, C., Miles, J., Sheng, X., Song, A., Waterman, R., Johnson, S. L., Dawid, I. B., Chevrette, M., Zon, L. I., McPherson, J., and Ekker, M. (2001). The LN54 Radiation Hybrid Map of Zebrafish Expressed Sequences. Genome Res. *11*, 2127-2132.

64. TAIR - Home Page - Available at: http://www.arabidopsis.org/.

65. Ensembl Genome Browser - Available at: http://www.ensembl.org/index.html.

66. Celniker, S. E., and Rubin, G. M. (2003). The Drosophila melanogaster genome. Annual Review of Genomics and Human Genetics *4*, 89-117.

67. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'eustachio, P., Fitch, D. H. A., Fulton, L. A., Fulton, R. E., Gri iths-Jones, S., Harris, T. W., Hillier, L. W., Kamath, R., Kuwabara, P. E., Mardis, E. R., Marra, M. A., Miner, T. L., Minx, P., Mullikin, J. C., Plumb, R. W., Rogers, J., Schein, J. E., Sohrmann, M., Spieth, J., Stajich, J. E., Wei, C., Willey, D., Wilson, R. K., Durbin, R., and Waterston, R. H. (2003). The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biol *1*, E45-E45.

68. The Arabidopsis Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

69. Nekrutenko, A., and Li, W. H. (2000). Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res *10*, 1986-95.

70. Vinogradov, A. E. (2005). Noncoding DNA, isochores and gene expression: nucleosome formation potential. Nucleic Acids Res *33*, 559-63.

71. Kent, W. J., and Zahler, A. M. (2000). Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res *10*, 1115-25.

72. Cockerill, P. N., and Garrard, W. T. (1986). Chromosomal loop anchorage sites appear to be evolutionarily conserved. FEBS Letters *204*, 5-7.

73. Tetko, I. V., Haberer, G., Rudd, S., Meyers, B., Mewes, H., and Mayer, K. F. X. (2006). Spatiotemporal Expression Control Correlates with Intragenic Scaffold Matrix Attachment Regions (S/MARs) in Arabidopsis thaliana. PLoS Comput Biol *2*, e21.

74. Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., and De Moor, B. (2004). Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics *5*, 34.

75. Louie, E., Ott, J., and Majewski, J. (2003). Nucleotide frequency variation across human genes. Genome Res *13*, 2594-601.

76. Walter, K., Abnizova, I., Elgar, G., and Gilks, W. R. (2005). Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. Trends Genet *21*, 436-40.

77. Zhang, L., Kasif, S., Cantor, C. R., and Broude, N. E. (2004). GC/AT-content spikes as genomic punctuation marks. Proc Natl Acad Sci U S A *101*, 16855-60.

78. MRSfinder - finding the MAR recognition signature - Available at: http://www.nematodes.org/bioinformatics/MRSfinder/.

79. WormBase WormBase. - Available at: ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/sequences/dna.

80. BioMart BioMart (MartView). - Available at: http://www.wormbase.org/biomart/martview/.

81. C. elegans (current release) - Available at: http://www.wormbase.org/db/seq/gbrowse/c_elegans/.

82. C. briggsae ftp site - Available at: ftp://ftp.wormbase.org/pub/wormbase/genomes/c_briggsae/sequences/dna/.

83. Workman, C., and Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res *27*, 4816-4822.

84. EMBOSS: marscan - Available at: http://emboss.sourceforge.net/apps/release/5.0/emboss/apps/marscan.html.

85. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet *25*, 25-9.

86. Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B. L., Couronne, O., Eisen, M. B., Visel, A., and Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499-502.

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

88. Stover, N. A., and Steele, R. E. (2001). Trans-spliced leader addition to mRNAs in a cnidarian. Proc. Natl. Acad. Sci. U.S.A *98*, 5693-8.

89. Vandenberghe, A. E., Meedel, T. H., and Hastings, K. E. (2001). mRNA 5'-leader trans-splicing in the chordates. Genes Dev. *15*, 294-303.

90. Pouchkina-Stantcheva, N. N., and Tunnacliffe, A. (2005). Spliced Leader RNA-Mediated trans-Splicing in Phylum Rotifera. Mol Biol Evol *22*, 1482-1489.

91. Krause, M., and Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in C. elegans. Cell *49*, 753-61.

92. WormBook - Available at: http://www.wormbook.org/toc_molecularbiology.html.

93. Nadon, R., and Shoemaker, J. (2002). Statistical issues with microarrays: processing and analysis. Trends in Genetics *18*, 265-271.

94. Genome Sciences Centre - C. elegans Resources - C. elegans SAGE - Available at: http://elegans.bcgsc.ca/home/sage.html.

95. Genomatix - understanding gene regulation - Available at: http://www.genomatix.de/.

96. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289-300.

97. WormBase - Home Page - Available at: http://ws150.wormbase.org/.

98. WormBase - WS180 Home Page - Available at: http://ws180.wormbase.org/.

99. Mizuno, M., and Kanehisa, M. (1994). Distribution profiles of GC content around the translation initiation site in different species. FEBS Lett *352*, 7-10.

100. Attardi, G. (2002). Role of mitochondrial DNA in human aging. Mitochondrion *2*, 27-37.

101. Jeong, S. Y., Rose, A., and Meier, I. (2003). MFP1 is a thylakoid-associated, nucleoid-binding protein with a coiled-coil structure. Nucl. Acids Res. *31*, 5175-5185.

102. Tan, S., Bartsch, D., Schwarz, E., and Bernard, H. (1998). Nuclear Matrix Attachment Regions of Human Papillomavirus Type 16 Point toward Conservation of These Genomic Elements in All Genital Papillomaviruses. J. Virol. *72*, 3610-3622.

Appendix

Anthony A and Blaxter M (2007). Association of the Matrix Attachment Region Recognition Signature with coding regions in *Caenorhabditis elegans. BMC Genomics,* **8:***418.*