Comparative genomic analysis as a tool for locating novel functional elements in *D. melanogaster*

Matthew Garrett

Jesus College

This dissertation is submitted for the degree of Doctor of Philosophy at the University of Cambridge

Declaration of originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text

Summary

Comparative genomic analysis as a tool for locating novel functional elements in *D. melanogaster*

Matthew Garrett

This thesis explores the use of comparative genomics techniques as a method for aiding the location of novel functional elements in *D. melanogaster*. The genomic era has brought with it a large number of techniques for predicting the location and identity of a range of functional elements, from transcription factor binding sites to protein coding genes. The primary difficulty associated with these methods is determining whether a prediction is genuine or a false positive. Comparative genomics makes use of the assumption that functional elements will be more evolutionarily constrained than non-functional ones, allowing the validation of predictions by examining the degree to which they are conserved.

The thesis starts with the examination of a dataset (Ma, unpublished results) representing the sequences of expressed small RNAs in *D. melanogaster*. After an examination of the quality of the sequences and what parameters are appropriate for its post-processing, it is used to identify novel tRNA genes which are then validated by comparative analysis. The degree of conservation

of a set of putative microRNAs earlier identified using a similar technique is then examined and used to determine the probability that the dataset represents genuine microRNAs.

Further investigation is carried out into the conservation of a specific spatial arrangement of genes within the genome with the aim of determining whether it is associated with functional relationships. Finally, the conservation of elements of the secondary structure of mRNA molecules is examined in an attempt to identify further genes with a specific subcellular oocyte localisation pattern during development: predictions from this are examined experimentally.

Throughout the thesis, comparative analysis is used to identify predictions that appear likely to be functional and worthy of further study. The availability of the genomes of 12 different species of *Drosophila* allows this to be achieved in a high level of detail, providing insight into a range of functional aspects of the *D. melanogaster* genome.

Acknowledgments

I would like to thank my supervisor, Gos Micklem, for his unceasing supply of ideas, enthusiasm, advice and his long-standing belief that I might make a competent PhD student. Thanks are also due to everyone else in the group (Semil Choksi, Monique Gupta, Lalitha Sundaram and Karen Ma) for all the assistance they've given me over the years, with special thanks to Karen for kindly providing the data analysed in chapters 2, 3 and 4. I am grateful to Francois Balloux for providing important feedback on my progress through the course.

Thanks also to Lucy Wheatley of the St. Johnston lab for showing me that I was broadly capable of carrying out experimental work, and Marion Martin for putting up with a large number of questions and not objecting too strongly to me filling the lab with fruitflies.

To the countless people who have provided moral support and helped keep me sane over the years – thank you.

Contents

Chapter 1

Introduction

The genomics era arguably began in 1972, with the sequencing of the bacteriophage MS2 coat protein gene (Min Jou et al., 1972). By 1976 MS2 had had its RNA genome entirely sequenced (Fiers et al., 1976), becoming the first organism with a known genome sequence. Bacteriophage Phi X 174 became the first complete DNA genome to be sequenced in 1977 (Sanger et al., 1977). More complex genomes took significantly longer to produce – *Haemophilus influenzae* was the first free-living organism to be entirely sequenced (Fleischmann et al., 1995), but progress since then has been rapid and aided by advances in sequencing technology.

Genomic sequencing is now sufficiently fast and cheap that Genbank (Benson et al., 2007) now contains over 81 gigabases of sequence from over 77 million individual sequences. This poses a significant problem. How can meaningful insights be drawn from this quantity of data? An analogy may be drawn with attempting to determine what makes a human simply by examining a single individual. By laborious examination of each component of the human body, it might be possible to identify which parts are necessary for survival and which are more cosmetic. Having a second human would reduce the amount of work required, as anything not common between the two would be less likely to be important. Given a sufficiently large number of humans, it would be possible to come up with a reasonably solid set of core aspects of "humanness", while noting that other features (like skin and hair colour, height and facial structure) vary widely.

Comparative genomics applies this approach to genomic sequences. Functional aspects of the genome should be subject to positive selection pressure, while the rest of the genome will be under neutral selection. Given adequate time for divergence to occur, conserved sequence is highly likely to be functional. When applied to groups of species, this provides a simple yet powerful mechanism for identifying functional sequences. By looking for well conserved sequences across a range of organisms, it is possible to come to conclusions about the fundamental genomic features that are required for life, with the distance over which a sequence is conserved giving an indication as to when in evolution it developed.

This thesis describes various approaches to identifying potentially functional aspects of the *D. melanogaster* genome, coupling this to comparative analysis to judge the effectiveness of the identification. Various techniques have been developed to aid this effort..

1.1 Comparative biology

The publication of the complete amino acid sequence of bovine insulin (Ryle et al., 1955) was a groundbreaking event in biology, although even the most optimistic researcher could hardly have imagined its long-term impact on science. A companion paper (Brown et al., 1955) discussed the sequence variations between the fully-sequenced bovine insulin and partially sequenced porcine and ovine insulin. The high level of sequence conservation provided insight into the functional nature of insulin, with only three residues showing variation between the studied species.

Perhaps insulin's greatest weakness in this burgeoning field was its absence outside the animal kingdom. When the sequence of equine cytochrome C was published (Margoliash et al., 1961), it was rapidly followed by human (Matsubara and Smith, 1962), pig, rabbit and chicken (Chan et al., 1963), tuna (Kreil, 1965) and yeast (Narita et al., 1963). This culminated in perhaps the first modern comparative sequence paper in the form of Margoliash (1963), containing some prescient wisdom:

> Similar considerations must be taken into account in attempts to ascribe functional importance to certain areas of the primary structure of a set of homologous proteins solely on the basis of their invariance over a large range of the evolutionary scale. Thus, for example, the available information does not make it possible to decide whether a section, such as that extending from residue 70 to residue 80, has remained invariant as a result of strict functional requirements or whether such constancy merely reflects a particular stability of the genetic material corresponding to this sequence. Indeed, the presence of apparent genetic "hot spots" implies genetic "cold spots."

By the end of the 60s, further sequence comparisons were taking place. An early analysis was the simultaneous publication of the structure of horse haemoglobin (Perutz et al., 1960) and sperm whale myoglobin (Kendrew et al., 1960). The fortuitous choice of sperm whales[1], coupled with a ready desire to identify the structure in human tissue, meant that comparative sequence-level analysis followed quickly. Watson and Kendrew (1961) compared the amino-acid level sequence of human haemoglobin and whale myoglobin, and in short order the whale sequence was being used to support analysis of the human sequence (Hill et al., 1969).

Of course, inter-species analysis was only part of the story. Before any protein had been completely sequenced, Linus Pauling and Harvey Itano had demonstrated that a single amino acid substitution could lead to sickle cell anemia (Pauling and Itano, 1949). Itano continued work in this field, demonstrating that a wide range of diseases could be explained by small mutations in haemoglobin (Itano et al., 1956). Comparative analysis across species would continue to provide insight into the core functionality of genetic elements, but analysis between individuals of the same species would provide great insight into what made people different – for better or for worse.

By the middle of the 1960s, enough sequences were being produced that it became necessary to start providing some level of organisation. Margaret Dayhoff and Richard Eck worked to produce the first edition of the Atlas of Protein Sequence in 1965, consisting of 65 sequences. Analysis of related sequences from distantly related organisms provided insight into the mechanisms involved in protein evolution, charting a plausible method for the development of ferredoxin from a more primitive but related sequence (Eck and Dayhoff, 1966).

[1] Human tissue contained too little myoglobin to form large crystals. Whales carry significantly more in order to allow them to remain submerged for longer periods of time, which simplified the crystal analysis

1.2 Genomic sequencing technology – from gels to commodity

Analysis of protein sequences inevitably resulted in some variation being missed due to the many-to-one mapping of nucleotide sequences to amino acids. Further, it meant that analysis of non-coding sequences was impossible. Using comparative biology to study the underlying mechanisms would require the sequencing of the genome itself.

The first DNA sequencing occurred in the late 1960s, when Wu and Kaiser (1968) determined the composition and a subset of the sequence of the complementary single-stranded ends of the phage λ DNA. This was achieved by adding DNA polymerase and individual radio-labeled bases to a solution of purified phage DNA. By determining which base was incorporated into the sequence it could be inferred that the template carried the complementary base. The phage DNA could then be purified and the experiment repeated to determine the next base.

This approach had the significant drawback that it could only be used on single-stranded sections of DNA that were adjacent to double-stranded sequences. The use of oligonucleotides as primers (Padmanabhan and Wu, 1972) made it possible to initiate sequencing at arbitrary points along the DNA, but at this stage sequencing was still a laborious and time-consuming task.

A breakthrough came with the development of the "plus and minus" technique for sequencing (Sanger and Coulson, 1975). This built on the use of primers as initiation sites for sequencing, but in contrast to the previous method of adding individual nucleotide the sequences were provided with all four and allowed to grow slowly. The reactions would then be stopped and sampled, with each sample containing a collection of sequences of random lengths. A second round of polymerase activity could then be started, this time in the presence of a radiolabeled form of only one nucleotide. These samples could then be run on an acrylamide gel, with the bands corresponding to sequences which incorporated a labeled base in the second round of polymerisation. This allowed sequences of up to approximately 50 bases to be identified in one procedure, making it viable to sequence bacterial genomes.

The primary disadvantage to this approach was that runs of identical bases would result in a single band. Identifying the number of bases that each band corresponded to was therefore a task requiring a certain degree of judgment and was the major factor preventing longer sequences from being accurately identified. A later adaptation of the protocol (Sanger et al., 1977) rectified this issue by utilising labeled bases that lacked the 3′ -OH group required for extension of the double stranded region. The polymerase reaction would be terminated at the point of inclusion of one of these labeled bases, resulting in a much sharper band and practical sequencing of up to around 400 bases.

One of the problems making it impractical to automate the entire sequencing process was the use of radiolabeled nucleotides to terminate the sequences. Smith et al. (1986) described the use of fluorescently labeled molecules, with each nucleotide fluorescing a different colour. This avoided the need to perform multiple electrophoresis runs and the risk of radioactive contamination, permitting the construction of machines that could perform a sequencing run in a little over 12 hours. By using large numbers of these machines in concert it became possible to sequence entire organisms in only a few years.

Up until this point, the general approach to sequencing had been to sequence one region of the genome and then use that sequence as the starting point to design primers to sequence the next section of the genome. Although shotgun sequencing had been described in the early years of sequencing (Anderson, 1981), it wasn't until the wider availability of high-powered computing resources in the 1990s that it became a preferred method of sequencing. Genomic DNA could be fragmented and cloned into plasmids of known sequence. These provided known starting points for the priming of the reaction and so could be sequenced without any manual primer design. The sequences thus obtained could then be reassembled by looking for overlapping regions, gradually building up the sequence of the entire genome.

Despite advances in automation and parallelisation, sequencing techniques have remained fundamentally identical to that developed by Sanger in the 1970s. The 454 pyrosequencing technique (Margulies et al., 2005) was the first of a new wave of massively parallel sequencing techniques. Short DNA sequences are derived by shearing genomic DNA and linking them to the surface of a bead. They are then amplified, resulting in beads covered in a large number of identical sequences. These beads are mounted on a slide and exposed in turn to each of four labeled nucleotides. As the nucleotides are incorporated by the polymerase, pyrophosphate is released and this in turn leads to an enzyme-mediated burst of light. By recording the wells which release a burst for each nucleotide it is possible to determine the template sequence.

The 454 approach suffers from the same drawback as Sanger's original approach. If a run of identical bases occurs in the template sequence, multiple nucleotides can be incorporated almost instantaneously. Unfortunately quantification of the bursts of light from multiple incorporations isn't especially accurate, especially for runs of more than four bases.

An approach more similar to the later Sanger technique is used in the

Solexa (now Illumina) sequencing method (Bennett et al., 2005). Fragments of template DNA are ligated to a linker sequence which attaches them to a slide. After several rounds of amplification, each spot contains a large number of identical template sequence. Fluorescently labeled nucleotides are then incorporated – however, unlike the 454 approach, they are modified and block further extension of the sequence. After the labeled nucleotides have been read, the fluorescent label is cleaved off the nucleotide and extension can occur once more. This prevents the problems associated with runs of identical residues that can occur with the 454 approach, but at the cost of a reduction in the length of individual sequences. Nevertheless, a single Solexa sequencing experiment is capable of producing up to 10Gb of sequence.

High-throughput sequencing techniques provide a rapid way of obtaining deeper sequence coverage. Traditional sequencing methods may be used to generate a rough scaffold adequate for aligning the shorter reads. In turn, the high-throughput techniques may be used to generate sufficient coverage of a region to provide confidence that the sequence is correct. As a consequence, even entirely de novo sequencing of an organism may take significantly less time and cost than today while achieving better coverage and lower error rate.

The level of coverage and relative economic viability of these high-throughput techniques provides an intriguing possibility for functional element discovery. Regions that are tightly conserved between two closely related species are more likely to be functional, but this approach provides little aid in locating the functional elements that distinguish the two species. However, given a sufficiently large sample set, similar levels of variance may be expected to be observed within the gene pool of a single species thus allowing the identification of species-specific functional elements. High-throughput sequencing techniques indicate that such an experiment may be economically viable in the near future.

1.3 Prior uses of comparative genomics

The concept of locating regulatory elements by comparative sequence analysis was introduced in Tagle et al. (1988). By aligning orthologous sequences from related species, functional sequences could be identified by the "footprints" of higher conservation they left behind. This phylogenetic footprinting hinged on the fact that functional sequences are more likely to be under selection pressure to retain their primary sequence. This analysis successfully identified several novel regulatory elements, demonstrating the ability of comparative genomics to locate functional sequences.

An analysis of several genes in the mouse (Hardison et al., 1997) was used as an argument in favour of sequencing the mouse genome, opening the door to full-genome comparative analysis between two complex organisms. The release of the mouse genome sequence (Waterston et al., 2002) provided the opportunity to use this type of analysis to improve the annotation of both genomes. Predicted genes that were present in mouse but not in humans often turned out to be spurious predictions or to be derived from viral insertions. Similarly, many predictions that occurred outside regions of synteny were determined to be pseudogenes. tRNA predictions were also validated by measuring their degree of conservation, this being significantly higher in functional tRNAs than in tRNA-like insertion elements. As with previous smaller-scale analyses, regions of higher than expected conservation in upstream regions were used to identify regulatory regions. However, while greater than average, the degree of conservation of regulatory regions was significantly lower than of coding regions.

Kellis et al. (2003) described the use of comparative genomics as a method for improving understanding of functional elements in *S. cerevisiae*. By aligning the *S. cerevisiae* genome with three others (*S. paradoxus*, *S. mikatae* and *S. bayanus*, divergent over a range of 5-20 million years), regions of conservation could be examined and used to identify whether a region was likely to be functional or not. This could be related to existing annotation in order to gain improved accuracy. Some 500 existing open reading frames were determined to be spurious by noting that orthologous regions of the other genomes had accumulated frameshifts and stop codons, indicating that the region was not under selective pressure to retain a coding sequence and strongly suggesting that there was no functional transcript. Gene start and end points, introns and new ORFs were annotated via similar means, resulting in alterations to approximately 15% of existing gene annotations.

This analysis demonstrated the power of comparative genomics in validating existing predictions. Not only were putative genes validated (or removed), the comparative analysis provided extra resolution in determining the extents of genes. It even suggested that some genes that had previously been thought to have been experimentally tested didn't exist. Instead, deletions that had been interpreted as impairing the functionality of a gene had probably been disrupting the promoter region of genuine adjacent genes.

This analysis was not limited to gene annotation. By examining the conservation of motifs by three different metrics (conservation in intergenic regions, higher levels of conservation in intergenic regions than in genic regions, and a significant difference between the levels of conservation upstream and downstream of genes), it was demonstrated that the majority of known regulatory elements could be located along with a number of novel elements.

It is likely that the analysis of Kellis et al benefited from the evolutionary closeness of the species analysed. An analysis including more distantly related species (Cliften et al., 2003) proved less sensitive, which could be interpreted in two ways. The first argument would be that any sequences conserved over the larger evolutionary distance are more likely to be functional. The alternative interpretation is that the enhanced divergence makes it more likely that alternative regulatory mechanisms will have evolved, reducing the probability that elements will be located this way. This was demonstrated by Liu et al. (2004), who found that the level of conservation of the known regulator elements between *S. cerevisiae* and *S. pombe* was no better than that of randomly selected intergenic sequence.

This is perhaps the primary risk in comparative genomics – that certain areas of the genome may be conserved by chance. Increasing the distance between the species examined increases the time available for the sequences to have diverged, reducing the probability that sequences under neutral selection will still bear a strong resemblance. The disadvantage of this approach is that it reduces the ability of comparative analysis to identify elements that have evolved more recently than the divergence between the organisms being analysed – the lack of conservation may be due to the sequence not being functional, or alternatively its absence may be one of the fundamental differences between the organisms being examined.

A refinement of phylogenetic footprinting, phylogenetic shadowing, was proposed by Boffelli et al. (2003). This method takes advantage of the divergence within a group of closely related species, providing the same level of resolution without requiring the same degree of historical divergence. Multiple independently evolving (but closely related) organisms will show a similar level of sequence diversity as a comparison between two more distantly related species. Estimates suggest that the sequencing of 4 or 5 closely related (ie, within around 40 million years of each other) species would provide an adequate level of intra-group diversity to be able to identify the majority of functional conservation, and hence provide a means of identifying the majority of regulatory elements while minimising the probability that conservation is due to chance alone.

Clark et al. (2003) went further, suggesting that even further detail could be provided by a "ladder and constellation" approach. Here, divergence points would be picked (the "rungs" of the "ladder") and multiple species sequenced at each point (the "constellations"). In this model, the divergence is as wide as in a phylogenetic shadowing approach but extra detail is available at the clusters of closely related species. This gives more insight as to whether divergence is accidental, or a functional alteration in a specific subgroup of the species examined. As a result, more information about functional but rapidly changing elements of the genome can be obtained.

A consequence of this proposal was the sequencing of an additional 10 species of Drosophila, joining the already sequenced *D. melanogaster* (Adams et al., 2000) and *D. pseudoobscura* (Richards et al., 2005). These represent a period of roughly 50-60 million years of divergence between the most distantly related species, down to under a million years for the most closely related. Though only recently completed (Clark et al., 2007), these have already been used to obtain a better understanding of functional elements (Stark et al., 2007).

1.4 Sequence alignment techniques

Comparative sequence analysis involves being able to identify the analogous sequences in multiple species. The process of matching sequences with each other is known as sequence alignment. There are two main approaches used for this. The first attempts to align either the entire sequence, or significant proportions thereof and is therefore known as global alignment. It is typified by the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). This consists of a matrix of alignment scores – positive for identical bases being aligned, negative for mismatches or introduced gaps. A dynamic programming approach is then taken to find the alignment that provides the optimal overall score. A variation on this, the Smith-Waterman algorithm (Smith and Waterman, 1981), differs primarily in scoring mismatches as 0 rather than a negative number. This approach allows for shorter areas of high similarity to be aligned even if that would otherwise result in a lower score for the overall alignment. This approach is therefore known as local alignment.

The preferred approach for sequence alignment depends on the research being performed. Given two gene regions from different species, a global alignment will give a strong overview of the genomic changes caused by evolutionary divergence – for example, a gene duplication event may be indicated by the introduction of a gap in one sequence that covers the entirety of the duplicate gene. However, if segments of the duplicate bear greater similarity to the ancestral gene, a local alignment may introduce additional gaps in order to provide a better indication as to these conserved elements. An example is shown in figure 1.1.

The majority of the research in this thesis consists of looking at small se-

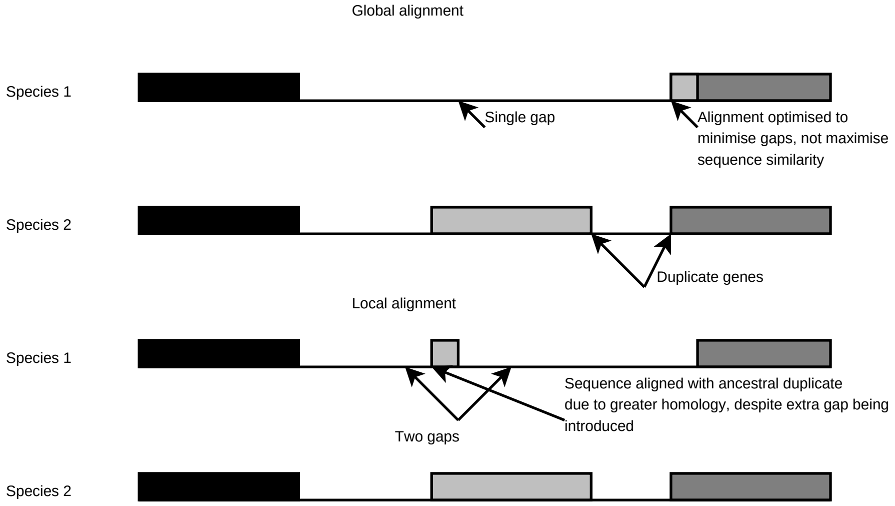


Figure 1.1: Comparison of global versus local alignments. A duplication event in species 2 has caused a new gene to be inserted. It has diverged over time, leaving the original gene with greater homology to the ancestral gene present in species 1. A global alignment will introduce a single gap, allowing visualisation of the duplication event as shown in the upper comparison. A local alignment may introduce multiple gaps in order to align segments of the duplicate gene that bear closer homology to the ancestral gene, as shown in the lower comparison

quences that are expected to be well conserved. Local alignment lends itself well to this form of investigation. Small sequences are prone to rearrangements in the "churn" of genetic divergence. Outside coding regions, there is little pressure on conservation of the sequence surrounding functional sequences. Global alignment would therefore tend to result in a failure to locate the conserved elements, leaving them poorly aligned in order to favour the alignment of larger conserved features.

The most commonly used local alignment tool is BLAST (Altschul et al., 1997). BLAST implements an algorithm similar to Smith-Waterman, but with additional heuristic steps to reduce the number of operations required. This means that BLAST is not guaranteed to give an optimal alignment. However, as a consequence it is able to perform around 50 times faster than a pure Smith-Waterman implementation.

1.5 Thesis overview

Chapter 2 introduces a dataset derived from high throughput sequencing of short RNAs from *D. melanogaster* embryos (Ma, unpublished results). The quality and error types of the sequencing reads are critically evaluated in order to identify the appropriate processing necessary for generating high-quality alignments with genomic sequences. This provides the basis for the following two chapters.

Chapter 3 describes a novel approach for identification of genetic elements by examining their distinctive expression profiles derived from the previously discussed alignments. This is used to locate unannotated tRNAs and validate the predictions of existing tRNA scanning applications, with comparative analysis used to validate the newly-discovered functional elements.

Chapter 4 examines the level of conservation of a set of microRNAs earlier identified earlier using a profile scanning technique similar to that discussed in chapter 3, along with additional criteria (Ambros et al. (2003), Ma, unpublished results). This conservation is used to evaluate the quality of the prediction technique and discuss potential mechanisms by which the observed conservation could have occurred.

Chapter 5 investigates a different kind of potentially functional relationship by examining conservation of gene nesting arrangements over the sequenced species, and uses this to estimate the rate of change of nesting arrangements. An evaluation of the significance of conserved nesting arrangements is also made.

Chapter 6 attempts to use conservational analysis as a tool for examining a relatively unexplored class of functional elements. Potentially functional RNA structure with a role in mRNA localisation are identified with novel technique for describing RNA structure, and a set of putatively localised genes identified.

Chapter 7 describes the application of experimental techniques in an in an attempt to validate the predictions made in chapter 6.

Chapter 2

Analysis of the Solexa high-throughput sequencing method

The raw data used for this analysis was provided by Karen Ma

2.1 Introduction

Microarrays have developed greatly since the early days of cDNAs being blotted onto filter paper (as introduced by Kulesh et al. (1987)) and small-scale gene expression examinations (such as Schena et al. (1995)). However, even when used to examine entire genomes (Lashkari et al., 1997), microarrays face certain fundamental issues – the difficulty in quantifying signals and difficulties with probe specificity and sensitivity. If the probes are targeted, then there must be some prior knowledge of the sequences in advance. Whole genome microarrays reduce this problem by containing probes derived from sequences regularly spaced across the whole genome, but the gaps in coverage make it difficult to locate small functional sequences.

Microarrays are therefore not an ideal tool for *de novo* functional element discovery. The ability to identify the actual gene expression of a cell at a given point in time would provide the potential for making predictions based on knowledge of which sequences are actually expressed.

The increasing availability of ultra-high throughput sequencing techniques such as those from 454 Life Sciences (Margulies et al., 2005) and the Solexa method from Illumina (Illumina, 2006) are able to provide this. The speed and number of samples that can be sequenced allows the determination of the sequence of RNA extracted from the cell and subject to minimal processing. In principle, the sensitivity allows for identification of sequences that are expressed even in very low copy numbers.

In an attempt to identify novel small RNAs in *D. melanogaster*, a dataset was generated by extracting RNA from embryos, size selecting it for sequences of between 20 and 30 bases, reverse-transcribing the sequences into DNA and

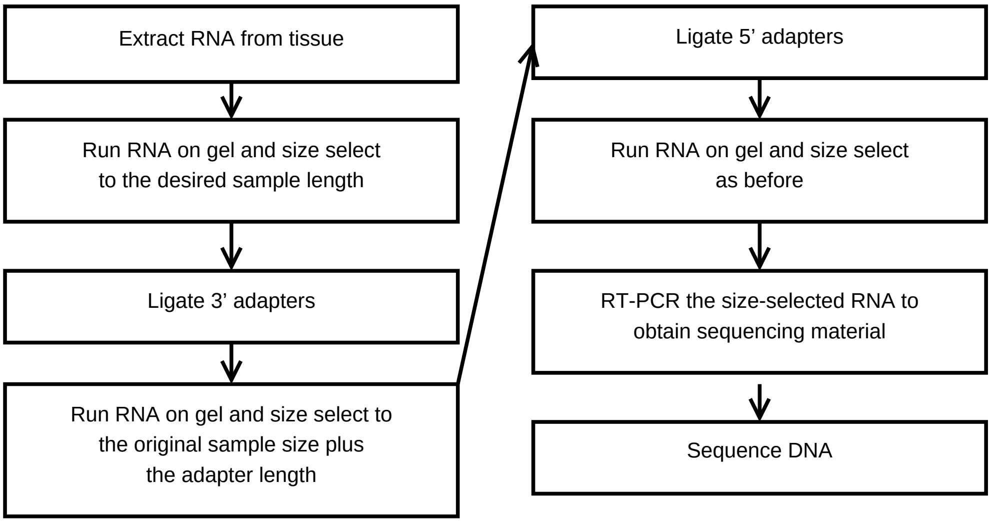

Extract RNA from tissue
Run RNA on gel and size select to the desired sample length
Ligate 3' adapters
Run RNA on gel and size select to the original sample size plus the adapter length
Ligate 5' adapters
Run RNA on gel and size select as before
RT-PCR the size-selected RNA to obtain sequencing material
Sequence DNA

Figure 2.1: Summary of source material preparation. More detailed discussion may be found in Appendix A.

then sequencing them via the Solexa method (Ma, unpublished results). The protocol for obtaining the source material is described in Appendix A, with an overview in figure 2.1. This chapter describes the initial analysis of the sequences, with the emphasis on determining the error rate in order to allow the maximum number of sequences to be confidently aligned to the reference genome.

The Solexa sequencing method

At the time of this work the Solexa sequencing process was designed for short sequences of DNA, up to around 30 bases[1]. This may be achieved by either size selection or fragmenting longer sequences. Known adapter sequences are ligated to the 5′ and 3′ ends of the DNA and these are denatured into single stranded fragments which are hybridised to primers covalently attached to flow cells. Amplification of the hybridised molecules occurs through a PCR-

[1] More recent work has allowed this to be extended to over 100 bases

like reaction in which all the primers are covalently attached to the flow cell.

A primer corresponding to the adapter sequence is then added to the flow cell and binds to the single stranded adapter at the unattached end of the sequence fragments. Labeled nucleotides that incorporate reversible terminators are then added, along with DNA polymerase. These are then incorporated into the sequence. The cell is then excited by laser light and the incorporated base identified. This allows the first base of the sequence to be identified. The terminator is then cleaved and the process of incorporation, measurement and cleavage/regeneration repeated to build up the complete sequence read.

Sources of error in the generation of the dataset

The results of the Solexa sequencing process may contain two classes of error – errors generated in the preparation stages and then correctly sequenced, and errors generated in the sequencing process itself. While it is the total error that is of most interest when it comes to interpreting the results, it is important to determine the relative proportions of the error sources in order to understand how the data can be used in further analysis.

To be able to define the importance of each of these two influences, it is necessary to posses known reference sequences. Two ideal candidates were identified. A large body of the sequenced material corresponded to the 2S ribosomal RNA, a 30 base sequence generated by cleavage of the 5.8S rRNA in *Drosophila* (Jordan et al., 1976). This was clearly derived from the original source material, and therefore had undergone every stage of the preparation process.

The second candidate sequence was that of the 3′ adapter ligated to the end of the source material. This 21-base sequence does not occur anywhere within

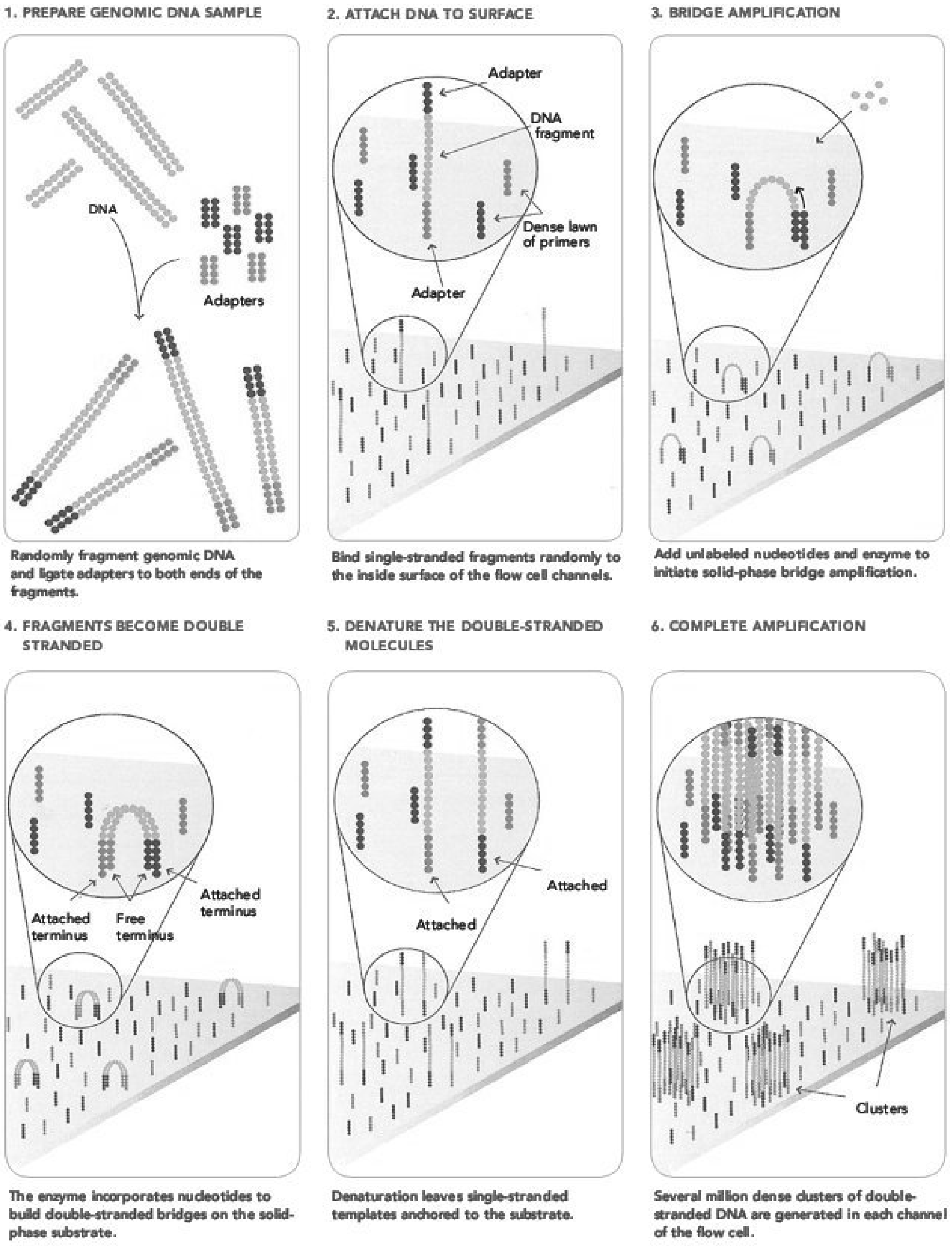

1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Free terminus
Attached terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 2.2: Diagrammatic representation of the Solexa sequencing process. Image copyright Illumina, Inc

the *D. melanogaster* genome, but is instead a synthetic oligomer of a sequence provided by Illumina. If the RNA fragment is shorter than the read length, then the sequencing process will read through the entire fragment and start sequencing this 3′ adapter. The raw material was size selected to be smaller than 30 bases, thereby making it likely that many of the original sequence fragments would be less than the 27 base read length. As a result, it would be expected that there be evidence of adapter sequence in many cases.

The 2S sequence is therefore suitable for use in determining the error rate of the entire process, with the adapter sequence being suitable for approximating the error rate of the sequencing itself. The difference between these rates should provide an estimate of the amount of error introduced during the preparation process.

Solexa provide per base quality estimates. This is an estimate of the probability of the base call being incorrect, and is based on a modified version of the scoring used by the Phred scorer used in conventional sequencing (Ewing et al., 1998). Rather than Phred's $Q = -10log_{10}(P_e)$, the Solexa scoring is $Q = 10log_{10}((1-P_e)/P_e)$ with Q being the quality score and P_e representing the probability of error. At higher quality scores (above 15 or so), the two equations give almost identical answers. The Solexa equation, however, provides a greater dynamic range and thereby makes it easier to differentiate between lower scores, representing lower qualities.

Unfortunately the data set examined did not include calibrated quality scores – that is, while quality scores were present, they did not provide any sort of absolute error rate. Instead, a lower quality value merely indicated that the probability of a base call being correct was lower than if there had been a higher quality score.

As a result, this analysis served three purposes. Firstly, it allowed an estimate of the error rate induced by the sequence preparation protocol. Secondly, it allowed an estimate of the error rate inherent in the sequencing itself. Finally, it allowed an estimate of the appropriate cutoff level when choosing which sequence data to trust in later stages of the analysis.

2.2 Methods and results

The two sequences to be tested were as follows. The 2S sequence of

```
tgcttggactacatatggttgagggttgta
```

was retrieved from Genbank (accession number GI:8456). The adapter sequence of

```
tcgtatgccgtcttctgcttg
```

was supplied by Solexa. Sequences from the dataset with significant alignment to these reference sequences were then retrieved with Blast (Altschul et al., 1997).

Each resulting dataset was then in turn analysed. A small perl application was developed to perform a gapped Needleman-Wunsch (Needleman and Wunsch, 1970) alignment between each of the sequences and the reference sequence.The number of correct matches was recorded, as was the number of incorrect matches. The number of correct and incorrect matches was noted for each different quality value, along with the proportion of correct and incorrect matches along the length of the sequence. These results were then tabulated in tables 2.1 and 2.2.

Quality	Incorrect	Correct	Error rate
-5	73547	2	0.99
-4	198	242	0.45
-3	1346	1624	0.45
-2	3656	5882	0.38
-1	6479	13180	0.33
0	24679	48047	0.34
1	32816	83111	0.28
2	24793	111494	0.18
3	19715	159561	0.11
4	15448	208407	0.069
5	12433	274114	0.043
6	10561	377586	0.027
7	9145	485434	0.018
8	7411	569917	0.013
9	6572	654712	0.0099
10	5756	804638	0.0071
11	5303	840636	0.0063
12	5126	971377	0.0052
13	4952	1097206	0.0044
14	4110	1027841	0.0040
15	4095	1167798	0.0035
16	4171	1361613	0.0031
17	3253	1332981	0.0024
17	3982	1408511	0.0028
19	2343	1006091	0.0023
20	4281	1619149	0.0026
21	1635	732701	0.0022
22	4418	2064341	0.0021
23	2491	1290095	0.0019
24	2484	1096458	0.0023
25	3217	1207858	0.0027
27	4830	2482986	0.0019
30	249127	80299275	0.0031

Table 2.1: Correct and incorrect base calls for each quality value in sequences with high-scoring alignments to the 2S reference sequence

Quality	Incorrect	Correct	Error rate
-5	13199	0	1
-4	8	35	0.19
-3	109	235	0.32
-2	318	792	0.29
-1	596	1665	0.26
0	4171	8650	0.33
1	3850	19091	0.17
2	2825	24426	0.10
3	2299	35483	0.060
4	1798	45566	0.038
5	1642	63385	0.025
6	1546	96551	0.016
7	1411	124961	0.011
8	1354	141606	0.0095
9	1366	170912	0.0080
10	1370	216368	0.0063
11	1305	224130	0.0058
12	1412	275127	0.0051
13	1535	326800	0.0047
14	1245	290238	0.0043
15	1442	326862	0.0044
16	1475	398383	0.0037
17	1392	351057	0.0039
18	1472	418960	0.0035
19	1061	246158	0.0043
20	1675	483111	0.0035
21	745	212099	0.0035
22	2209	600378	0.0037
23	1088	396530	0.0027
24	1044	368817	0.0028
25	1325	377289	0.0035
27	2056	695275	0.0029
30	73471	37557273	0.0020

Table 2.2: Correct and incorrect base calls for each quality value in sequences with high-scoring alignments to the 3′ adapter reference sequence

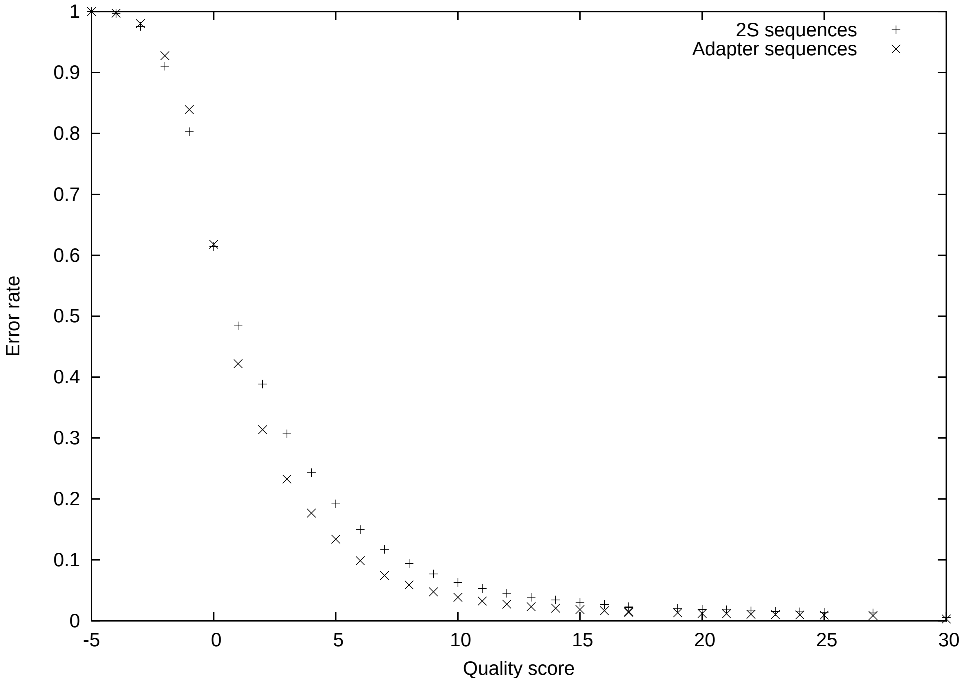


Figure 2.3: Per-base error rate for bases at each quality level

2.3 Discussion

Based on the high number of bases with Q30 scores from both sets of data, it is possible to assign a tentative error rate to each step of the process. Examining the values from the adapter-containing dataset, a reasonable estimate for the raw error rate of sequence data flagged as high quality is 0.2%. The 2S data suggests a figure of 0.3%. This difference is consistent with the hypothesis that some extra error would be introduced during the preparation process.

These results were then examined to see if the error rate was continuous over sequence length. Figures 2.4 and 2.5 show plots of error rate against position in the query sequence - ie, the position in the read that is being aligned.

Both figures 2.4 and 2.5 show a significant positive trend, indicating that there is an increase in error rate towards the end of the sequence. Assuming that a linear increase is a representative model of the error, we can approxi-

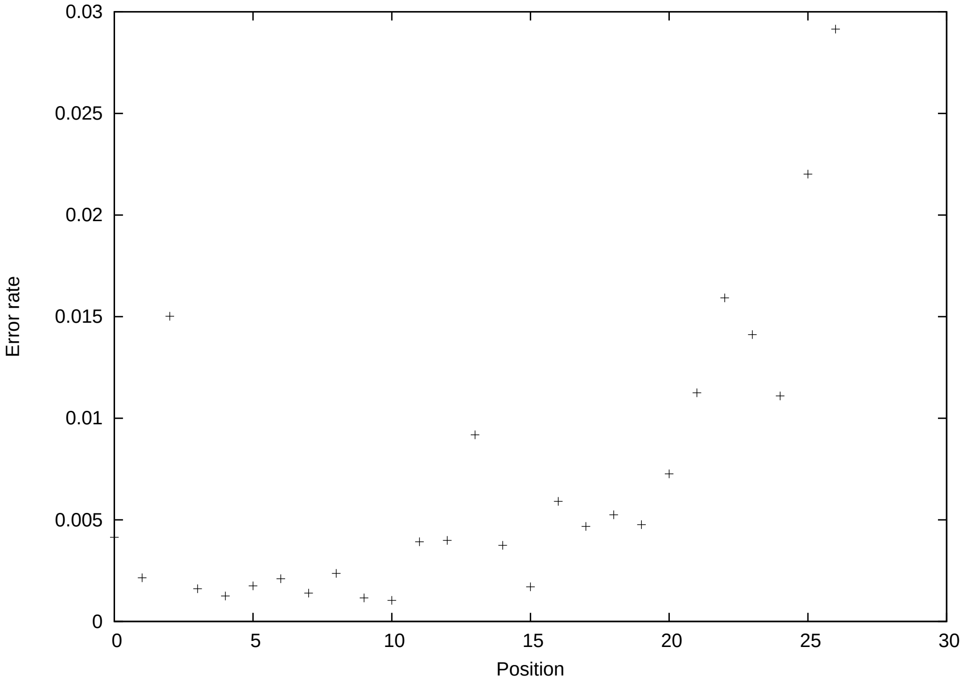


Figure 2.4: Error rate plotted against position in the query for aligned 2S sequences. Linear regression provides a slope of 5.79×10^{-4} with a variance of 5.75×10^{-10}

mate that on average a base at the end of a 27 base read is some 6 times more likely to be incorrect than a base at the beginning of the read. This represents a relatively small absolute increase - the probability of error per base is still less than 1%. This low rate suggests that the Solexa sequencing method is not highly prone to losing synchronisation. The lac of accuracy of this estimate is due to the difference in these values between the two datasets analysed. There is no obvious explanation for these differences.

Figures 2.6 and 2.7 show the error rate per position in the reference sequence. An interesting observation is the presence of two large peaks in the 2S data. The peak at position 14 can be easily explained. Of the 16 occurrences of the 2S rRNA gene precursor in the *D. melanogaster* genome, 4 share the same t to c point mutation. However, only around 1% of the expressed sequences

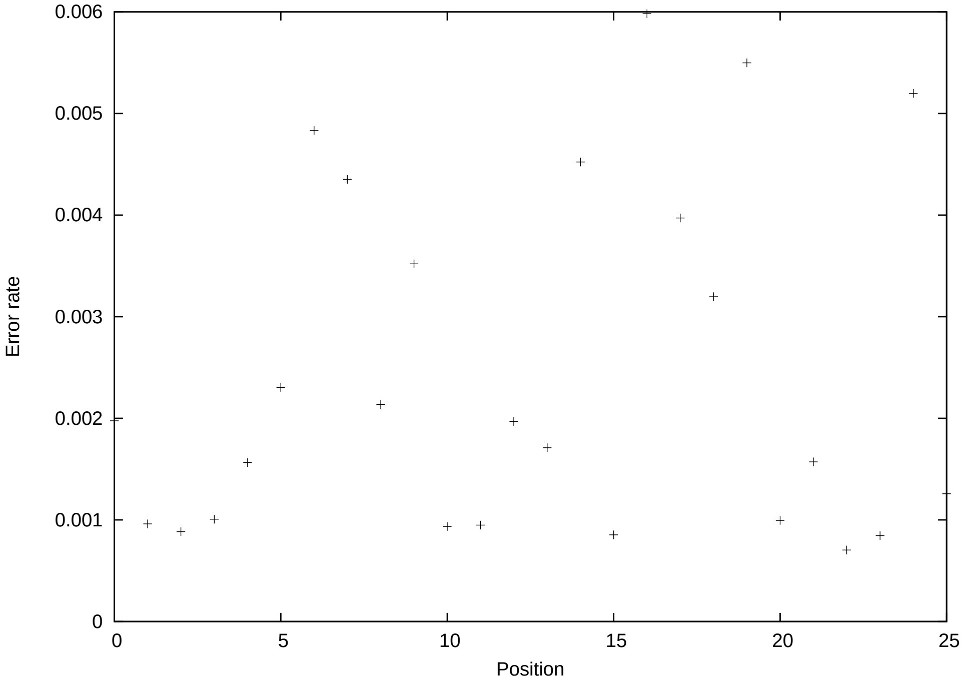


Figure 2.5: Error rate plotted against position in the query for aligned 3′ adapter sequences. Linear regression provides a slope of 1.46×10^{-4} with a variance of 2.86×10^{-10}

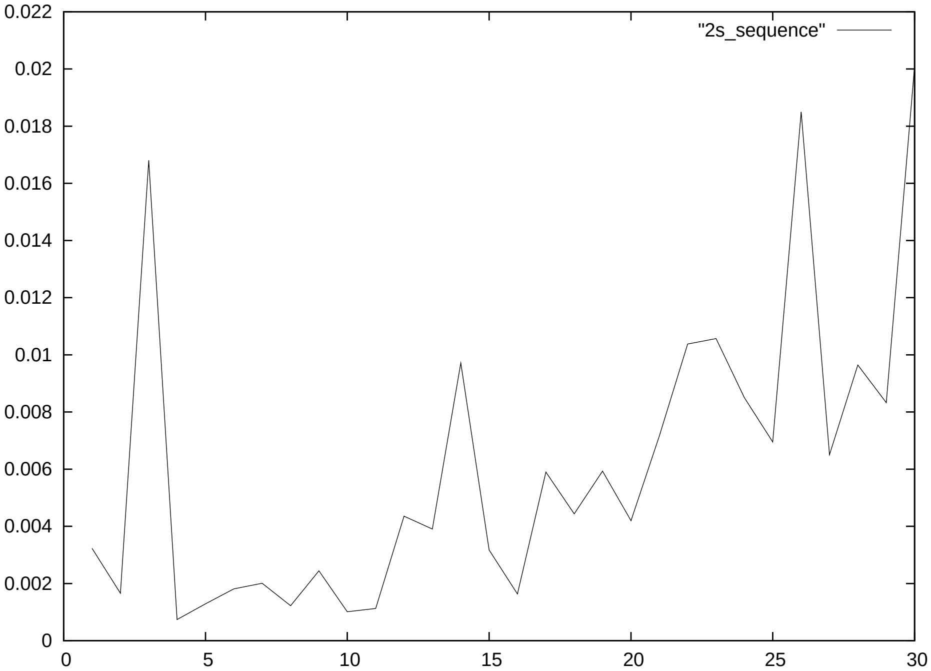


Figure 2.6: Error rate plotted against position for aligned 2S sequences. Note peaks at positions 3 and 14

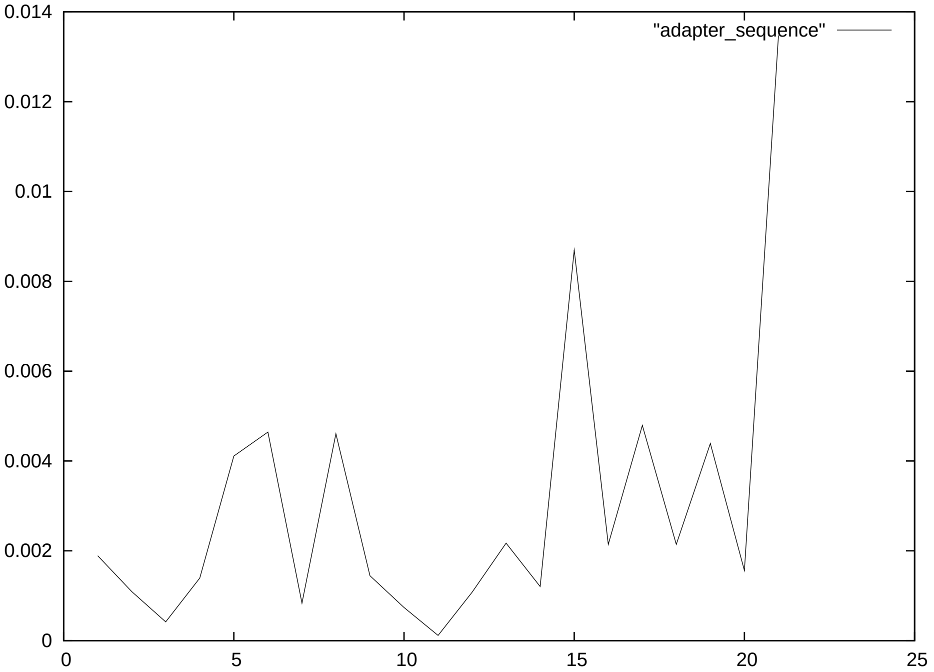


Figure 2.7: Error rate plotted against position for aligned 3′ adapter sequences

contain evidence of this. This may indicate that these altered sequences have been selected against and are expressed at a lower level during embryonic development. Interestingly, the Genbank EST library provides no evidence for this sequence. It is likely that previous sequencing methods have not been sufficiently deep to distinguish between these seemingly genuine reads and sequencing errors. Attempting to align this sequence against other drosophilids shows a similar polymorphism in *D. willistoni* (though missing the first base) but nothing in any other species.

The peak at position 3 is less easily explained. The *D. melanogaster* sequence provides no support for a mutation at this point. However, as shown in table 2.3, despite there being a significantly higher error rate at this position almost all of the error can be assigned to c to t transitions. Discounting these, the error rate for this position would be no higher than in surrounding positions.

Position	A	C	G	T
2	1500	1357	3889849	3582
3	1681	3896342	1812	62026
4	629	1582	675	3908891

Table 2.3: Number of reads with each base in the area surrounding position three of the 2S rRNA sequence. Note the much increased error rate in position 3, along with the strong bias towards a c to t transition

Given the absence of any especially strong evidence for a significant bias in miscallings, the parsimonious explanation is that this peak represents a genuine divergence from the reference sequence.

The Ma dataset was generated from embryos collected from the Oregon R strain of *D.melanogaster*, rather than the y[1]; cn[1] bw[1] sp[1] strain used in the sequencing effort. Therefore it is reasonable to propose that the Oregon R strain contains one or more copies of the 2S rRNA sequence with an additional c to t transition at position three. This is consistent with cytosine deaminating into uracil (cytosine being the least stable amino acid – Frederico et al. (1990)) without being corrected, with the uracil in turn being interpreted as a tyrosine during DNA replication. This could conceivably have occurred since the development of this specific strain in the laboratory, as the reduced stress environment may mean that the loss in some functionality of a 2S rRNA might not have a significant impact upon viability. *D. ananassae* shows a similar sequence (though with further mutations in the first two bases), demonstrating the plausibility of this being a genuine polymorphism.

As well as validating the overall quality of the Solexa sequencing, the error data is important in order to determine the threshold at which the data is trusted. This is necessary in order to allow the sequence reads to be aligned against the genome with confidence. Choosing too low a threshold will result in data containing errors which may then result in misalignments. Too high a threshold will risk discarding too much correct sequence, resulting in sequences that can now be aligned to multiple locations in the genome.

Two approaches may be taken to this analysis. The first is to treat the quality data as a classifier – that is, a means of telling whether a given base may be trusted or not. The sole parameter of this classifier would be the quality level threshold used to determine the threshold.

Figures 2.8 and 2.9 show plots of the true and false positive rate as the quality threshold at which bases are discarded is varied, a graph known as a receiver operating characteristic (or ROC) curve. In this case the true positive rate indicates the proportion of incorrect bases that would be discarded at a given threshold, while the false positive rate indicates the proportion of correct bases that would be discarded. Both curves follow a similar shape, indicating that the ability to use the quality scores to determine whether a base is correct or not is fairly consistent across the two data sets. In both cases, the point at which the true positive rate is no longer significantly increased by discarding more bases (while the false positive rate *is*) corresponds to a quality threshold of around 12. In other words, the classifier is most accurate between quality scores of -5 and 12, at which point 40% of incorrect bases can be discarded while discarding only 1% of correct bases. Higher quality scores provide less power for differentiating between correct and miscalled bases in comparison.

However, this still leaves the problem of where on this line to place the threshold. Choosing a higher threshold will save more correct data, but will discard a smaller quantity of incorrect data. The appropriate cut-off point may be aided by considering the information content of the sequences.

Each base in a sequence may have one of four different values. This may be considered to be 2 bits of information ($2^2 = 4$)[2]. A 25 base sequence would therefore have 2×25, or 50 bits of information. In other words, there are 2^{50} possible ways of writing a sequence of 25 bases.

The mean length of a sequence in the dataset after the removal of adapter sequence is approximately 25 bases. The *D. melanogaster* genome contains around 140 megabases, or approximately 2^{27} bases. The expected number of occurrences of a specific 25 base sequence in the genome[3] is therefore $2^{27}/2^{50}$, or 2^{-23} (1.19×10^{-7}). This is sufficiently small that finding a 25 base sequence that correctly aligns is likely to be due to the sequence having been derived from the genome, rather than having arisen by chance.

Determining the appropriate proportion of bases to discard is therefore influenced by the amount of information that will be lost in the process. Removing a single base from a 25 base sequence will still leave 48 bits of data, and will therefore have little impact upon the probability of correctly aligning the sequence against the genome. As shown in table 2.4, setting the threshold to 27 would result in the loss of approximately 23% of all good sequence. This would translate to an average loss of 6 bases from each 25 base sequence, reducing the information content to from 50 bits to 38 bits. At this level a random sequence might be expected to be found 2^{-11} times in the genome, or 4.88×10^{-4} – that is, after this clipping approximately one in 2048 25 base sequences might be expected to align to the genome by chance rather than because they originated from that location. 97% of sequences are 21 bases or longer. At 21 bases, the loss of 23% of bases would result in a total length of

[2] This is dependent upon the frequency of all bases being approximately equal, as can be seen with a simple thought experiment – if the genome was 99% C and G nucleotides, then vast majority of bases would be either C or G. As a consequence, the information content of most bases would be closer to 1. The true ratio of bases in *D. melanogaster* is around 56% A or T nucleotides, a figure which does not alter these calculations to any significant extent

[3] Assuming random sequences

16 bases and an information content of 2^{32}, giving each sequence a one in 32 chance of aligning by chance.

This suggests that clipping at a quality threshold of 27 would result in between 1 in 32 and 1 in 2048 sequences now mapping to more than one location in the genome. There is therefore a reasonable argument for simply discarding any data with a quality value of below 30 – the expected loss of specificity is on the order of 2% in the worst case (that is, around 2% of sequences that would previously have aligned uniquely may no longer do so), but around 50% of incorrect bases will be discarded and so the accuracy of the remaining sequence should be significantly greater.

There are therefore two arguments for determining the threshold value. One would be to set the threshold to 12, at the peak of the effectiveness of the classifier. The other would be to discard all data with a quality of less than 30 in order to discard as much incorrect sequence data as possible.

After realigning sequences with differing levels of quality masking (ie, replacing all bases below a certain quality with a wildcard able to match any base), the results shown in figures 2.10 and 2.11 were obtained.

Figure 2.10 is consistent with the ROC analysis – setting the quality threshold to 12 would include the area of greatest gain in aligned sequences. Figure 2.11 shows a value slightly worse than the expected worst case calculation for loss of specificity. The peak proportion of uniquely aligned sequences (that is, sequences which are aligned to a unique location on the genome) is 0.596 at a quality threshold of 17. At a quality threshold of 27 (ie, discarding all bases with a quality score of 27 or lower), the proportion is 0.574. This indicates that around 3.7% (or 1 in 27) of sequences lost specificity. The calculated value assumed an approximately random distribution of bases. In reality, gene

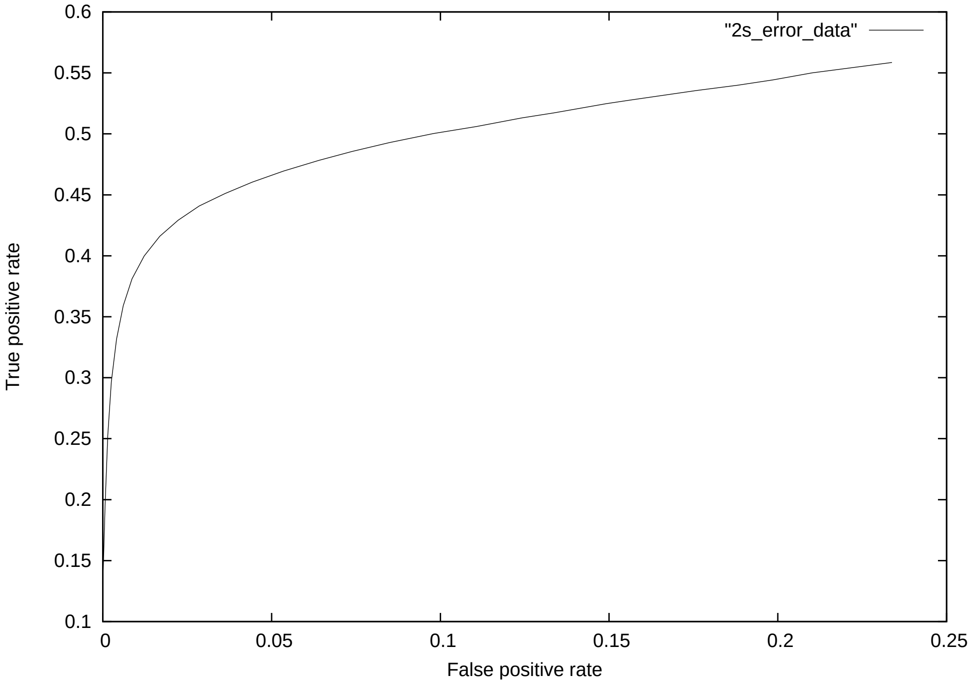


Figure 2.8: ROC curve plot of true and false positive rates for sequences aligned against the 2S rRNA sequence

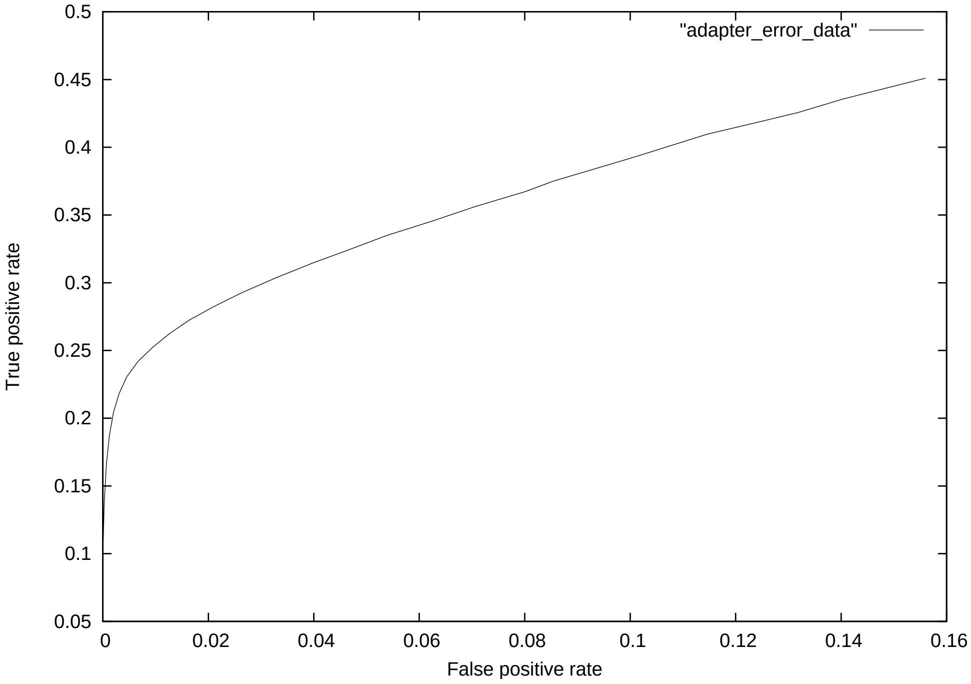


Figure 2.9: ROC curve plot of true and false positive rates for sequences aligned against the 3′ adapter sequence

Quality	Proportion of good bases lost
-5	1.9e-08
-4	2.3e-06
-3	1.8e-05
-2	7.4e-05
-1	0.00020
0	0.00066
1	0.0015
2	0.0025
3	0.0040
4	0.0060
5	0.0086
6	0.012
7	0.017
8	0.022
9	0.029
10	0.036
11	0.044
12	0.054
13	0.064
14	0.074
15	0.085
16	0.098
17	0.11
18	0.12
19	0.13
20	0.15
21	0.16
22	0.18
23	0.19
24	0.20
25	0.21
27	0.23
28	0.23
29	0.23
30	1

Table 2.4: The proportion of good bases lost at different quality threshold values

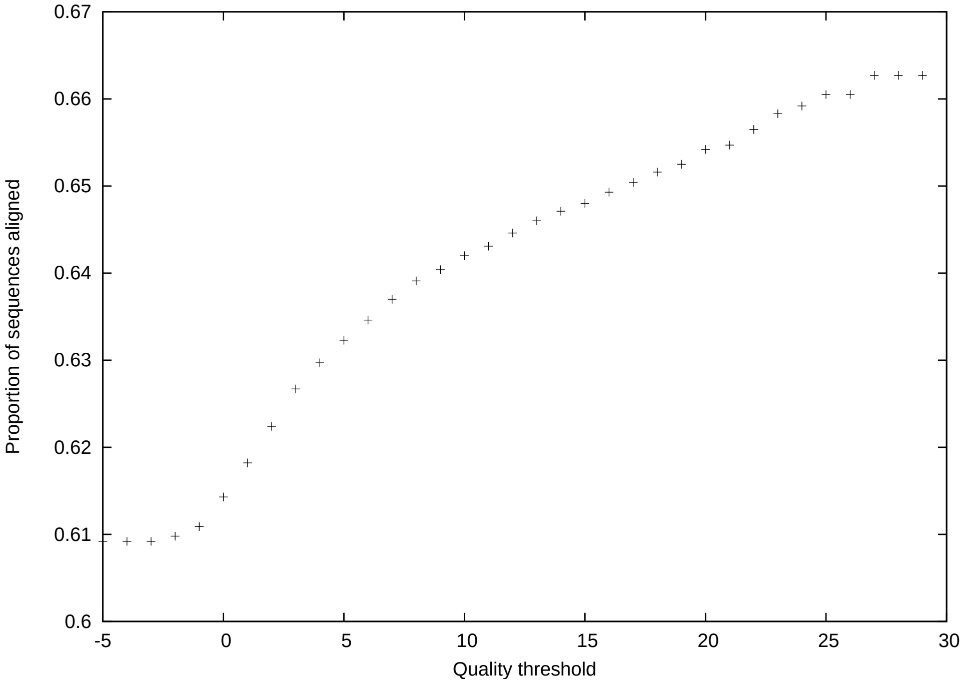


Figure 2.10: Total proportion of sequences successfully aligned against the reference genome without mismatches after base removal at the indicated threshold

duplication and reuse of similar domains in multiple proteins will mean that the primary sequence of functional genes may well be similar to that of other functional genes or pseudogenes. As a result, even small losses in information content may cause these sequences to align to multiple locations.

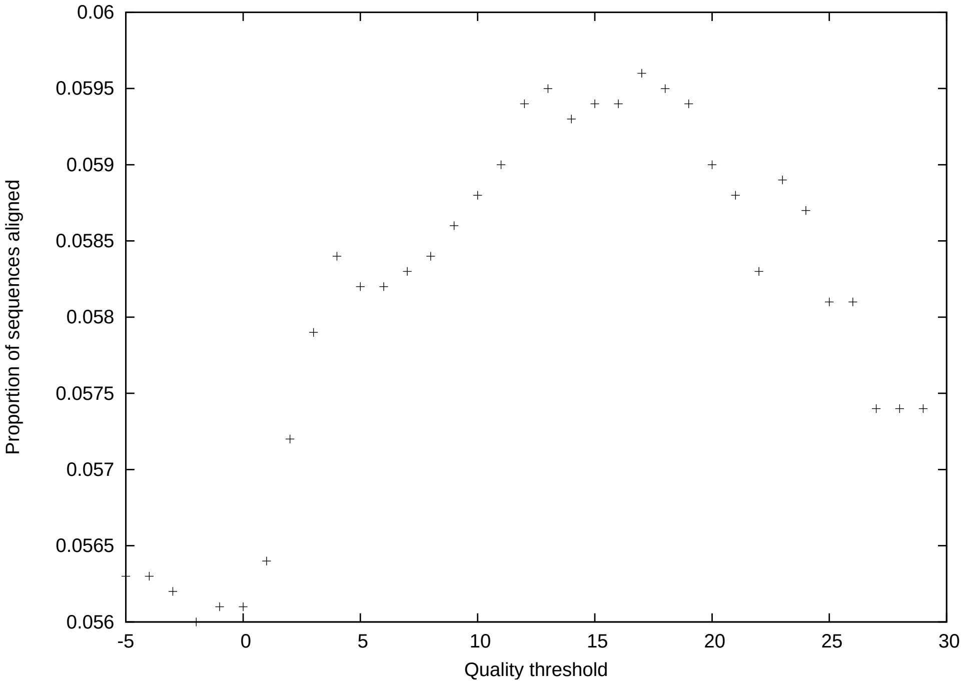


Figure 2.11: Total proportion of sequences successfully aligned against a unique location on the reference genome without mismatches after base removal at the indicated threshold

2.4 Conclusion

The average quality of the Solexa sequencing method is high, with the majority of called bases having an error rate of around 0.3%. 1/3 of this appears to be attributable to the preparation technique, as an error rate of only 0.2% is seen in the synthetic DNA used as the adapter sequence. This error rate is sufficiently low that it permits the identification of a previously undescribed variant of the 2S rRNA gene that is expressed at a low level in *D. melanogaster*, as well as providing evidence for a novel mutation in at least one 2S rRNA in the strain of *D. melanogaster* used for the preparation of the sequences. This high sensitivity demonstrates the power of high-throughput sequencing in examining gene expression.

By examining the quality levels, it is possible to draw conclusions about the appropriate level of sequence to discard. Setting the quality threshold to 17 appears to be optimal for ensuring that as many sequences as possible are uniquely aligned to the reference genome. Setting the quality threshold to 29 allows an extra 1.2% of sequences to be aligned, but in the process results in 3.7% fewer sequences being uniquely aligned. Determining the precise point to set this threshold may depend on the nature of the analysis being carried out. A lower threshold (ie, keeping more data) would discard more sequences that might otherwise be aligned. Increasing the quality score at which bases are discarded increases the probability that the sequences will be aligned, at the cost of a certain loss of specificity. In effect, this would lead to a decrease in false negatives (that is, areas that are expressed but whose reads are discarded) at the cost of an increase in false positives (that is, areas that are not expressed but which now have reads aligned against them). The tuning of this parameter is therefore influenced by experimental design.

The divergence in the apparent sequence of some copies of the 2S rRNA demonstrates the ability for comparative genomics to pinpoint sequence divergence, but also the rate at which sequences can change in even closely related species. *D. simulans* carries 25 copies of the 2S rRNA (compared to 16 in the closely related *D. melanogaster*), but none show the point mutation present in 4 of the 16 copies carried by *D. melanogaster*. Functional areas which demonstrate greater than expected divergence may indicate the presence of increased selection pressure, providing important evidence in determining what causes the functional differences between related species.

Chapter 3

Sequencing depth profiles as a tool for locating genes

The raw data used for this analysis was provided by Karen Ma

3.1 Introduction

The Solexa sequencing method allows a broad picture of expression levels of short RNAs to be built up. However, it is not then straightforward to assign predicted roles to each of these.

The sequences aligned against the reference genome may be thought of as a frequency graph. A single read aligned against a given position would represent a small bump against the background, while several thousand reads would represent a significant peak. The shape of these peaks then provides information about the transcript – the degradation products of an exon might be expected to be represented by a fairly flat peak stretching the entire length of the exon, while a tRNA would be a short peak which may or may not contain a gap indicating the presence of an intron. Similar methods have been used in the past for the discovery of other genetic elements (Lu et al., 2005) – here it is applied to tRNA discovery.

tRNA

Transfer RNA (tRNA) genes typically account for a relatively large number of genes in eukaryotic genomes. Release 5.3 of the *D. melanogaster* genome as retrieved from Flybase (FlyBase Consortium., 2003) contains 314 annotated tRNAs. tRNAs all fulfill the same purpose – that is, they are responsible for transferring amino acids to the ribosome and recognising the appropriate codon sequence in order to allow translation of an mRNA. This constrains them to sharing a similar secondary sequence, a property that has been taken advantage of by numerous applications designed to predict tRNA genes by genomic sequence alone.

The first of these prediction applications was developed by Roger Staden in 1980 (Staden, 1980). The abstract included the claim that "This program obviates the need to map the tRNA genes", which with hindsight was somewhat at odds with the rash of similar applications that were produced over the following decades. tRNAscan-SE (Lowe and Eddy, 1997) and Aragorn (Laslett and Canback, 2004) are now the most commonly used applications, with tRNAscan offering an extremely favourable combination of sensitivity and selectivity. These tools have sufficiently high performance that the need for mapping of the tRNA genes has, effectively, been obviated.

tRNAscan SE has a claimed false positive rate of less than 0.00007 per megabase. If this is accurate, the expected number of false positives in the *D. melanogaster* genome would be fewer than 1. Aragorn performs approximately an order of magnitude worse than tRNAscan-SE in this respect, but would still not be expected to produce any false positives on a genome of this size.

In both cases, these figures are determined from examining the performance of the software on random sequence. Even with careful selection of sequence parameters, such as maintenance of di- and tri-nucleotide frequencies, a random sequence will not match the information content of an actual genome. In particular, degenerate tRNA genes may still bear a close resemblance to genuine tRNAs even if they are no longer expressed. Estimating the ability of an algorithm to distinguish between tRNA pseudogenes and genuine tRNA genes is difficult in the absence of good prior knowledge of which sections of genome are functional and which are not.

High-throughput sequencing may provide an insight into this dilemma. It is now practical to sequence the RNA content of a cell, including tRNAs. These sequences may then be aligned to the host genome and used to validate predictions made by applications such as tRNAscan-SE or Aragorn, as described within this chapter.

3.2 Methods

As discussed in the previous chapter, sequences of 27 bases and under were derived from RNA extracted from *D. melanogaster* embryos (Ma, unpublished results). Though the size selection would be expected to exclude tRNAs[1], inspection revealed that a large number of exons of multiple types were represented in the sequences. The most likely explanations for this are either that the sequences contained degradation products or that the extraction protocol fragmented a proportion of full-length RNAs into shorter segments.

These sequences were aligned against the reference *D. melanogaster* genome. The number of reads aligning to each location was recorded, along with whether the read was uniquely aligned to this location or could be mapped to multiple locations on the genome. When plotted as a graph, this alignment presents a view of the sequence with peaks at locations with mapped reads. The presence of a peak implies that the corresponding region of the genome is expressed, with the height of the peak ideally being related to the relative expression level of the region. In this case, due to the shortness of the reads, many genes were only represented by degradation products. As a result of this somewhat more stochastic process, for longer sequences the number of aligned reads may vary greatly over their length even given a constant expression level.

An application was written to examine the profile generated after sequence alignment, with the aim of generating sequences that bore some resemblance to tRNAs. This was achieved by scanning the length of each chromosome in both the forward and reverse strands, looking for runs of aligned sequence that fit the desired criteria – in this case, that the length of the putative tRNA be be-

[1] The shortest annotated tRNA in *D. melanogaster* is 60 nucleotides, with the longest being 185

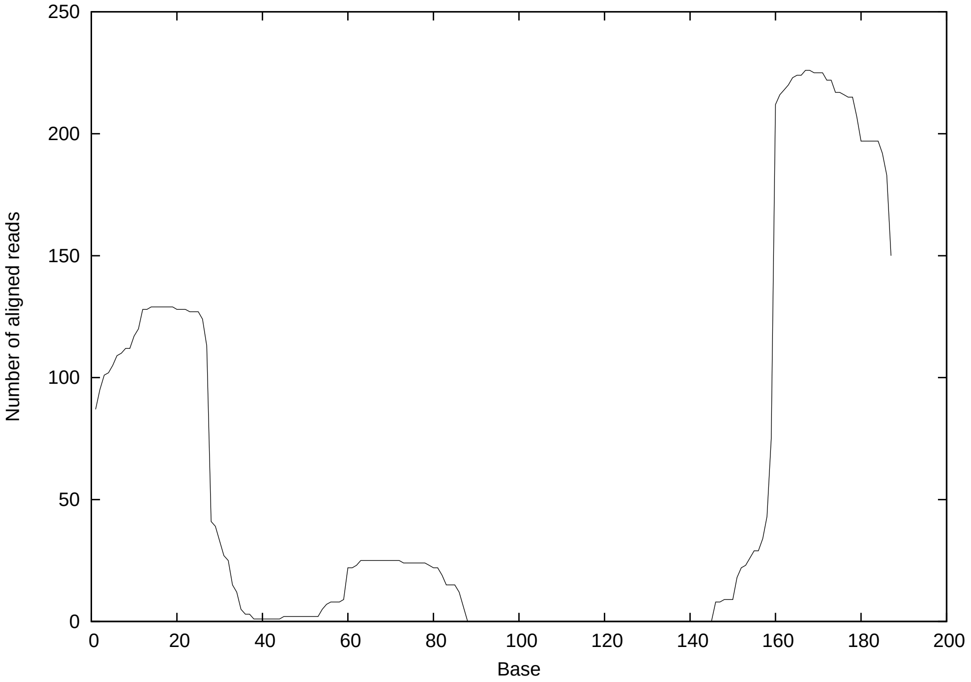


Figure 3.1: Graphed profile of a tRNA containing an intron between base 85 and 144, indicated by the absence of sequence coverage at this point

tween 40 and 200 bases, that it contain no more than one intron (as determined by the sequence count dropping to 0) and that the profile be supported by at least 5 individual sequences at some point during its length in order to reduce the probability that it was due only to spurious alignments. The first two criteria were simply based on examination of existing annotations of tRNAs. The final criterion was used in order to ensure confidence that the peak was not simply an artifact generated by a faulty sequence being misaligned. Figures 3.2, 3.3, 3.4 and 3.5 show a graphical representation of these constraints.

Once these peaks had been generated, a further step was undertaken in order to filter out a number of peaks that were considered more likely to belong to transposable elements than genuine tRNAs. This consisted of extracting the 100 nucleotides surrounding each peak and blasting them against the genome. For each putative tRNA, if the number of results with an e value of 0.0001 or

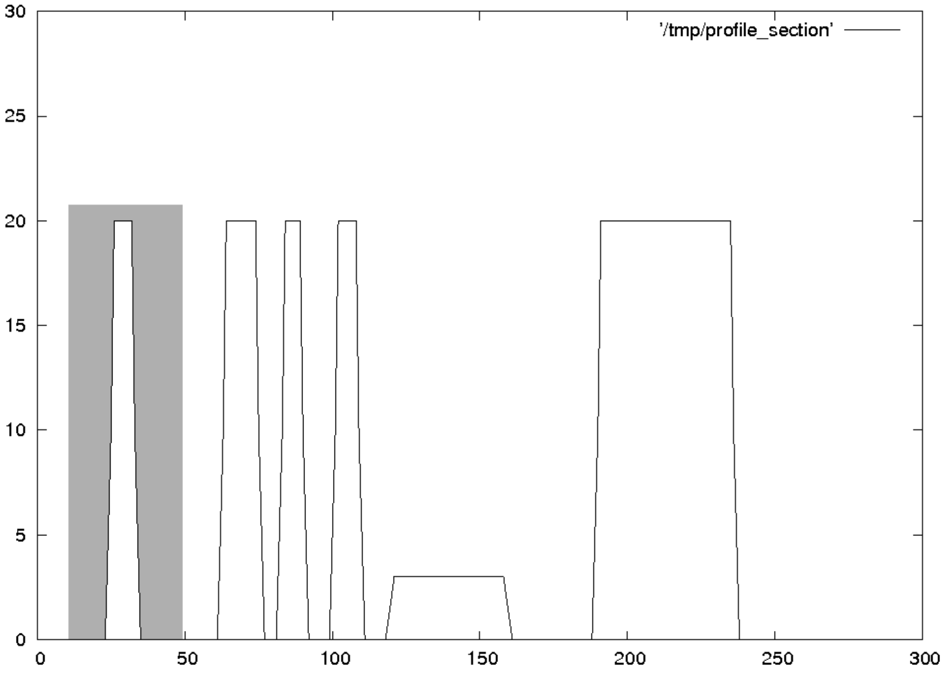


Figure 3.2: This peak is too short to be accepted

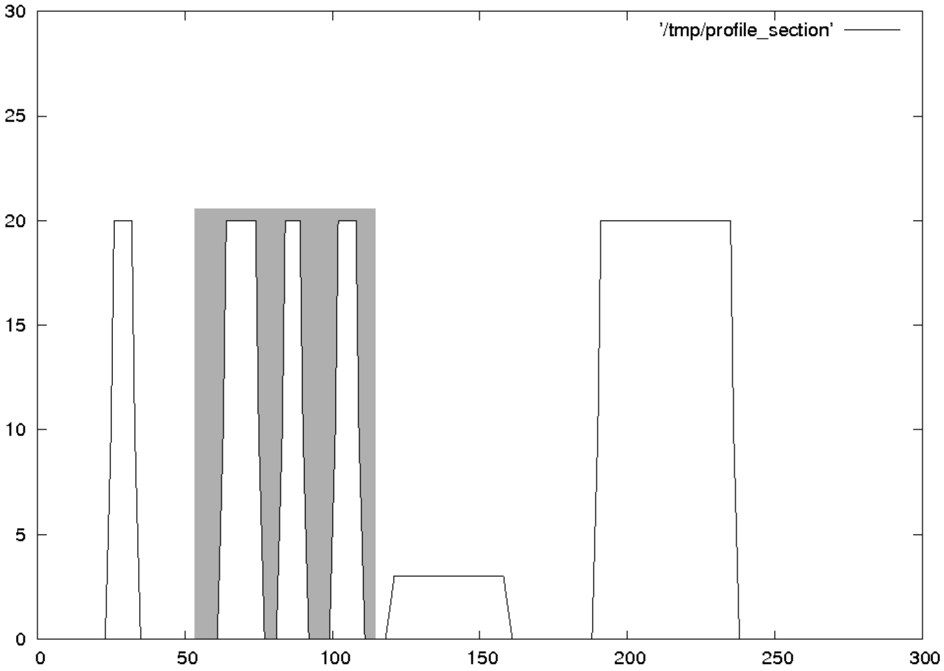


Figure 3.3: This peak contains too many introns to be accepted

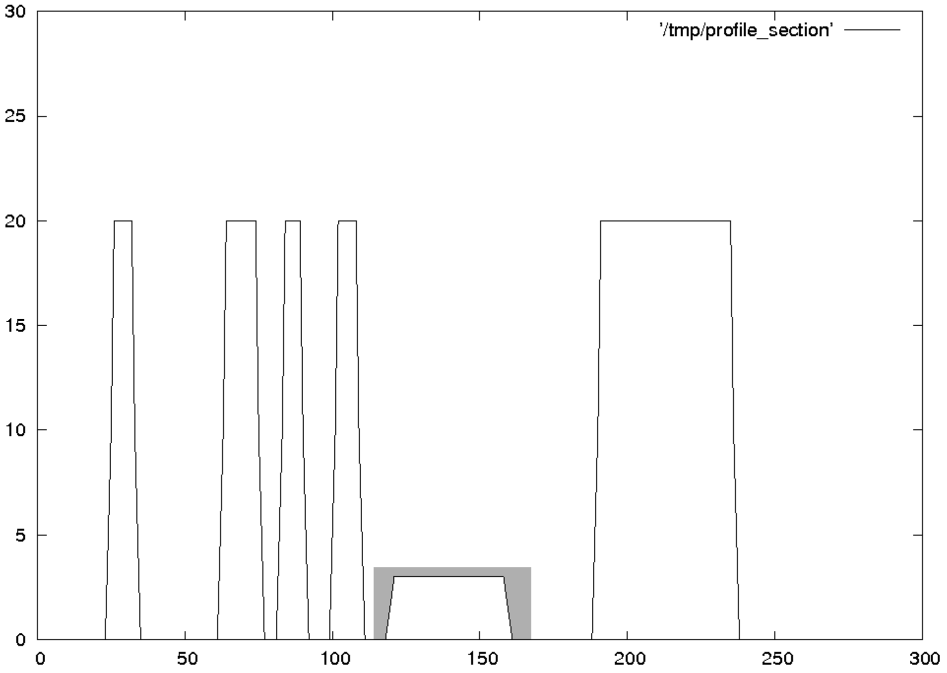


Figure 3.4: This peak is supported by too few sequences to be accepted

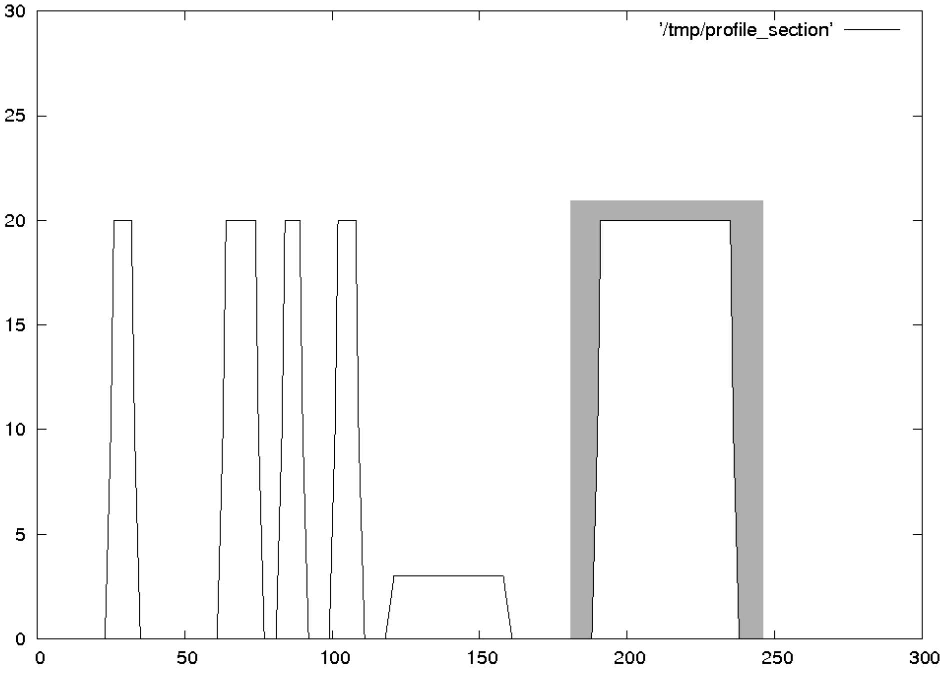


Figure 3.5: This peak is just right

lower was greater than 10, it was considered probable that the sequence was a false positive.

Those putative tRNAs that passed the above criteria were then tested by extracting the genomic sequence and running tRNAscan and Aragorn against them. Sequences which were supported by both the profile data and at least one of the prediction applications were considered as probable genuine tRNAs.

3.3 Results

14601 potential tRNAs were identified before the blast analysis stage. After apparently repetitive sequences were rejected, this number was decreased to 3925.

When checked against the 314 existing annotated tRNAs, 275 of them were positively identified. Closer examination revealed that 10 more had been positively identified and then discarded at the filtering step. Of these, three were located directly next to insertion elements, explaining the high frequency of hits. The others appeared to be cases where the peak was detected as shorter than in reality, with the result that the flanking sequence included sections of tRNA. As a consequence, the blast filtering picked up on other tRNAs and flagged the sequence as repetitive.

Sensitivity of the profile scanning approach

Reducing the stringency of the blast filtering increases the number of putative tRNAs to 4972, covering 280 of the 314 annotated transcripts. Of the remaining 34, 3 were discarded in the filtering step as described above and 22 are annotated as being present in the mitochondrial genome - an area not examined in this analysis. 9 were annotated as being present on the nuclear chromosomes but were not found. The small disparity between these figures and the number of putative tRNAs that matched annotated sequence appears to be down to the peak finding application misinterpreting large introns as a gap between two peaks, resulting in a single tRNA being misidentified as two putative tRNAs.

The 9 missing nuclear tRNAs were missed by the peak detection algorithm due to patchy coverage along their length, resulting in either no peak, a peak of inadequate length or a peak appearing to contain more than one intron. These rejections are therefore likely to be a consequence of the sequencing process not being optimised for sequences the length of tRNAs, rather than a failure of the algorithm itself.

Selectivity of the profile scanning approach

Of the 4972 sequences identified as potential tRNAs, 282 aligned with the 314 existing annotated tRNAs. Two of these corresponded to tRNAs that had already been identified – in these cases, the software had incorrectly identified peaks with large introns as two independent peaks. Therefore, the selectivity of the profile scanning approach on its own is a mere 5.6%. This corresponds to an unacceptably high false positive rate of 94.3%.

In conjunction with post-analysis with either tRNAscan-SE or Aragorn, the number of sequences identified as potential tRNAs drops to 268 (tRNAscan) or 265 (Aragorn). These figures correspond to a true positive rate of approximately 95%. The published figures for tRNAscan-SE suggest a true positive rate of 99.5%, while those for Aragorn suggest a rate of 98.2%. The disparity may be explained by the sequence coverage being sufficiently poor in places that peak boundaries were smaller than the genuine tRNA boundaries. As a consequence, the sequence passed to the prediction applications would be smaller than the genuine tRNA and the application would fail to identify it as a genuine tRNA.

Both tRNAscan-se and Aragorn identified two sequences in the putative tRNA dataset that had not previously been identified as tRNAs, shown in table 3.3. The profiles of these putative tRNAs include sequences that uniquely map

Chromosome	Start	End	Strand
2R	7292203	7292304	+
2R	7292609	7292916	-

Table 3.1: The chromosomal coordinates and strand of potential novel tRNAs identified by profile scanning technique and supported by tRNAscan-SE and Aragorn

to that location, strongly supporting the hypothesis that these are functional genes rather than pseudogenes with strong homology to functional tRNAs. The false positive rate is therefore between 0 and 0.7%, giving a selectivity of between 99.3% and 100%.

Sensitivity and selectivity of Aragorn

When run against the entire fly genome, Aragorn successfully identifies 290 of the 314 annotated tRNAs in *D. melanogaster*, giving a sensitivity of 92.4%. It also locates 5 unannotated sequences. Of these, two are supported by the profile analysis approach. The other three are unsupported by sequence evidence. This suggests a selectivity of between between 98.3% and 98.9%.

Comparative sequence analysis

In order to gain a better idea of whether the 2 predicted tRNAs were accurate, their sequences were aligned against the the other 11 sequenced Drosophila. Similar sequences were found in all other species, with the most relevant hits in all cases being around 1.5 kilobases upstream of the putative orthologue for the *D. melanogaster* gene CG7759. This matches the location of the sequences in *D. melanogaster*, demonstrating strong sequence conservation across a significant period of time.

3.4 Discussion

The combination of supporting profile data and computational tRNA prediction strongly supports the hypothesis that the sequences listed in table 3.3 are genuine new tRNAs. If this is accurate, then the selectivity of the dual approach would be on the order of 100% – that is, any results found by both the profile analysis and the computational predictions are highly likely to be true tRNAs. The sensitivity of the approach is primarily bounded by the sensitivity of the tRNA prediction applications, as any tRNAs not predicted by these applications will be rejected as spurious. In the event of the tRNA prediction applications not being able to predict all genuine tRNAs, the combined approach will also reject some potentially genuine results. This may be viewed as an argument for lowering the specificity of the prediction algorithms when used in this situation, as while this may lead to an increase in the number of tRNA-like pseudogenes picked up by the prediction algorithms, it is unlikely that these would also be supported by the profile analysis. The loss in selectivity of the prediction algorithms would therefore be compensated for.

The sensitivity of the profile analysis technique was also limited by the absence of good sequence coverage of some tRNAs. This limitation is likely to have been a result of the size selection involved in the initial RNA extraction protocol. A practical approach aimed at locating tRNAs would involve different size selection criteria, and would therefore be likely to provide better coverage of these areas. This would also enhance the sensitivity of the approach, reducing the number of genuine tRNAs that were missed. The relatively low coverage of some tRNAs in the profile analysis makes it impractical to conclude that tRNAs with no coverage are misannotated. A more in-depth study would make it possible to determine whether these tRNAs are genuine or not.

The high level of conservation across all available Drosophila sequences is strongly suggestive of functional conservation. The sequences of the two predicted tRNAs are highly similar[2], which in conjunction with their tight colocation in the upstream region of CG7759 suggests an ancestral duplication event. This arrangement could not be located in any other sequenced insect, suggesting that they developed some time after the *Drosophoilidae* and *Culicidae* split some 250 million years ago (Winter et al., 2007).

[2] But show adequate divergence to be certain that both are expressed

3.5 Conclusion

Sequence expression profile analysis provides strong evidence for determining whether a tRNA prediction is valid or not, though the sequence coverage of some tRNAs was sufficiently low in this case to make it impossible to conclude that any previously annotated tRNAs were spurious.

The combination of profile analysis and tRNA prediction validates the identification of two previously unannotated tRNAs, providing experimental support for predictions made by the existing tRNA prediction applications. Applying conservational analysis to these sequences supports this conclusion, and strongly suggests that these tRNAs developed within the past 250 million years.

It is therefore possible to conclude that the combination of profile analysis and computational prediction algorithms provides a powerful tool for the identification and annotation of functional elements.

Chapter 4

Conservational analysis of a dataset of putative microRNAs

The set of putative microRNAs used for this analysis was provided by Karen Ma

4.1 Introduction

In 1993, Lee et al. (1993) noted that the lin-4 gene in *C. Elegans* was required in order to control the temporal expression of the LIN-14 protein. Further examination revealed that lin-4 did not encode a protein, but instead generated a 22-base RNA complementary to the 3′ UTR of lin-14. It was suggested that the short RNA bound to the lin-14 transcript, forming a double-stranded RNA and inhibiting translation. Controlling the expression of lin-4 would therefore influence the translation of lin-14, providing the observed temporal control of LIN-14 production.

By 2001, almost 100 of these small RNAs had been discovered (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001) and the term microRNA introduced to describe them. The region surrounding these transcripts showed a tendency to fold into a stem-loop structure with the mature transcript located in one arm. Subsequent work showed that processing of this hairpin by the Dicer gene product, an RNAse III enzyme, would result in the mature transcript (Bernstein et al., 2001).

As more genomic sequence became available, computational prediction tools started to appear. MIRscan (Lim et al., 2003) and miRseeker (Lai et al., 2003) both adopted similar techniques based on locating areas of predicted hairpin formation and measuring conservation, with miRseeker suggesting around 100 microRNAs in the Drosophila genomes. While these applications showed high specificity, their reliance on conservation inherently reduced their ability to locate species-specific microRNAs.

High throughput sequencing allows an alternate approach to be taken. Rather than relying on conservation to reduce the false positive rate, it is possible to limit predictions to sequences that are expressed *in vivo*. Predictions based on these data will therefore be supported by biological evidence without limiting the method's sensitivity for species-specific microRNAs.

A dataset of putative microRNAs[1] was obtained using a method similar to that described in Chapter 3, using criteria derived from Ambros et al. (2003) with additional criteria to improve sensitivity and specificity(Ma, unpublished work). Their conservation amongst Drosophila genomes was then tested in order to gain an understanding of the distribution of the predictions.

[1] Kindly provided by Karen Ma

4.2 Methods

The coordinates of putative microRNAs were used to extract the corresponding sequence from the *D. melanogaster* genome, along with 100 bases of flanking sequence on each side. These sequences were aligned against the sequences of the other 11 genomes. Any significant hits were examined to ensure that the seed region of the microRNA was conserved, as this is believed to be the region of the microRNA that provides specificity in target binding (Lewis et al., 2003). The species furthest from *D. melanogaster*, while still retaining conservation, was recorded. This was repeated with the set of previously known microRNAs.

As a control set, 1000 randomly chosen sequences were extracted from each of the intronic and intergenic regions of the *D. melanogaster* genome. These sets were treated separately in order to ensure that the analysis was not biased by any different conservation pressures in intronic and intergenic regions. Each of these consisted of 22 bases in order to approximate the length of the putative microRNAs. These sequences were then aligned against the other 11 genomes with the same constraints as the putative microRNAs, as described above.

The results from the control set were used to generate a distribution via bootstrapping: for each iteration, 100 of the samples were chosen and the proportion falling into each category of conservation recorded. This was repeated 1000 times, providing a range of figures for each category. The mean and standard deviation of each category were then calculated and used to plot 95% confidence limits.

Some of the putative microRNAs were identified as overlapping microRNAs predicted by Ruby et al. (2007). A second dataset without these overlapping predictions was also generated and examined in the same way.

Table 4.1: Key to distance from *D. melanogaster*, as estimated by Tamura et al. (2004)

Millions of years	Species
5.4	*D. simulans, D. sechellia*
12.8	*D. yakuba, D. erecta*
44.2	*D. ananassae*
54.9	*D. pseudoobscura, D. persimilis*
62.2	*D. willistoni*
62.9	*D. mojavensis, D. virilis, D. grimshawi*

4.3 Results

Figures 4.1 and 4.2 show plots of the conservation of previously known microRNA dataset and putative microRNA dataset against the randomly selected sequences in intronic and intergenic regions, respectively. In both cases the previously known microRNAs are significantly better conserved than the random sequences across the entire range of species. In species as far away as *D. yakuba* and *D. erecta* the putative microRNAs are better conserved than the random sequences. However, beyond this the putative microRNAs are no better conserved than random sequence.

Performing a Chi-squared test upon the number of putative microRNAs conserved at each boundary (using the distribution of previously known microRNAs to generate the expected values) provides a probability of < 0.0001, strongly indicating that the distribution of the putative microRNAs is different to that of previously known microRNAs.

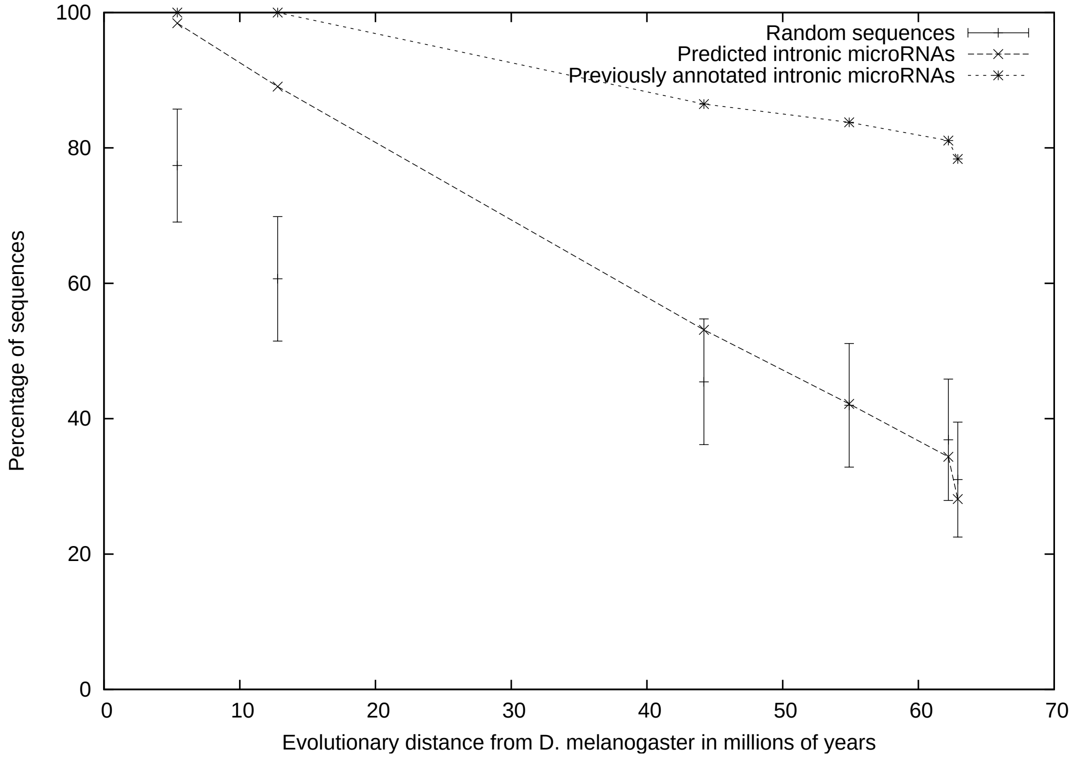


Figure 4.1: Conservation of intronic sequences. Evolutionary distance is millions of years of divergence from *D. melanogaster*, as estimated by Tamura et al. (2004) and shown in table 4.1. Error bars indicate 95% confidence limits on conservation of random sequences derived from *D. melanogaster*. Plotted against these are previously known MicroRNAs and the putative MicroRNA dataset

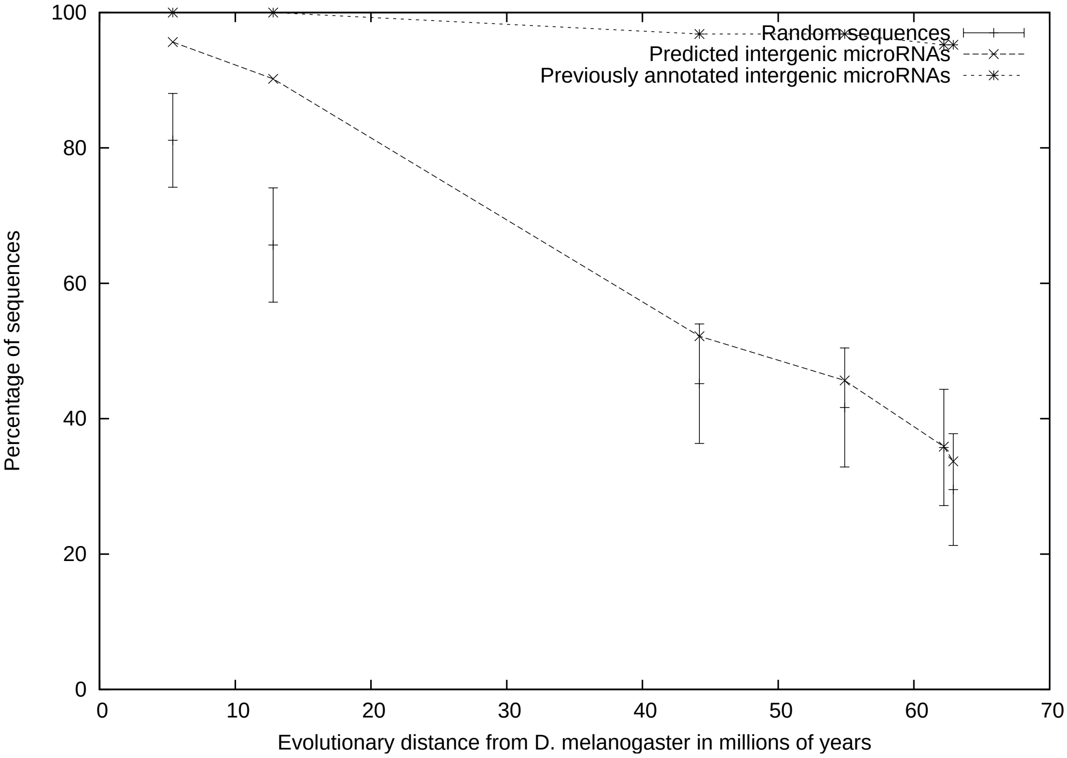


Figure 4.2: Conservation of intergenic sequences. Evolutionary distance is millions of years of divergence from *D. melanogaster*, as estimated by Tamura et al. (2004) and shown in table 4.1. Error bars indicate 95% confidence limits on conservation of random sequences derived from *D. melanogaster*. Plotted against these are previously known MicroRNAs and the putative MicroRNA dataset

4.4 Discussion

Of the previously annotated microRNAs, 89% are conserved across the entire range of sequenced Drosophilids with 7% being limited to the Melanogaster subgroup and the remainder being somewhere in between. However, recent studies (Berezikov et al., 2006; Lu et al., 2006; Ruby et al., 2006, 2007) have located evidence for less well conserved microRNAs. Ruby et al. (2007) describe a set of 58 novel microRNAs in *D. melanogaster* which is less well conserved than the previously known sequences (as shown in figure 4.5). Using the same method for conservational analysis as described previously, only 67% of these new predicted sequences are conserved over the entire range. 14% are limited to the Melanogaster subgroup.

Any attempt to use conservational analysis to determine the plausibility of predicted microRNAs being functional must therefore consider what the "true" distribution is. Ruby et al. (2007) suggest that the distribution of their identified microRNAs differs due to previous analyses using conservational studies to validate their predictions – that is, if a predicted microRNA could not be found in other species, it would tend to be regarded as spurious. They further note that many of these lineage-specific microRNAs tend to be expressed at lower levels, making it harder for them to be observed before the availability of ultra high-throughput sequencing techniques. The result of this would be that the existing annotations will tend to be enriched for sequences that are well conserved, rendering it inevitable that the conservation profile for previously annotated microRNAs will show high levels of conservation.

The studied set of putative microRNAs contained several sequences that were also present in the set identified by Ruby et al. (2007). Removing these

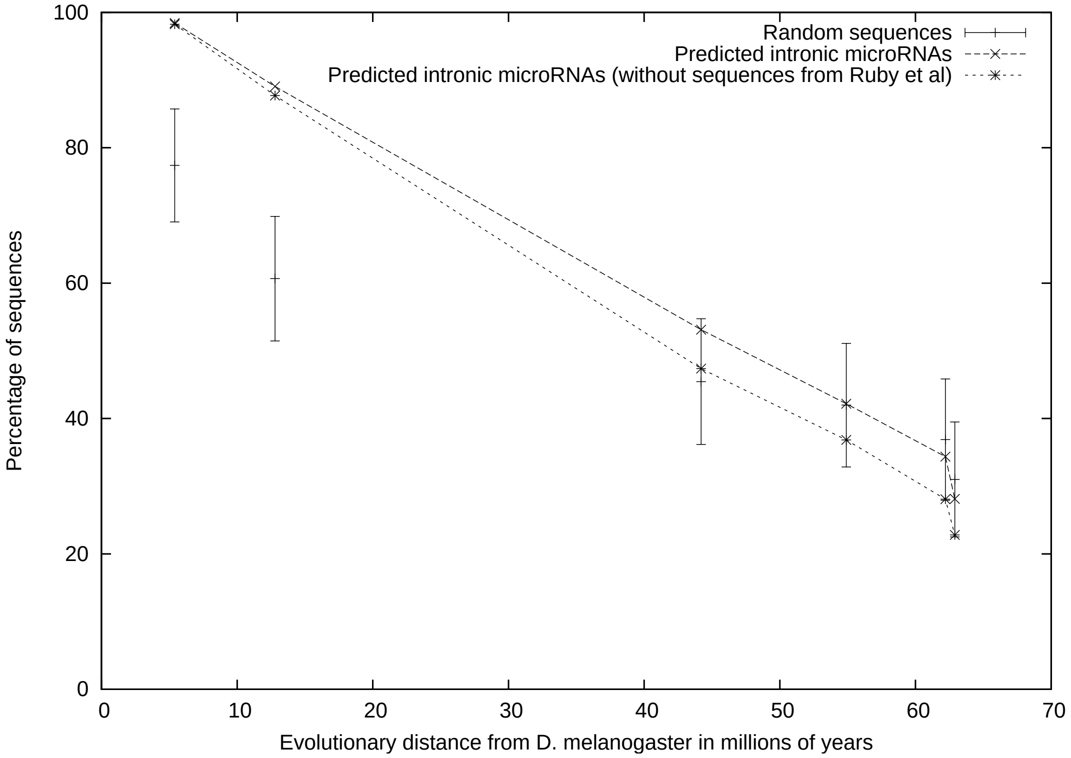


Figure 4.3: Conservation of intronic sequences. Evolutionary distance is millions of years of divergence from *D. melanogaster*, as estimated by Tamura et al. (2004) and shown in table 4.1. Error bars indicate 95% confidence limits on conservation of random sequences derived from *D. melanogaster*. Plotted against these are putative microRNAs with and without the overlapping data from Ruby et al. (2007).

from the dataset results in figures 4.3 and 4.4 for intronic and intergenic sequences, respectively. Removing these overlapping sequences results in little significant change to the results, though it is notable that the proportion of sequences conserved across all 12 species is reduced. This is accompanied by an increase in the proportion of sequences that are only conserved as far as *D. yakuba* and *D. erecta*.

How these results are to be interpreted depends on assumptions made about microRNA evolution. The vast majority of known microRNAs in *D. melanogaster* are well conserved across multiple species of Drosophila, and

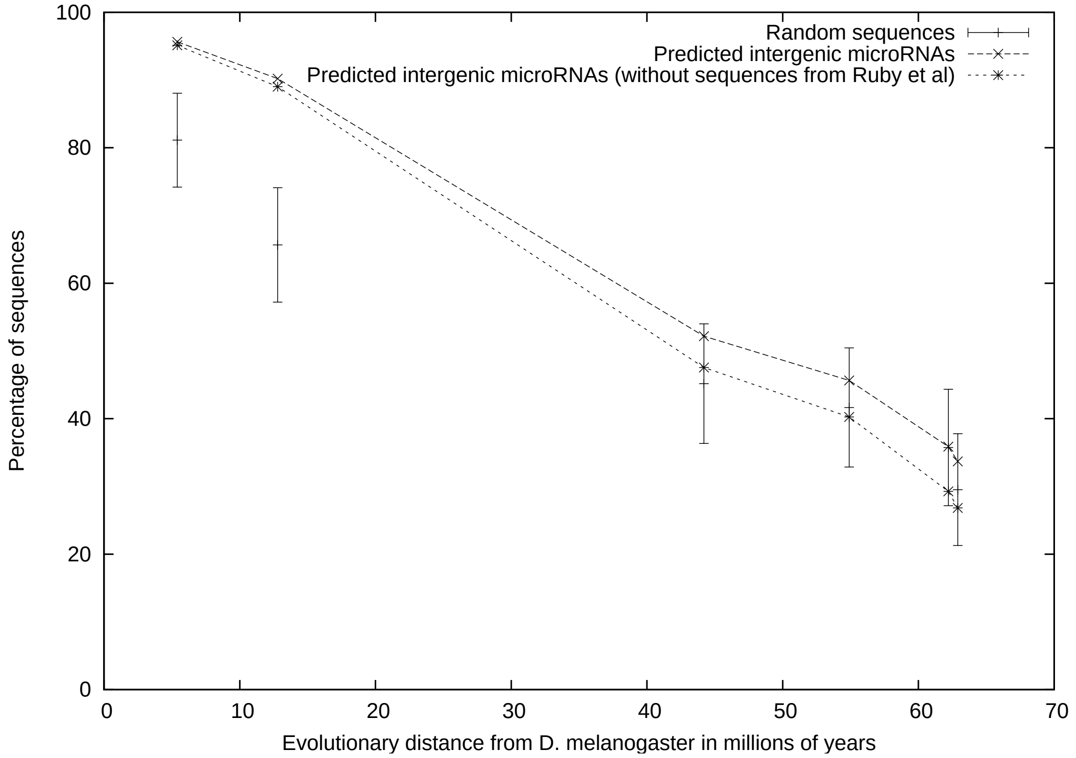


Figure 4.4: Conservation of intergenic sequences. Evolutionary distance is in millions of years of divergence from *D. melanogaster*, as estimated by Tamura et al. (2004) and shown in table 4.1. Error bars indicate 95% confidence limits on conservation of random sequences derived from *D. melanogaster*. Plotted against these are putative microRNAs with and without the overlapping data from Ruby et al. (2007).

there is therefore a temptation to conclude that any algorithm that generates predictions without these properties is poor.

However, it is also plausible that there is more evolutionary flux in microRNAs than has previously been believed. This hypothesis would imply that a the existence of a large number of poorly conserved putative microRNAs suggests that they evolved relatively recently in the organism's history. The results are not inconsistent with this hypothesis – conservation as far away as *D. yakuba* and *D. erecta* is high, after which the results become statistically indistinguishable from the randomly selected sequences. The significant difference between conservation of the putative microRNAs and conservation of randomly chosen sequences in these closely related species implies a certain degree of functional conservation.

In the species close to *D. melanogaster*, the difference between conservation of the random sequences and the putative set is on the order of 15-20%. This indicates that 15-20% of the putative sequences are better conserved than expected by chance alone, implying that at least that number represent some functional conservation – the true number may be much higher. This in turn implies the evolution of at least 20-30 microRNAs between the divergences of *D. ananassae* and *D. erecta*, or between approximately 40 and 12 million years ago (Tamura et al., 2004). This would imply an average rate of microRNA gain of approximately one per one to two million years, a figure that does not seem biologically implausible.

The implication that the putative microRNA set may contain sequences that evolved in relatively recent history is marginally supported by examining the level of conservation in more evolutionarily divergent species. With the overlapping data from Ruby et al. (2007) removed (figures 4.3 and 4.4), the results are verging on being *less* well conserved than randomly selected sequence in the more distantly related species. This is consistent with the evolution of new microRNAs in recent history – as sequences become functional, they will be subject to positive selection pressure and will evolve more quickly. This will manifest itself as a lower rate of conservation than random sequences under neutral selection pressure.

This leaves the question of why the level of conservation of the predicted microRNAs drops off where it does. The *D. pseudoobscura* genome was available significantly earlier than that of other non-Melanogaster drosophilids, and hence was used heavily in conservational analysis when attempting to predict and identify microRNAs. A consequence of this would be that existing sets of annotated microRNAs would be biased towards those conserved at least as far as *D. pseudoobscura*. If most well-conserved microRNAs have already been identified, then any studies that filter out existing annotated microRNAs would therefore appear deficient in microRNAs that are conserved in *D. pseudoobscura* and correspondingly enriched in microRNAs that are conserved only in more closely related species.

The time between the divergence of *D. pseudoobscura* and the divergence of *D. ananassae* is estimated to be around 10 million years, though the margin of error in this estimate is sufficiently large that the splits could in fact have been roughly contemporaneous (Tamura et al., 2004). If the time period between the two divergence events is small, then the probability of new microRNAs developing between those events is also small. This would be consistent with the most precipitous drop in conservation of the predictions being observed at the Melanogaster/Ananassae split rather than the Melanogaster/Pseudoobscura split. A similar but much less pronounced drop can be seen in the plot of the

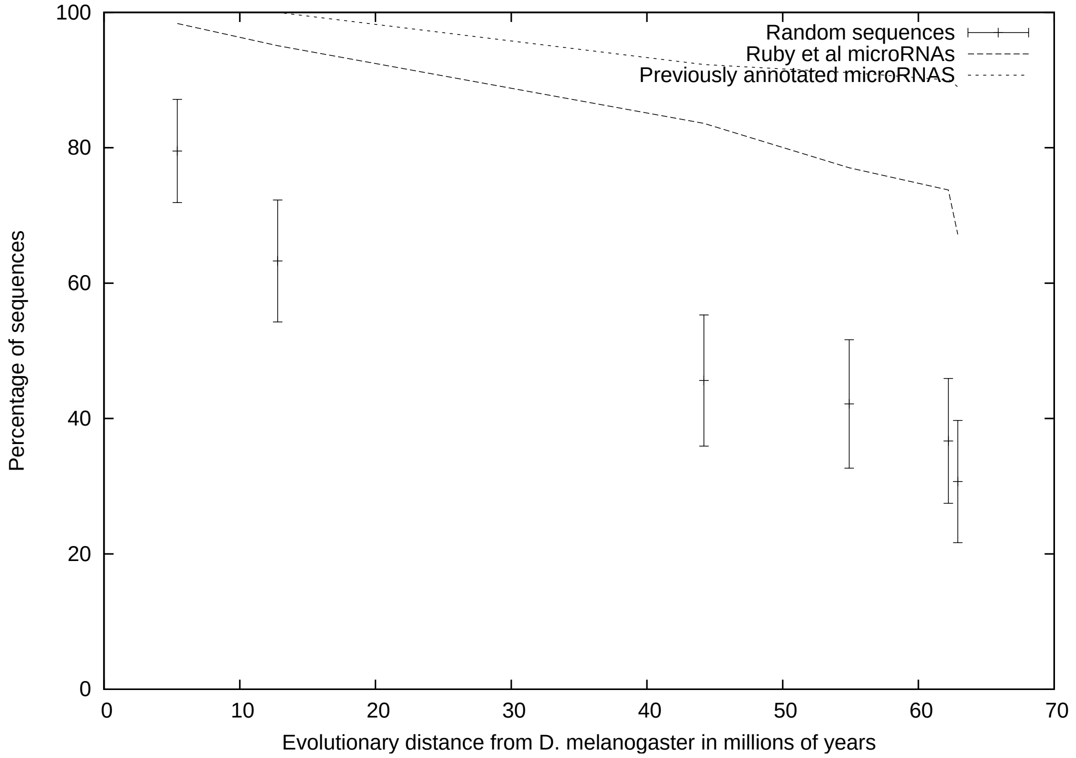


Figure 4.5: Conservation of sequences from predictions in Ruby et al. (2007), plotted against that of previously annotated microRNAs. Evolutionary distance is in terms of evolutionary forks away from *D. melanogaster*, as shown in table 4.1. Error bars indicate 95% confidence limits on conservation of random sequences derived from *D. melanogaster*.

data from Ruby et al. (2007), as shown in figure 4.5. The same considerations would apply to this set of data.

Much of this analysis is based on assumptions about the evolutionary history of the organisms considered. As these speciation events occurred in the distant past, we can only infer this history from phylogenetic analysis of the organisms' genomes. One of the assumptions made in phylogenetic analysis is that the rate of genomic change is broadly constant over time and across the different species concerned. If this is not true, it is possible that speciation events that look broadly simultaneous may in fact have occurred significant distances apart – or vice-versa.

This limitation does not pose a significant concern to this analysis. The aim was to identify whether the analysed sequences were conserved at a significantly different rate to random sequences in the same organisms. Unless there is any reason to think that this genomic change preferentially targeted sequences other than the putative microRNAs, the results remain valid. "Millions of years" in this case could instead be considered a shorthand for "Degree of genomic change".

Phylogenetic analysis also risks misjudging the order or structure of speciation events. This risk is greater where horizontal transfer may result in analysed genes passing between two otherwise distantly related species, resulted in them being considered closely related. For more distantly related species, it is also possible that convergent evolution may drive coding sequences back towards each other.

These issues are also unlikely to have affected this analysis. While phylogenetic trees based on small quantities of sequences stand a higher chance of misidentifying species relationships, Drosophila benefits from being well sequenced. Phylogenetic analysis can therefore take into account sections of genome, both coding and non-coding. Misidentification of species relationships would therefore require large quantities of genomic sequence from distantly related species to converge. Any argument that this is a likely outcome of genomic change stretches credulity. It is therefore reasonable to assume that the generally held history of Drosophila speciation is accurate, even if the timings of certain events hold some uncertainty.

4.5 Conclusion

The comparative genomics analysis allows us to say with a high level of confidence that the putative microRNAs are not from the same population as previously known microRNAs in *D. melanogaster*. The higher than expected level of conservation in species closely related to *D. melanogaster* is consistent with a set of more species specific microRNAs being identified, with conservation suggesting at least 20-30 of them as functional. This number is likely to be much higher, given the sequence evidence supporting each of them.

Chapter 5

Conservation analysis of gene nesting relationships

This chapter is a development following on from collaborative work published in Hudson et al. (2007)

5.1 Introduction

The hypothesis that conservation of genetic features will be higher when the features are functional is not limited to the primary sequences of genes. This chapter explores the use of conservational analysis as applied to a different type of feature – the relative arrangements of genes.

The first identified case of a eukaryotic gene being located within the intron of another gene was the discovery that the *Pcp* gene in *D. melanogaster* was contained within an intron of *ade3* (Henikoff et al., 1986). Further investigation noted that this arrangement (including non-coding elements) was conserved for over 50 million years, as far as *D. pseudoobscura* [1], suggesting a possible functional relationship or linking of expression. Hudson et al. (2007) described a conserved nesting relationship between *kay* and *fig* which has survived many independent chromosomal rearrangement events, strongly arguing for a functional relationship *fig* is hypothesised to be involved in *kay* regulation, strongly supporting the concept that they share some level of interdependent regulation, perhaps being driven divergently from a common promoter.

In humans, Yu et al. (2005) located 373 nested genes showing around 58% conservation to mouse, 28% to chicken and 15% to *Takifugu Rubripes*. 73% of the nested pairs examined via microarray analysis showed significant negative correlation in expression, suggesting that nesting may be used as a mechanism for avoiding coexpression of genes either by blocking the transcription apparatus, or by spliced introns from one gene forming dsRNA with the other.

The combination of high-quality genomic annotation for *D. melanogaster* and sequences for many closely related species has made it possible to con-

[1] though not in all other Diptera (Clark and Henikoff, 1992)

template whole-genome identification of nested genes and detailed analysis of the level of conservation. This chapter will attempt to determine the behaviour and functional significance of nested genes.

5.2 Methods

GFF files containing tabulated data of chromosomal features were obtained from http://www.flybase.org and used to determine the location of protein-coding genes in *D. melanogaster*. Genes which were entirely contained within intronic portions of other genes were identified, while non-nested and partially-nested genes were excluded. This resulted in a set of 1034 nested genes. The peptide sequences of each coding exon were identified and aligned against each of the other 11 sequences species of *Drosophila* using tlbastn (Altschul et al., 1997) in order to determine the extents of the orthologous genes in these remote species. The orthologous regions in each species were then examined to identify whether they retained the same nesting arrangement as in *D. melanogaster*.

GFF files were obtained from ftp://ftp.ensembl.org and used to determine the location of protein-coding genes in the human genome. Genes which were entirely contained within intronic portions of other genes were identified, while non-nested and partially-nested genes were excluded. Orthology information between human genes and *D. melanogaster* was obtained from http://inparanoid.sbc.su.se/ (Remm et al., 2001) and used to translate the set of nested human genes into the set of nested *D. melanogaster* orthologues. This list was then compared to the set of nested *D. melanogaster* genes.

5.3 Results

The tblastn approach was applied to *D. melanogaster* as a means of judging sensitivity. 605 nested genes were recovered, indicating a sensitivity of around 58.5%. When applied to *D. pseudoobscura*, 391 conserved nested genes were recovered. This compares favourably to an approach using existing gene annotations in *D. pseudoobscura* which identified only 215 conserved pairs (Hudson et al., 2007), indicating that current levels of gene annotation in non-*Melanogaster* organisms are probably lacking.

Of these 605 sets of nested genes in *D. melanogaster*, 433 unique container genes were identified. Most of these genes contained only a single nested gene, but 96 contained more than one. Flo-2 contained 14 genes, the most of any single gene. The number of pairs conserved in each of the other species is shown in figure 5.1.

The relatively poor sensitivity (58.5% in *D. melanogaster*) of the search tool was examined to determine the sources of the errors. 113 of the nesting pairs could not be unambiguously identified, resulting in a failure to identify the nesting. 114 were identified as being either on separate chromosomes, or spread widely on the same chromosome. 202 were identified as being located nearby on the same chromosome, suggesting that the extents of the genes had been misidentified.

Using these figures to estimate the false classification rate allowed a similar analysis to be performed on the other results, providing an estimate of the "true" number of nested pairs. This is shown in figure 5.2.

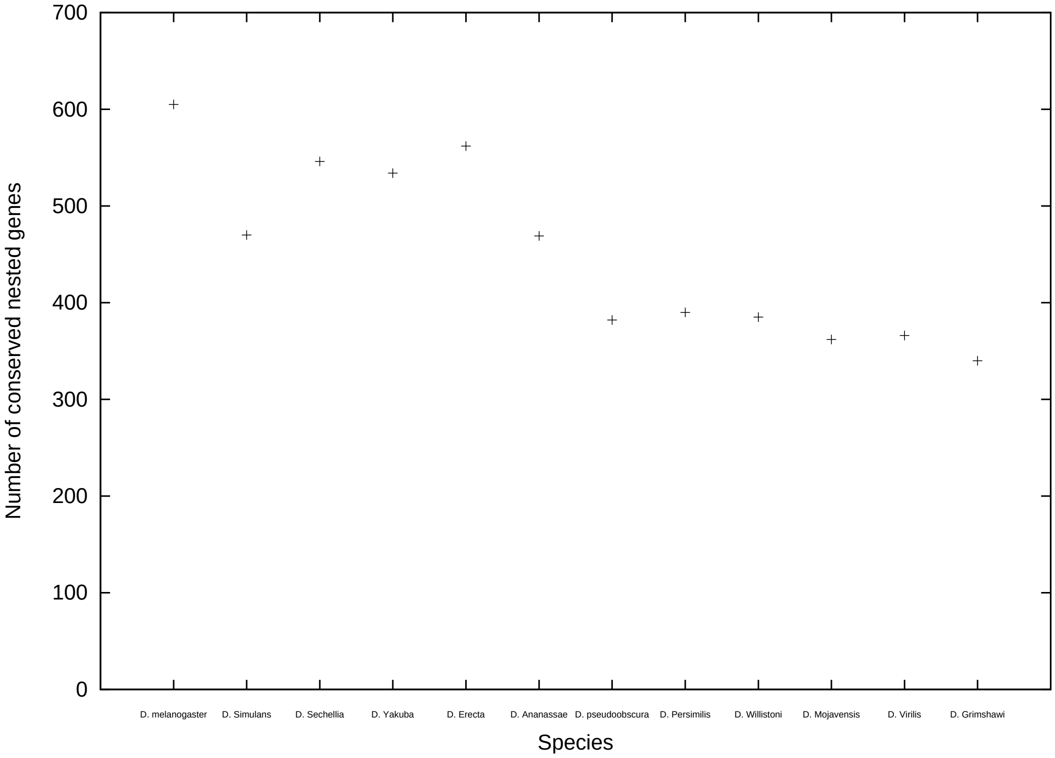


Figure 5.1: Conservation of nested genes. The number of gene nesting relationships in *D. melanogaster* that are conserved in each of the other sequenced species

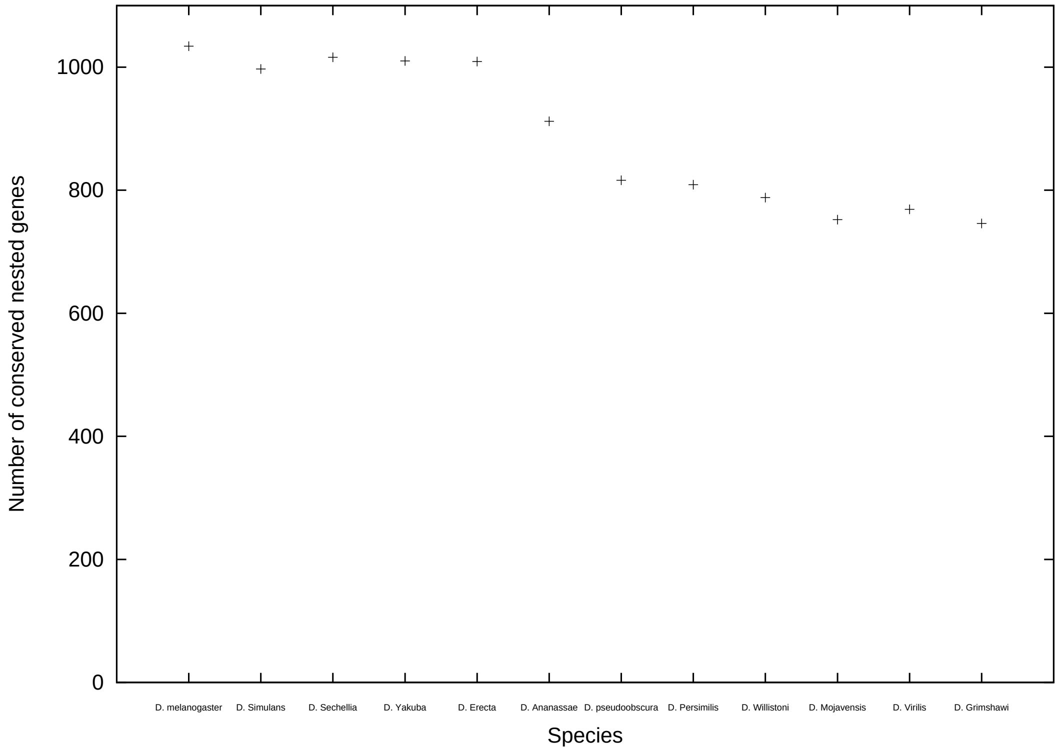


Figure 5.2: Conservation of nested genes. The estimated true number of gene nesting relationships in *D. melanogaster* that are conserved in each of the other sequenced species

5.4 Discussion

Modeling nested gene conservation

As shown in in figure 5.3 the number of conserved nesting arrangements tends to decrease with evolutionary distance. An estimate of the lower bound of the rate of loss of these nested genes can be made by examining *D. yakuba* and *D. erecta*. Using the data derived purely from the blast analysis[2], the number of pairs conserved between *D. melanogaster* and *D. yakuba* is 562, with the number conserved between *D. melanogaster* and *D. erecta* being 534. A more interesting figure is the total number of pairs carried within these two species, which is 574.

This implies that the common ancestor of *D. erecta* and *D. yakuba* carried at least 574 of the 605 pairs in the blast-identified subset of the pairs present in *D. melanogaster*[3]. *D. yakuba* and *D. erecta* diverged some 10.4 million years ago (Tamura et al., 2004). Assuming that all nested pairs have equal probability of being disturbed, this is consistent with an exponential decay with a half-life of between 99.7 and 341.2 million years.

D. melanogaster separated from the *D. yakuba* and *D. erecta* families around 12.8 million years ago, giving a gap of approximately 2.4 million years between this event and the speciation of *D. yakuba* and *D. erecta*. Working backwards from the figure of 574 nested genes in the common ancestor, we can calculate that the common ancestor of *D. melanogaster* and these species carried between

[2] This analysis has been carried out with the directly measured numbers rather than the estimates of the true numbers, due to the increased uncertainty inherent in the estimates and the inability to determine which pairs were conserved between a given species and *D. melanogaster*

[3] This ignores the case where both *D. yakuba* and *D. erecta* have independently lost the same nesting arrangement, but the effect of this is likely to have little significance on the estimated values

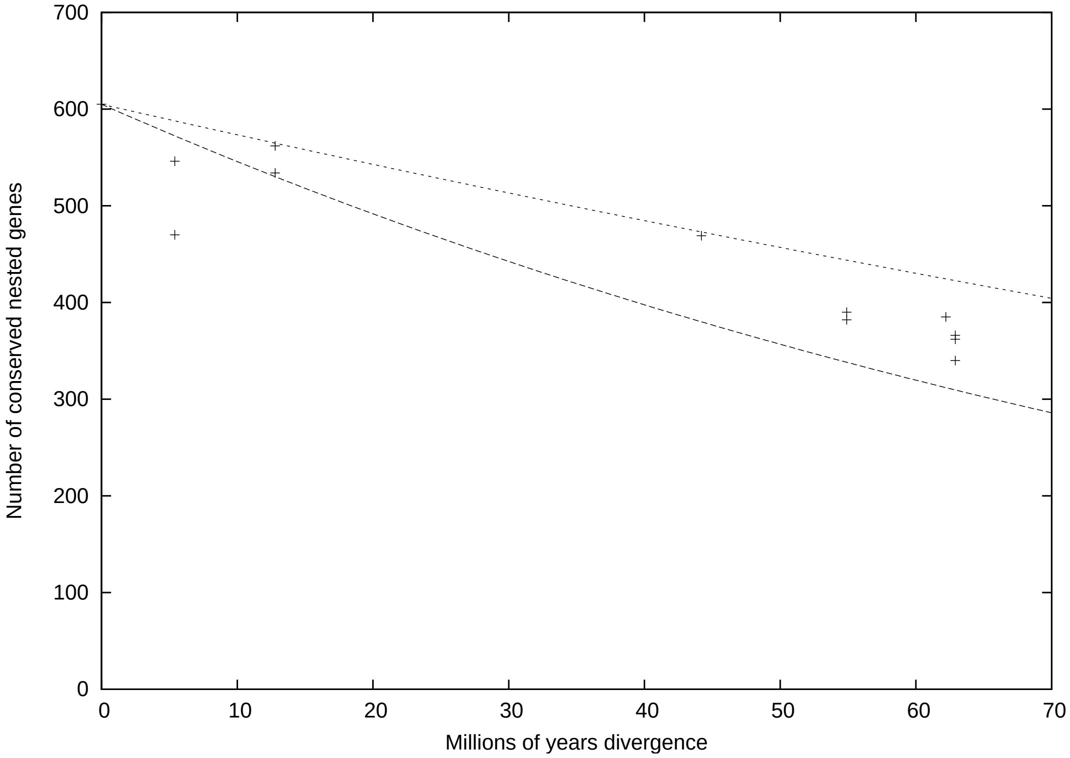


Figure 5.3: Conservation of nested genes along with upper and lower bounds of predicted conservation rate

577 and 584 nested genes that are present in *D. melanogaster*. This implies that 20-25 new pairs were gained by *D. melanogaster* over this 12.8 million year period, or a rate of gain of approximately 2 nested pairs per million years. This rate of gain is likely to be broadly independent of the existing number of nested genes, and so can be modeled as a direct linear increase. Performing the same analysis with *D. mojavensis*, *D. virilis* and *D. grimshawi* provides a comparable estimate.

Combining these two rates results in a formula of $n = (605 - 2x)e^{rx}$, where n is the number of nesting arrangements, x the number of millions of years of divergence and r the halflife of a nesting arrangement in millions of years. This is plotted in figure 5.3. The majority of results are within the range identified above, with the notable exceptions of *D. simulans* and *D. sechellia*. Both these genomes have been noted as being of poorer quality than the other sequenced genomes, with a larger number of artifacts (Hahn et al., 2007). Examining the data for *D. simulans* shows a disproportionately high number of nested genes that appear to have unambiguously altered their nesting status, either by moving to different chromosomes or being widely separated on the same chromosome. This indicates either a genuine acceleration in disruption of nesting arrangements or is an artifact of poor quality assembly.

Applying these figures to the estimated number of nested gene pairs results in the graph shown in figure 5.4. This suggests that the estimate of the rate of loss is excessive, with the half life perhaps being closer to the 340 million years figure. However, making a firmer estimate is difficult without a better idea as to which genes are actually conserved. This will require an improvement in the quality of genomic annotation.

Functional conservation of nested genes

Perhaps worthy of note is the sole pair identified as being conserved between *D. melanogaster* and human. The predicted model of nesting suggests that the number of nested genes in common between *D. melanogaster* and any other organism would decrease to 1 after approximately 180 to 520 million years of divergence, a figure not grossly inconsistent with the estimated divergence of 600 million years predicted to have passed since the branching of the Ecdysozoa and Deuterostomia. Nevertheless, the existence of a conserved pair is not a highly likely event.

The containing gene is *Brf*, a necessary component of the RNA polymerase III transcription machinery (Takada et al., 2000) and computationally identi-

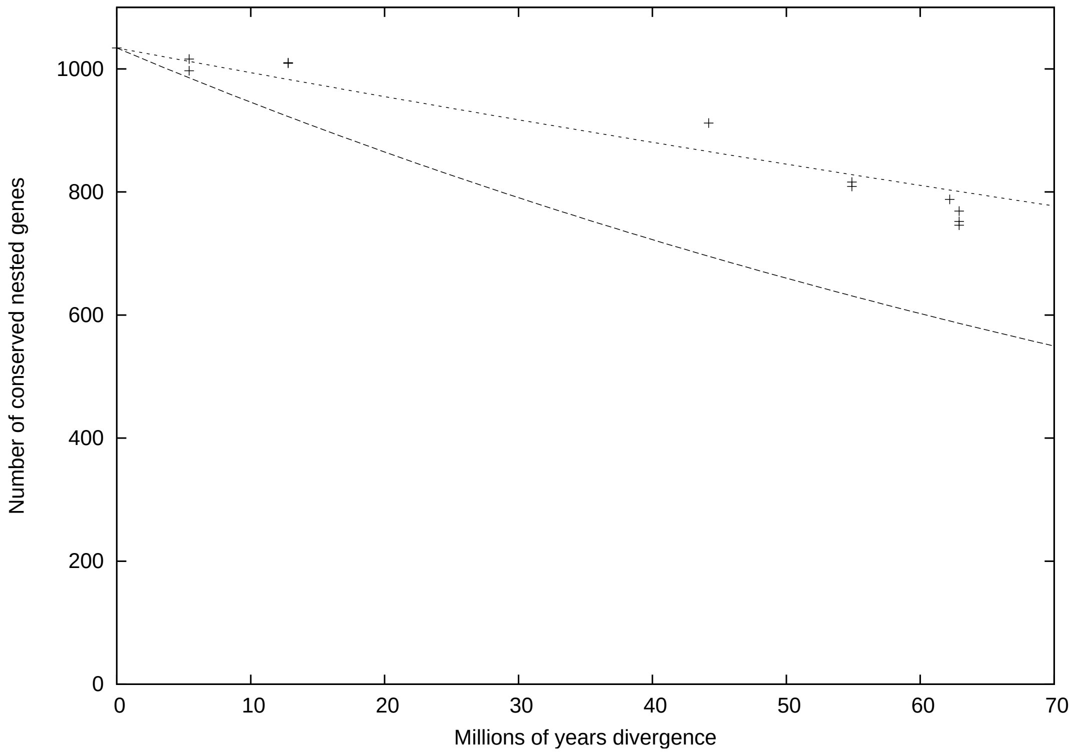


Figure 5.4: Estimated true number of nested genes in species, along with upper and lower bounds of predicted conservation rate

fied as being involved in RNA polymerase II activity (Thomas et al., 2003). The nested gene (*CG5319*) is also annotated as associating with RNA polymerase II promoters, though this is purely through sequence homology. *BTBD6*, the human orthologue of *CG5313*, has no functional annotation but two similar proteins in humans (*BTBD1* and *BTBD2*) have been experimentally identified as associating with topoisomerase I (Xu et al., 2002). The C terminus of *BTBD1* has been identified as sufficient for binding of topoisomerase I, and this region shows high levels of homology to *CG5319*.

The degree of sequence conservation suggests that *CG5319* (and, by extension, *BTBD6*) will also bind with topoisomerase I. Topoisomerase I is essential for gene expression in higher eukaryotes, though not in yeast (Lee et al., 1993), and has been shown to enhance TFIID-TFIIA complex assembly during transcription activation (Shykind et al., 1997). This argues for *CG5319*'s role in polymerase II activation.

According to GOToolBoxMartin et al. (2004), the probability of two genes associated with RNA polymerase II activity being randomly chosen is < 0.01, and in conjunction with the degree of conservation that this nesting relationship possesses provides a compelling argument in favour of this nesting relationship being functional. Expression data extracted from Genenote (Yanai et al., 2005) suggests that in humans, both genes are expressed in all tissue types with the *CG5319* orthologue being expressed at a lower level. However, this does not rule out the possibility that expression of one gene downregulates the other – examination of expression levels at different stages of the cell cycle would be required to determine that.

5.5 Conclusions

Conservational analysis has provided insight into the development of a model for the development and retention of nested gene pairs in Drosophila. In itself, this does not give any particular insight into whether nesting arrangements are functional. However, a high proportion of the total number of nesting arrangements in *D. melanogaster* are conserved over the entire set of sequenced drosophilids, indicating conservation over some 60 million years of evolution. While it is tempting to conclude that this slow rate of clearance is evidence for functional relationships, the ability to model the conservation rate as a random process suggests that this may not be the case. However, extending this analysis to humans provides a nesting relationship that has survived around 600 million years of divergence and shows suggestive evidence of a functional relationship. In combination with the high level of conservation of the *kay* and *fig* arrangement shown by Hudson et al. (2007), this provides a strong indication that some proportion of the nesting arrangements are under positive selection pressure.

Chapter 6

Locating functional RNA structural elements via folding and comparative genomics

6.1 Introduction

Comparative analysis of genes often concentrates on either the entire gene itself, or gross structure such as exons or protein domains. This ignores other functional elements contained within genes. This chapter explores one instance of this – RNA secondary structural elements used for transcript localisation.

In *E. coli*, a typical mRNA molecule may be responsible for the production of a protein molecule approximately every 3 seconds. In *D. melanogaster*, the average halflife of an mRNA molecule is somewhere in the region of 80 minutes (Ma and Huen, 2005). As a consequence, a single mRNA molecule could be capable of producing well over 1000 protein molecules during its lifetime. Localisation of the mRNA molecule therefore provides significant efficiency benefits over localisation of individual proteins.

However, efficiency is not the only argument for RNA localisation. In order for an egg to develop into an adult organism, it is necessary for it to have some asymmetry. This is commonly accomplished by localising RNA in specific areas of the egg. When the egg is fertilised and translation starts, the protein formed will also be localised and so, as the egg begins to divide into multiple cells, can affect gene expression and cell fate. The initial RNA localisation results in differing cell fates depending on where in the egg the cells are located, accomplishing the goal of setting up an initial axis.

Localisation's role in cell fate determination is not limited to axis formation. Asymmetric cell division is common during development, with each of the offspring developing into different types of cell. This can be achieved in a similar manner to axis formation. RNA is localised to one end of the cell so that it is segregated into a single daughter cell on division. Translation of this RNA again results in a protein that can influence cell fate, and its presence in one cell triggers a different development pathway to the other daughter.

It is fairly easy to understand how RNA localisation can be useful in development. Slightly more confusing is the role of RNA localisation in somatic cells. As an example, RNA localisation in the neurones has been implicated in learning (Miller et al., 2002). It is hypothesised that neuronal stimulation may lead to differing localisation of RNAs, with transcripts clustering around the dendrites. Translation of these transcripts may then alter the behaviour of synaptic junctions and allow long-term adaptation to the stimulus.

Examples of mRNA localisation

mRNA localisation was first observed in Ascidian eggs (Jeffery et al., 1983), where β actin mRNA is localised to the myoplasm. As understanding of the process of development and visualisation techniques improved, it became possible to find many more examples of localisation.

k10 is a protein required for *D. melanogaster* development. Flies carrying mutated copies are viable, but mutant females are sterile. This can be explained by the observation that k10 mRNA is laid down in the oocyte by the mother. Closer examination reveals that the k10 mRNA is localised to the anterior end of the egg, suggesting a role in polarisation of the embryo. (Cheung et al., 1992). Indeed, incorrect localisation of k10 mRNA results in a failure of gurken mRNA to localise to the dorsal side of the anterior of the egg (Schupbach, 1987; Haenlin et al., 1995). Correct localisation of gurken to the dorsal side of the egg is required for dorsal-ventral axis formation.

Interestingly, localisation of gurken is not limited to this dorsal-ventral role. Earlier in development, gurken mRNA is localised at the posterior of the embryo and translated gurken protein is required for induction of posterior cell fate (Gonzales-Reyes et al., 1995). Gurken highlights an important aspect of mRNA localisation – it is not a straightforward process. An idealised mechanism might resemble a postal system. Transcripts would carry a small sequence identifying the region they should be localised to, and the embryo would contain machinery for recognising those sequences and carrying the transcripts to the appropriate region. As the case of gurken shows, life is not that simple. The same transcript may be localised to different regions of the embryo at different times. Even localisation to a single location may be achieved through multiple stages, as can be seen in bicoid (McGregor, 2005).

There are many other maternally localised mRNAs in *D. melanogaster*. bicoid is localised to the anterior pole during oogenesis, and is responsible for anterior structure generation (Berleth et al., 1988). While bicoid mRNA is tightly localised to the anterior of the cell, this is less true of its protein. After translation, this slowly diffuses towards the posterior of the cell. The resulting gradient influences cell fate along the anterior-posterior axis and so plays a significant role in axis formation.

bicoid also demonstrates another important feature of mRNA localisation. While localisation will result in localised protein expression, not all localised protein expression is due to localisation of the corresponding mRNA. caudal mRNA is expressed constitutively, but caudal protein is not. This is because bicoid protein binds to caudal mRNA and inhibits its translation (Rivera-Pomar et al., 1996). Since bicoid protein is present in a gradient along the anterior-posterior axis, caudal is expressed in the opposite gradient.

Since bicoid is sufficient to result in the formation of the anterior-posterior axis, it may come as something of a surprise that several other mRNAs are localised to the posterior of the cell. The most obvious is oskar, whose localisation is necessary for formation of the pole plasm (the progenitor of the germ cells) and abdomen (Lehmann and Nusslein-Volhard, 1986). oskar protein acts indirectly: rather than encoding a transcription factor, oskar influences the localisation of other mRNAs and proteins (Breitwieser et al., 1996).

Perhaps the best characterised of these is nanos, the mRNA responsible for encoding the protein that does actually trigger the posterior fate (Wang and Lehmann, 1991). Translation of nanos is triggered after fertilisation, and the protein diffuses across the embryo in a similar manner to bicoid. Interestingly, in a similar way to bicoid's inhibition of caudal, nanos will inhibit transcription of bicoid (Wharton and Struhl, 1991). nanos thus strengthens the localisation of bicoid mRNA to the anterior of the cell.

Correct formation of the pole cells requires further mRNA localisation. germ cell-less protein induces some amount of pole cell formation (Jongens et al., 1992), but is not sufficient – mislocalisation to the anterior of the oocyte results in the production of polar buds, but not full pole cells (Jongens et al., 1994). Pgc is a non-coding transcript that is localised to the posterior of the cell, and is thought to be involved in the correct formation of the polar granules required for pole cell formation (Nakamura et al., 1996).

An interesting participant in multiple pathways, and a localised mRNA itself, is Orb. It codes for the oo18 RNA-binding protein which contains an element very similar to the terminal localisation sequence (TLS) in k10 (Serano and Cohen, 1995), which implies localisation to the anterior. However, orb is believed to be necessary for the translation of Oskar mRNA when it arrives at the posterior of the cell (Chang et al., 1999). orb is also responsible for enabling translation of K10 and gurken mRNA, playing a role in the formation of the dorso-ventral axis (Neuman-Silberberg and Schupbach, 1996). Orb appears to influence its own translation and localisation (Tan et al., 2001), but it is clear that it localises to different parts of the cell at different times.

How does localisation occur?

In order for localisation of mRNAs to occur, there must be some means by which the localisation machinery can recognise the transcripts. In several mRNAs, this has been traced to small regions of sequence. So far these have tended to be located in the 3′ UTR[1] and fold into distinctive secondary structure.

One of the best characterised localisation elements is that of K10. It has been identified as a 44 base stem-loop structure (Serano and Cohen, 1995) which is both necessary and sufficient for localisation. Mutations which disrupt the structure of the element have been found to impair localisation, indicating that the secondary structure is more important than the primary sequence. Closer examination has revealed that mutations that alter the shape of the minor groove impair localisation much more strongly than those that do not (Cohen et al., 2005).

In contrast to DNA (where the minor groove is small and inaccessible), the minor groove of double stranded RNA is the larger of the two grooves of the helix. It is therefore hypothesised that the proteins that bind to this structure in order to mediate the localisation identify it through interactions with the minor groove.

Orb has a similar localisation pattern to K10, and marks the only currently

[1] Though this is not always true, as can be seen in gurken

known occurrence of localisation element homology. The primary sequence of the Orb element is similar to that of K10, especially at the top of the helical section of the stem-loop structure. In contrast, the structure of the minor groove is almost identical – 14 out of 17 base pairs are the same in this respect.

Unfortunately, so far no other elements have been identified as having homology to the K10 TLS. Even so, Bicoid has a well characterised and rather more complex structure. A large region of the 3′ UTR is predicted to fold into a structure including 5 stem loops (Seeger and Kaufman, 1990; MacDonald, 1990). Different elements of this structure appear to be responsible for different stages of localisation. Initial localisation is disrupted if the structure of the distal section of stem-loop V is disrupted (Macdonald and Kerr, 1998). Its interaction with Staufen requires stem-loop III and the distal regions of stem-loop IV and V to be structurally intact, although primary sequence conservation is not required (Ferrandon et al., 1997). Stem-loop III is capable of base-pairing with other bicoid molecules, and this also appears to be required for Staufen recruitment (Ferrandon et al., 1997).

gurken's localisation is, like bicoid's, a multi-stage process. 5′ UTR signals are required for localisation in late oogenesis (Saunders and Cohen, 1999) and 3′ UTR sequence for complete dorso-anterior localisation (Thio et al., 2000). These have not been identified in detail, but a conserved stem-loop structure in the open reading frame has been determined as necessary for localisation (Bor et al., 2005). Interestingly, the I factor retrotransposon RNA includes a structure with similar secondary structure (though very little primary sequence conservation) and localises in a similar manner to gurken. This supports the idea that secondary structure is more important than primary sequence.

Computational prediction of RNA structure

In contrast to protein structure, RNA secondary structure can be predicted with a reasonable degree of accuracy. The most commonly used method is known as the Zuker algorithm (Zuker, 2000). This is based on a simple thermodynamic model. As single stranded RNA folds back on itself and forms helical sections, hydrogen bonds are formed between the paired bases. This process releases energy. The more bonds formed, the more energy released and the greater the stability of the structure.

The assumption is made that the structure formed is the one that forms the most bonds and is therefore the most stable. Given a list of energy values corresponding to each possible base pairing, finding the secondary structure is then "merely" a matter of finding the optimal set of base pairs.

The problem is effectively one of dynamic programming (Bellman, 1957). A grid can be built up with the sequence along both axes. Since the case being examined is that of a sequence folding back on itself, half the grid may immediately be masked off and ignored [2]. The grid is now filled in following the diagonals. For every possible base pair, the score is incremented appropriately. Since the value attached to a base pair is independent of the surrounding context, this can be done without keeping significant quantities of information about the state of the structure. As a result, the time taken by the algorithm is only proportional to the square of the sequence length.

At this point, finding the optimal structure is simply a matter of starting at the top right corner of the grid and tracking back through the scores. The result is guaranteed to be the optimal structure.

[2] This is because base 1 pairing with base 10 is equivalent to base 10 pairing with base 1. The latter case is ignored

The above is a slightly simplified version of the algorithm. In reality, it is also necessary to keep track of points where the structure may bifurcate, for example by one stem loop structure branching off another. For a sequence of N bases, there are effectively N-4 possible bifurcation points. These must each be examined and the global score recalculated, resulting in an algorithm that takes overall time proportional to the cube of the sequence length. To make matters even more complicated, modern implementations take into account the fact that the energy released by a given pair of bases bonding is influenced by its neighbours. Even so it is practical to simulate the folding of every *D. melanogaster* transcript on a modern desktop computer, the process taking around a week.

Existing algorithms are capable of predicting RNA secondary structure with an accuracy of between 50 and 70 percent when compared to experimentally determined structure (Eddy, 2004). The single biggest flaw is their inability to predict pseudoknots[3]. This is primarily for performance reasons – currently, accurate algorithms that take pseudoknots into account are unusably slow except for simple cases.

The other flaw is that the energy values used are not entirely accurate. These must be experimentally determined, and may vary under different experimental conditions. Attempts have been made to work around this, and algorithms which take the chemical modifiability of the RNA molecules into account exist (Mathews et al., 2004). However, this requires significant experimental work and is mostly suited to aiding experimental verification of predicted structures rather than enhancing the initial predictions.

On a genome-wide scale, the Zuker algorithm is currently the only viable

[3] Where bases in the loop of a stem-loop structure pair with bases outside that stem-loop

method for producing predicted secondary structures. As a consequence, it is the one that has been used in this work.

While most known functional mRNA secondary structures are under 50 bases long, it is not necessarily the case that each 50 base sequence will fold in the same way in isolation when compared to the context of surrounding mRNA sequence. In the context of the Zuker algorithm, possible pairing in one region may be given up in order to sequester some of the bases concerned into another, more energetically favourable structure. *In vivo*, it may be more reasonable to think of this as there being some degree of thermodynamic "churn", finally resulting in the sequence settling into the most thermodynamically favourable structure.

As a result, accuracy of structural prediction may be improved by simulating folding the entire mRNA. This has an unfortunate side-effect – as previously noted, simulated folding algorithms scale to order N^3, where N is the number of bases in the sequence to be folded. As a result, the time taken to fold the sequence increases significantly. This effect was mitigated by splitting up the workload and running multiple folding calculations in parallel.

6.2 Methods

In order to obtain a reasonable estimate of the number and location of mRNA structural elements over the entire genome, a simple software application was written. This read in each mRNA from a FASTA file containing the entirety of each spliced transcript in *D. melanogaster*. Each of these sequences was folded using the Zuker algorithm.

Once a structure had been obtained for each mRNA, it was cut into segments representing individual structural elements. The algorithm for producing these elements ran as follows:

- The 5′ end of the structure was examined. If it was unpaired, it was ignored and the next base chosen until a paired piece of structure was found.
- The first base that formed part of a structure formed, by definition, part of the "left hand side" of a structure. Any individual element should contain the same number of "left hand" and "right hand" bases in order to provide a consistent structure.
- Once the first structural base had been found, the next base was examined. If it was also a "left hand side" base, this step was repeated. For each "left hand side" base found, a counter was incremented.
- Once the first "right hand side" base was located, the counter was decremented. For each following "right hand side" base, the counter was further decremented. If another "left hand side" base was located before the counter reached zero, the 5′ base was incremented and the process restarted.

- If the counter reached zero, the sequence between the point where the counter was initiated and where the counter reached zero represented a minimal structural element. The start and end points were recorded, along with the structure inbetween. The 5′ start point was then moved to the end of the structural element and the process restarted.

This algorithm guaranteed that individual stem-loop structural elements would be obtained. In the case of a structural element consisting of a branched stem-loop structure, each individual stem-loop structure would be recorded separately. As a result, some of the large scale structure of the folded mRNA may have been lost.

This was considered to be an acceptable compromise. So far, each functional localisation element has been determined to consist of a stem-loop structure, occasionally with additional non-base paired bubbles protruding from the sides. More complex elements such as the Bicoid localisation element consist of much larger structures, but each independent functional element of the larger complex conforms to this criterion.

6.3 Results

Across the entire genome, this methodology produced 1396992 individual structural elements. The vast majority of these (approximately 90%) were under 50 base pairs long, with the average element length being 29 bases (standard deviation of 21) and the average stem length being 9 bases (standard deviation of 7)

6.4 Discussion

Identification of families of mRNA structural elements

As previously discussed, there are indications that functionality of mRNA localisation sequences depends on their structure rather than their primary sequence. Divergent sequences may therefore slowly evolve away from each other while maintaining secondary structure – alternatively, convergent evolution may drive structures towards similarity.

This leaves us with the problem of searching for structures that match any particular structural query. Traditionally, there have been two slightly different approaches used for this:

- A sequence and accompanying secondary structure are used to generate a profile, which may incorporate aspects of the primary sequence such as base composition. This profile is then used to search a database of sequences. An attempt is made to fit each sequence to the profile's structure, and (in a similar way to BLAST) scored on the degree of manipulation required in order to fit the sequence to the search profile. Examples of this are RSEARCH (Klein and Eddy, 2003) and ERPIN (Gautheret and Lambert, 2001). RSEARCH is optimised for querying single sequences against a database, whereas ERPIN is optimised for querying a multiple alignment and consensus structure against a database.

- Alternatively, a model can be constructed to describe the pattern of the structure. This model may contain information about the number of bases in the structure, bases that must be conserved and bases that must covary. Patsearch (Pesole et al., 2000) is typical of this.

In the case of functional localisation elements, there are very few pre existing families. The only case of two known elements sharing a common functional pathway is between K10 and ORB. In order to test the practicality of a consensus structure based approach, the K10 and ORB elements were aligned using CLUSTALW and then passed to RNAalifold (part of the Vienna package, an implementation of the Zucker algorithm (Hofacker, 2003)). This provided a consensus structure. This was then passed to ERPIN, which was instructed to ignore bubbles and the loop at the top of the stem - instead, alignment was to be based purely on how closely sequence could be fitted to the helical stem segment of the structure.

The results were not promising. Even at low cutoff thresholds, ERPIN was unable to identify any homologous sequence other than that found in the original queries. ERPIN has been successful in cases where a larger set of information is available [4], but does not seem appropriate in this case. The lack of already known families also makes it impractical to build the models that Patsearch requires.

As a result, further work was concentrated on RSEARCH. RSEARCH automatically creates a covariance model (a tree describing the base pairing, which then indicates which bases must vary together in order for the same structure to be formed) from the structure it is provided with, and then proceeds to find optimal alignments between the target sequences and this covariance model. Naively, RSEARCH generates a table with the query sequence on one axis and the target sequence on the other axis. In a BLAST search this table would then be filled with scores based on whether bases matched each other.

[4] For example, the iron response elements involved in iron metabolism – these have been identified in several genes and across multiple organisms, allowing much better determination of which information is functionally important

This is made more difficult with RNA structural alignment, as matching two bases (and thereby constraining them) may alter the ability of a distant base to align.

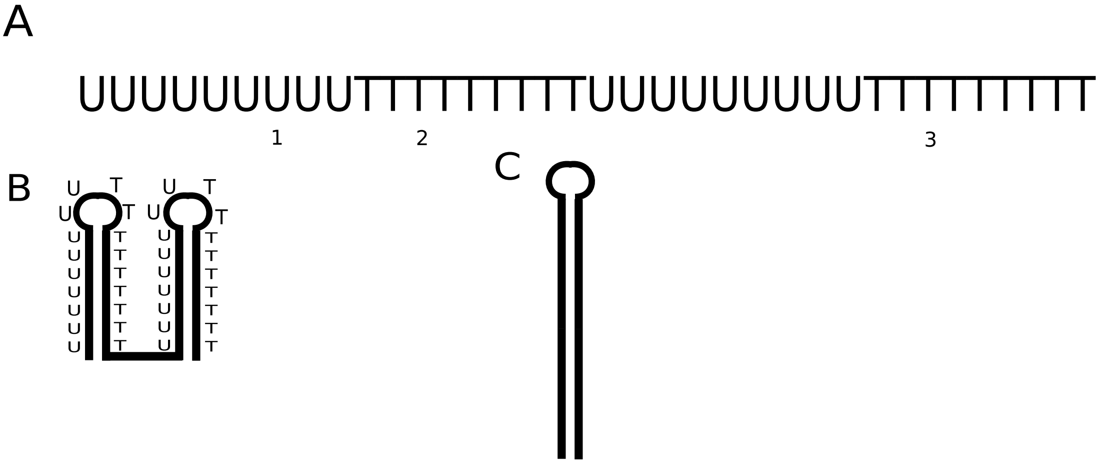


Figure 6.1: (A) shows a simple mRNA sequence. If base 1 pairs with base 2, the structure is constrained to that in (B). If base 1 pairs with base 3, the structure is constrained to (C)

Effectively, aligning any two bases may result in the entire table's scores being rewritten. As a consequence the algorithmic complexity is much greater and RSEARCH takes more processor time and more memory to perform a search than BLAST. In order to make this manageable, an assumption is made that any matches will not be significantly longer than the query. If a 50 base structure is queried against a 3000 base target sequence, RSEARCH will not attempt to form base pairs between the first and last base in the target. Instead, it will (by default) only look at bases up to 100 bases[5] away.

RSEARCH was used to attempt to match each of the structural elements located against every other element. Utilising a 16-way subset of a Linux-based Opteron cluster, runs typically took a week.

[5] Double the query length

The output was then processed in an attempt to identify families of homologous elements. Firstly, each high-scoring RSEARCH match was simplified into a pair of genes – the subject and the query. Each pair was then checked in order to ensure that the same match was obtained when the role of the sequences was reversed.

As a consequence of this, each gene could be considered to be connected to a number of other genes. A threshold was picked, and any gene with fewer connections than this threshold was removed. The removal of a gene may have reduced another gene's number of connections below the threshold – as a consequence, the process was iterated until the number of genes remaining stabilised. This tended to form small clusters of genes with high levels of interconnection and structural similarity.

This was performed for two significance levels of RSEARCH results. Examining RSEARCH results of 95% significance or more generated 78 groups, with a mean size of 4 and a standard deviation of 3.2. Looking at results with 99.9% significance or more generated 23 groups, with a mean size of 3 and a standard deviation of 1.8. In both cases, the median number of members was 2.

Each family was then queried via Flybase's bulk query interface and the set of molecular function GO annotations inspected by eye. The largest families were characterised by members containing highly repetitive DNA and having no functional relationship.

Two more interesting families were noted, each containing only two members. The first consisted of CG4819 and CG31054, both believed to be involved in the small nuclear ribonucleoprotein complex. These were both found to contain a 44 base stem-loop structure with a high degree of conservation between the two transcripts. However, closer examination revealed that the CG31054 transcript is duplicated in its entirety in the CG4849 5′ UTR. This suggests either a recent duplication event or an annotation error of some description.

The second family consisted of CG9455 and CG9456. In this case. closer examination revealed that the conserved structure is contained within a region of overlap between the two transcripts (Misra et al., 2002).

The issue appears to be that RSEARCH and other similar algorithms are concerned with finding near-precise matches for as much of the query structure as possible. This works well when trying to find precisely defined RNA structural elements, but less well when only small subsets of the features in the structure are necessary for functionality. Rather than work on modifying these complex algorithms, it was decided to focus on a simple, straightforward approach to exposing the information that was considered important in the structure.

A novel mechanism for identifying homologous structure

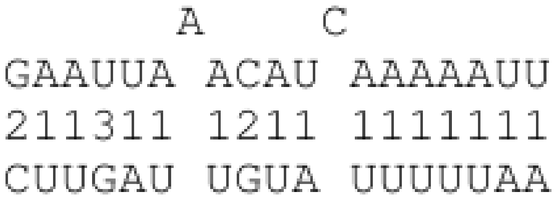
Cohen et al. (2005) showed that much of the functionality of the K10 localisation element is dependent on the shape of the minor groove of the helical section of the stem loop structure. To a first approximation, U:A and A:U basepairs are identical to each other in this respect, and the same applies to C:G and G:C basepairs.

Looking purely at the helical section of the structure, it is therefore possible to assign each base pair into one of three categories:

- U:A or A:U
- C:G or G:C

- non Crick-Watson pairing

In order to investigate this more closely, a method was developed to describe each RNA structural element in terms of the shape of its minor groove. Software was written to take each element and its accompanying structure. The sequences was then aligned to the structure and each base pair identified. Starting at the base of the stem, each base pair was then encoded as a number. A:U or U:A pairs were assigned a value of 1, C:G and G:C pairs a value of 2 and any other pairs a value of 3.



```
      A    C
GAAUUA ACAU AAAAAUU
211311 1211 1111111
CUUGAU UGUA UUUUUAA
```

Figure 6.2: The stem structure of the K10 TLS. Numbers represent the minor groove shape of the structure.

Cohen et al determined that the 3rd, 5th, 8th and 10th base pairs of the K10 helix impaired localisation if mutated in such a way as to alter the shape of the minor groove, while reducing the stem length below 14 base pairs also impaired functionality. This can be written as

....2.1..1.1..

where a . represents an unconstrained base pair, a 1 represents a base pair that must be A:U or U:A, and a 2 a base pair that must be C:G or G:C. This is semantically identical to a Unix regular expression, and as such a database of sequences tagged with these “structural alphabets” can be queried for matches using tools such as grep.

Out of around 140,000 RNA structural elements across the entirety of the *D. melanogaster* transcriptome, 20,681 match this query. This is a large proportion of the genome, and so this filtering is not obviously helpful in itself.

In order to improve the usefulness of this technique, the primary sequence of each element was BLASTed against the *D. pseudoobscura* genome. *D. pseudoobscura* was chosen for two reasons:

* It diverged from *D. melanogaster* 40-60 million years ago, and so has had ample opportunity for non-functional sequence to undergo significant change.

* *D. pseudoobscura* has the most completely sequenced genome of any of the non-Melanogaster Drosophilids.

It was assumed that any functional element should show structural conservation in *D. pseudoobscura*, along with primary sequence conservation. The top *D. pseudoobscura* hit for each *D. melanogaster* query was folded using the Zucker algorithm and then converted into the same structural alphabet as previously. These were then checked against the original query to ensure that the conserved primary sequence folded into a structure that matched the original search conditions.

This procedure reduced the number of matches to 56, corresponding to 43 unique genes [6]. Unsurprisingly, this included both K10 and ORB.

The marked genes were examined for GO annotations. Several were found to be interesting, and perhaps worthy of further investigation. As well as Orb and K10, two other genes involved in dorso-ventral patterning were found: Wingless and Dorsal. Five more genes were found to be involved in protein localisation, seven were involved in cytoskeletal organisation and eight in neurogenesis.

This information was collated with GOstat (Beissbarth and Speed, 2004), a tool which examines the annotation on a list of provided genes and provides

[6] In the case of multiple transcripts, each was searched independently and would therefore count towards the former score

information as to which annotation terms are over or under-represented. GO-stat showed several tags as significantly overrepresented[7], including dorso-ventral patterning, cytoskeletal organisation and neurogenesis (table 6.1).

[7] These statistics were generated without Orb and K10 being included in the list of query genes, as their inclusion was guaranteed in the first place. Including them results in an increased significance of several hits, as seen in table 6.2

6.5 Conclusions

The combination of structural analysis, conservational genomics and searches based on the minor groove shape of the K10 TLS have revealed a number of genes. Functional analysis (based on GO annotation) provides an indication that these a marginally significant overrepresentation of certain functional annotations. If the conserved structure is functional, the similarity to the minor groove shape of the K10 TLS could indicate an interaction with the same localisation pathway as K10 and ORB. Experimental examination of the localisation of these genes in oocytes would be required to demonstrate this.

Table 6.1: GOstat results, without K10 and ORB

GO Annotation	Genes	P
neurogenesis	KEK2 DL OTK TM1 NERFIN-1 WG CG31694 HIW	0.0306
morphogen activity	WG DL	0.0306
Notch binding	WG DL	0.0306
cytoskeleton organization and biogenesis	KHC-73 CG3121 CHER WG DL BTV ACT5C	0.0477
dorsal/ventral pattern formation, imaginal disc	WG DL	0.0477

Table 6.2: GOstat results

GO Annotation	Genes	P
dorsal/ventral pattern formation	WG DL ORB FS(1)K10	0.017
neurogenesis	KEK2 DL OTK TM1 NERFIN-1 WG CG31694 HIW	0.017
morphogen activity	WG DL	0.017
Notch binding	WG DL	0.017
oocyte axis determination	DL ORB TM1 FS(1)K10	0.017
axis specification	WG DL ORB TM1 FS(1)K10	0.017
female gamete generation	DL ORB TM1 FS(1)K10	0.017
dorsal/ventral pattern formation, imaginal disc	WG DL	0.0296
cytoskeleton organization and biogenesis	KHC-73 CG3121 CHER WG DL BTV ACT5C	0.0296
ovarian follicle cell development (sensu Insecta)	CHER WG DL FS(1)K10	0.0296
oogenesis (sensu Insecta)	CHER WG DL ORB TM1 FS(1)K10	0.0296
oogenesis	CHER WG DL ORB TM1 FS(1)K10	0.0296
dorsal/ventral axis specification	DL ORB FS(1)K10	0.0296
anterior/posterior axis specification	WG DL ORB TM1	0.0296
anterior/posterior pattern formation	WG DL ORB TM1	0.0335
female gamete generation	CHER WG DL ORB TM1 FS(1)K10	0.0335
intracellular mRNA localization	ORB TM1 FS(1)K10	0.0335
protein localization	KHC-73 CHER BTV ORB CG31158 FS(1)K10 CYP33	0.0335
pattern specification	WG DL ORB TM1 FS(1)K10	0.0335
gametogenesis	CHER WG DL ORB TM1 FS(1)K10 ACT5C	0.0335
cytoskeleton	KHC-73 CG3121 BTV TM1 ACT5C	0.0335
oocyte axis determination (sensu Insecta)	ORB TM1 FS(1)K10	0.0335
oocyte construction (sensu Insecta)	ORB TM1 FS(1)K10	0.0335
sexual reproduction	CHER WG DL ORB TM1 FS(1)K10 ACT5C	0.0335
reproduction	CHER WG DL ORB TM1 FS(1)K10 ACT5C	0.0335
structural constituent of cytoskeleton	KHC-73 CG3121 CHER BTV ACT5C	0.0335
ommatidial rotation	WG DL	0.0335
oocyte anterior/posterior axis determination	DL ORB TM1	0.035

Chapter 7

Experimental validation of predicted localisation elements

7.1 Introduction

The sheer number of identified structural elements in *D. melanogaster* (some 140,000) is sufficiently large that picking some arbitrary subset of them may produce apparently significant results. As a consequence, it was felt that experimental validation was required to determine whether the set of putatively localised genes described in chapter 6 actually demonstrated any subcellular localisation. Based on the hypothesis that the minor groove motif of the *k10* TLS is responsible for the correct localisation of *k10* and *Orb*, it would be expected that these transcripts would also be localised in a similar manner. Examining *D. melanogaster* oocytes with in situ staining should demonstrate any subcellular localisation, if present.

7.2 Methods

Of the 41 unique genes identified in chapter 6, 29 were identified as being expressed in ovarian tissue using data from FlyAtlas (Chintapalli et al., 2007) and are show in table 7.1. The largest exon in each of these genes was identified and primers designed using the Primer3 package (Rozen and Skaletsky, 2000) to produce products of between 500 and 1000 bases. For each gene, one primer was prefixed with the T3 promoter sequence and the other with the T7 promoter sequence.

Genomic DNA was extracted from *D. melanogaster* using the protocol described in appendix A. This material was then used as the basis for 29 PCRs. The PCR product consisted of exonic material prefixed with the T7 promoter sequence on one strand and the T3 promoter sequence on the other. Single-stranded RNA probes were then grown from each strand of the product, producing a set of 29 antisense probes and 29 sense control probes.

Ovaries were extracted from a number of female *D. melanogaster* from the Oregon R strain and separated into individual egg chambers. In-situ staining was then performed using the protocol described in appendix A and the stained egg chambers examined under a light microscope for evidence of localisation.

CG2519
CG3173
CG3218
CG3570
CG3937
CG4027
CG4886
CG4898
CG5452
CG6438
CG6667
CG6818
CG7865
CG8002
CG8183
CG8291
CG8967
CG10868
CG11107
CG11416
CG11614
CG11897
CG11986
CG13148
CG31133
CG31158
CG31694
CG32592
CG32791

Table 7.1: Genes identified through structural screening and positively identified as being expressed in ovarian tissue

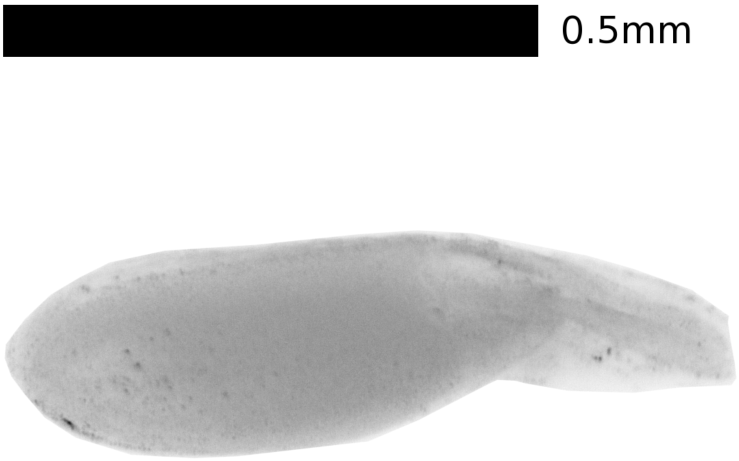


Figure 7.1: Typical negative control staining, in this case using a sense *CG32592* probe. Stage 10 oocyte – at this stage, *k10* is still localised to the anterior of the oocyte.

7.3 Results

Negative controls were performed using the sense versions of the generated probes. All showed no evidence of staining. Figure 7.1 shows a typical example. k10 was used as a positive control. Figure 7.2 shows anterior localisation of staining, consistent with the expected localisation pattern.

Results from other probes could be split into two categories – those which showed ubiquitous staining (eg figure 7.3)and those which showed no staining (eg figure 7.4). No localised staining was observed. These results are summarised in table 7.2.

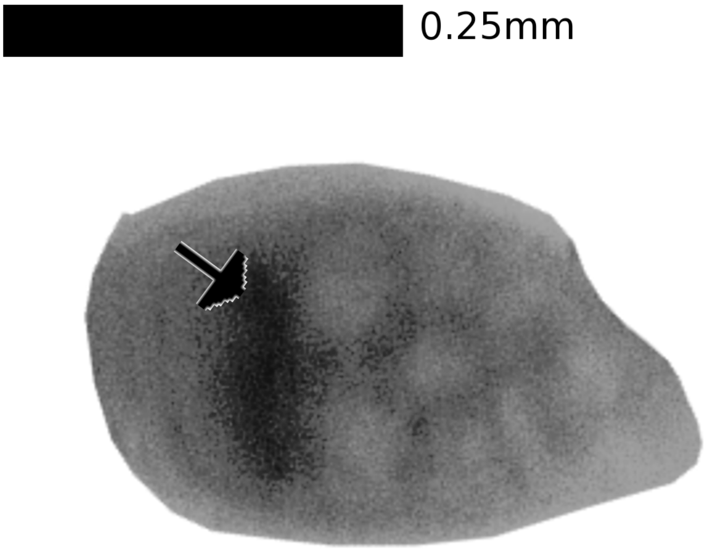


Figure 7.2: *k10* localisation. Stage 9 oocyte. Probe is DIG-labeled anti-*k10*, with NBT/BCIP staining. Arrow indicates anterior localisation of *k10* transcript. Image has been enhanced for clarity.

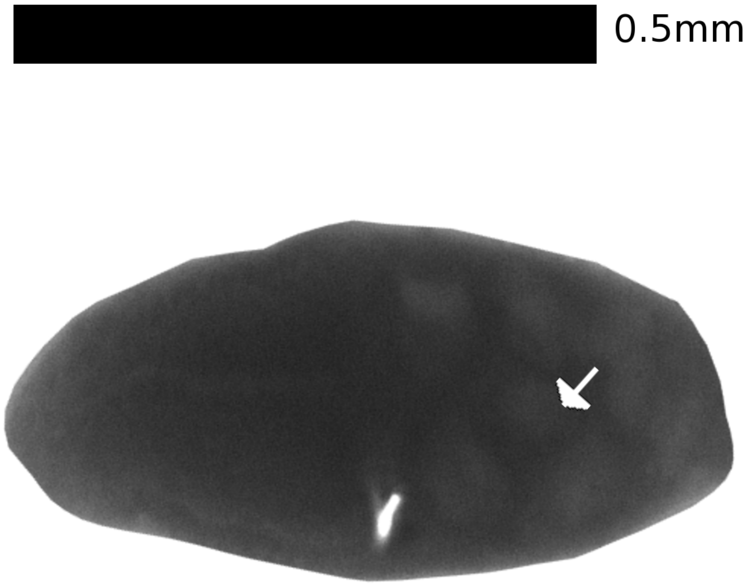


Figure 7.3: Ubiquitous staining. Stage 9 oocyte. Probe is DIG-labeled anti-*CG8183*, with NBT/BCIP staining. Arrow indicates nucleus of nurse cell, showing weakened staining.

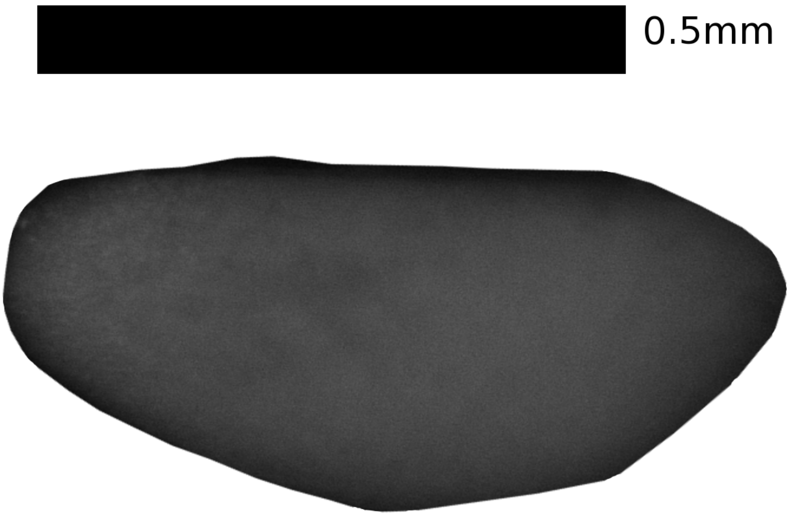


Figure 7.4: Absence of staining. Stage 10 oocyte. Probe is DIG-labeled anti-*CG3570*, with NBT/BCIP staining. Image colours have been inverted for clarity.

Gene	Staining
CG2519	None
CG3173	None
CG3218	Anterior localisation
CG3570	None
CG3937	Ubiquitous
CG4027	Ubiquitous
CG4886	Ubiquitous
CG4898	Ubiquitous
CG5452	None
CG6438	Ubiquitous
CG6667	Ubiquitous
CG6818	None
CG7865	None
CG8002	None
CG8183	Ubiquitous
CG8291	Ubiquitous
CG8967	None
CG10868	Anterior localisation
CG11107	None
CG11416	Ubiquitous
CG11614	Ubiquitous
CG11897	Ubiquitous
CG11986	None
CG13148	Ubiquitous
CG31133	None
CG31158	None
CG31694	None
CG32592	Ubiquitous
CG32791	None

Table 7.2: Staining patterns of experimentally tested genes

7.4 Discussion

The failure to observe subcellular localisation that mirrors that of *k10* indicates that the presence of a structural element containing the consensus structure described in chapter 6 is not sufficient to obtain this localisation pattern.

One of the supporting aspects of the original RNA structural analysis was the GOstat analysis. As with all statistical analysis, the results must be interpreted carefully. Performing the GOstat analysis again with the subset of the genes that are expressed in the ovaries gives no hits that are significant at the 5% level if *k10* and *orb* are excluded, and the results with K10 are shown in table 7.3.

Table 7.3: GOStat results

Go Annotation	Genes	P
oocyte anterior/posterior axis determination	TM1 ORB DL FS(1)K10	0.0106
oocyte axis determination	TM1 ORB DL FS(1)K10	0.0106
oocyte construction	TM1 ORB DL FS(1)K10	0.0106
oocyte development	TM1 ORB DL FS(1)K10	0.0106
oocyte differentiation	TM1 ORB dl FS(1)K10	0.0106
cell development	TM1 OTK ORB HIW ACT5C CHER FS(1)K10 DL	0.0106
anterior/posterior axis specification	TM1 ORB DL FS(1)K10	0.0106
localization	TM1 OTK ORB ACT5C CHER KHC-73 CYP33 DL FS(1)K10	0.0106
germarium-derived egg chamber formation	CHER ORB DL	0.0106

pole plasm mRNA localization	TM1 ORB FS(1)K10	0.0106
pole plasm RNA localization	TM1 ORB FS(1)K10	0.0106
protein binding	CHER KHC-73 TM1 PNGASE OTK ORB HIW DL	0.0109
germ cell development	TM1 ORB DL FS(1)K10	0.0109
pole plasm assembly	TM1 ORB FS(1)K10	0.0109
dorsal/ventral axis specification	ORB DL FS(1)K10	0.0109
anterior/posterior pattern formation	TM1 ORB DL FS(1)K10	0.0109
intracellular mRNA localization	TM1 ORB FS(1)K10	0.014
cell differentiation	TM1 OTK ORB HIW ACT5C CHER FS(1)K10 DL	0.0144
cellular developmental process	TM1 OTK ORB HIW ACT5C CHER FS(1)K10 DL	0.0163
axis specification	TM1 ORB DL FS(1)K10	0.0163
dorsal/ventral pattern formation	ORB DL FS(1)K10	0.0207
RNA localization	TM1 ORB FS(1)K10	0.0216
cytoplasm organization and biogenesis	TM1 ORB FS(1)K10	0.025
cellular component organization and biogenesis	TM1 OTK ORB HIW ACT5C CHER KHC-73 CYP33 FS(1)K10 DL	0.025
gamete generation	ACT5C CHER TM1 ORB DL FS(1)K10	0.025

macromolecule localization	CHER TM1 ORB FS(1)K10	0.025
sexual reproduction	ACT5C CHER TM1 ORB DL FS(1)K10	0.025
pyrimidine ribonucleoside monophosphate biosynthetic process	DNK	0.025
deoxynucleoside kinase activity	DNK	0.025
hatching behavior	AMON	0.025

Many of these results are more significant than previously. This is due to the sample set being smaller, but still being effectively constrained to include *k10* and *orb*. A more recent version of GOStat includes a correction method (Benjamini and Yekutieli, 2001) that accounts for dependencies between results, such as those introduced by the constraint on *k10* and *orb*. Applying this correction to the same dataset reveals no statistically significant results.

However, the occurrence of Tropomyosin and Dorsal in the dataset is still worthy of further investigation. Erdelyi et al. (1995) demonstrated that Tropomyosin is vital for anterior-posterior axis formation, but, in contrast to *k10*, Tropomyosin is localised to the posterior of the embryo (Hales et al., 1994). Further, this localisation appears limited to the embryo – it appears ubiquitously in the oocyte (above data, Hales et al. (1994)). Similarly, Dorsal is a maternally expressed mRNA involved in axis formation, but undergoes protein localisation rather than mRNA localisation (Rushlow et al., 1989).

A remaining question concerns the disparity between the observed lack of expression of certain genes despite FlyAtlas indicating that they are upregulated in ovarian tissue. Plausible explanations for this include the expression being limited to ovarian tissue other than the oocytes or the fact that the FlyAtlas protocol requires the collection of 1500ng of RNA before microarray analysis. This may result in the protocol being more sensitive than in-situ analysis in cases where the expression is low.

7.5 Conclusions

Performing in situ staining did not successfully demonstrate that the presence of a conserved structural element with a minor groove pattern similar to that of the *k10* TLS is sufficient to obtain a localisation pattern similar to that of *k10* and *orb*. More recent research has indicated TLS recognition is mediated by Egalitarian (Dienstbier et al., 2009), a dynein binding protein. Egalitarian has previously been implicated in the localisation of other mRNAs that lack the consensus sequence described here (Bullock and Ish-Horowicz, 2001). This strongly implies that the minor groove consensus sequence in the *k10* and *orb* localisation sequences is a function of sequence similarity constrained by the requirement to form a stable stem structure, rather than a direct result of functional conservation. There is thus no reason to believe that the minor groove pattern is sufficient for localisation – the failure to observe localisation in other transcripts carrying it is therefore unsurprising.

Chapter 8

Conclusion

This thesis has described multiple techniques utilised in attempting to locate novel functional elements in *D. melanogaster*, utilising comparative analysis as a mechanism for determining the likelihood of the significance of these elements. Two novel techniques have been described, along with validation of the discovery of two previously unannotated tRNAs and strong indications that a dataset of putative microRNAs contains functional sequences.

This thesis has concentrated on the utility of comparative genomics in providing further information about each of the putative elements located. In chapter 2, this demonstrated the apparent recent accumulation of a variant of the canonical drosophila 2S rRNA sequence in *D. melanogaster* along with evidence for a strain-specific mutation. The comparative analysis also demonstrated the plausibility of these mutations by showing similar (though not identical) accumulations in other species.

Despite these discoveries, the primary aim of chapter 2 was to investigate the quality level of the Solexa sequencing in order to provide a solid dataset for the following chapters. This was achieved by careful examination of sequences generated at two different stages of the sequencing process, allowing a fine-grained analysis of the sources of error. A third of the error appears to be derived from the preparation of the sequences, with the remaining two thirds being introduced by the sequencing itself. Both these error rates are low, providing confidence in the correctness of peaks analysed in the following chapters.

In chapter 3, an analysis of the alignment profile generated in chapter 2 was carried out with the aim of locating tRNA genes. This demonstrated that the expression profile of tRNA genes could be distinguished from that of most other expressed regions of the genome. In combination with existing tRNA prediction techniques, this provided robust evidence for the existence of two previously unannotated tRNA genes. Yet more evidence was provided by the new genes being present in the conserved syntenic region of all 11 other sequenced species. This more traditional use of conservational analysis makes use of the fact that well-conserved sequences are highly likely to be functional.

Similar techniques were put to use in chapter 4, where the measuring of conservation was used to judge whether an analysis of experimentally obtained sequence data predicted genuine microRNAs. The results were consistent with the dataset containing some number of novel microRNAs, though indicated that they were likely to be restricted to species more closely related to *D. melanogaster* than *D. pseudoobscura*. This enrichment may an artifact of the use of conservational analysis in previous studies of microRNA, demonstrating the risks of relying on this technique in order to reduce false positives from other prediction techniques. While the presence of conservation can be strongly indicative of functionality, its absence does not inherently indicate that the sequence in question has no function. Sequencing a range of closely related species makes it less likely that genuine functional elements will be ignored due to lack of conservation, but does not remove the possibility entirely.

While conservational analysis is usually explained in terms of primary sequence conservation, it also follows that other functional features will also be conserved. Chapter 5 uses conservational analysis to judge the significance of gene arrangement, rather than examining the conservation of the genes themselves. This demonstrated that the pattern of conservation was essentially consistent with the gain and loss of nested genes being a random process, but found that the only genes with a conserved nesting arrangement in both *D. melanogaster* and humans had a plausible functional relationship. It is therefore possible that a subset of nesting arrangements are functional.

Finally, chapters 6 and 7 attempted to use conservational analysis to identify structural mRNA elements that were considered likely to be functional, making use of the fact that secondary structure conservation implies primary sequence conservation in closely related species. Experimental investigation failed to demonstrate any link between these elements and their predicted function, indicating another risk of conservational analysis – the presence of conservation may be strongly indicative of a functional relationship, but does not guarantee that there is one. In combination with experimental or well-tested prediction algorithms, however, comparative genomics can provide a great deal of information.

This thesis has demonstrated that conservational genomics is a powerful tool for both the validation and rejection of putative functional elements. This makes it ideal for use in judging the efficacy of novel techniques for predicting the presence of these elements, providing the potential for rapid improvement in computational analysis tools.

Appendix A

Protocols

A.1 Preparing microRNA samples for Solexa sequencing

This protocol was provided by Illumina, Inc.

1. Purifying 20-30nt size-range from 10ug of Drosophila total RNA (1ug/uL)

   (a) Remove the comb from the 15% TBU gel and rinse out the wells thoroughly with 1X TBE.

   (b) Pre-run the 15% TBU gel for 15-30 min at 200V, and wash the wells using 1X TBE.

   (c) Mix 10μL (10μg) of total RNA with 10μL of 2X formamide loading dye in a 200uL PCR tube. Heat the sample at 65C for 5 min, spin down, and load the sample into one well.

   (d) Mix 2μL of 10bp ladder with 2μL of 2X loading dye in another 200uL PCR tube, and load into another well without heating.

(e) Run the gel at 200V for 1 hour, and stain the gel with 1X TBE / EtBr for 2 min.

(f) Cut out the corresponding gel band (20-30nt) and transfer to a 0.5mL tube with 4-5 pores punctured by a 21 gauge needle on the bottom.

(g) Set this tube into a 2ml round-bottom Eppendorf tube, and spin the gel through the hole into the 2mL tube at full speed for 2 min.

(h) Add 300μL of 0.3M NaCl to the tube, and elute the DNA by rotating the tube gently at room temperature for 4 hours.

(i) Transfer the elute and the gel debris onto the top of a Spin-X filter, and spin at full speed for 2 minutes.

(j) Add 750 μL of 100% EtOH and 3μL of glycogen to the sample, and incubate at -80C for 30 minutes.

(k) Spin down at˜14K rpm for 25 minutes at 4C in a microcentrifuge.

(l) Carefully remove supernatant and wash pellet with 750 μL of room temperature 75% EtOH. Allow the RNA pellet to air dry then dissolve the RNA in total of 4.7 μL of DEPC-treated water.

2. 5 prime Adaptor Ligation and Purification

   (a) Set up the 5 prime Adaptor ligation reaction:

(b)

Reagent	Amount
Purified 20-30nt RNA	4.7 μL
5pmole/μL 5 primeRNA Adaptor (28nt)	1.3 μL
10X Ligation Buffer	1 μL
T4 RNA Ligase (Ambion, 5U/uL)	2 μL
RNase Out (Invitrogen)	1 μL
Total reaction volume (μL)	10 μL

(c) Set up the 5 prime Adaptor ligation reaction for mir168 control:

(d)

Reagent	Amount
mir168 at 10pmole/μL	5 μL
100μM 5 primeRNA Adaptor (28nt)	2 μL
10X Ligation Buffer	1.1 μL
T4 RNA Ligase (Ambion, 5U/uL)	2 μL
RNase Out (Invitrogen)	1 μL
Total reaction volume (μL)	11.1 μL

(e) Incubate at room temperature for 6 hours.

(f) Stop reaction by adding 10 μL 2x Loading Dye. Heat sample/loading buffer at 65C for 5 minutes prior to loading.

(g) Prerun the 15% TBU gel for 15-30 minutes at 200V. Wash the wells with 1X TBE.

(h) Load 1μg (1μL) of 10bp DNA ladder (1μL of 10bp ladder + 1μL of 2X loading dye). Do not heat the ladder at 65C.

(i) Load the samples into other wells. Run gel at 200V for 1 hour. Stain the gel with 1X TBE / EtBr.

(j) Cut out the corresponding gel band (40-60nt) and transfer to a 0.5mL tube with 4-5 pores punctured by a 21 gauge needle on the bottom.

(k) Set this tube into a 2ml round-bottom Eppendorf tube, and spin the gel through the hole into the 2mL tube at full speed for 2 min.

(l) Add 300μL of 0.3M NaCl to the tube, and elute the DNA by rotating the tube gently at room temperature for 4 hours.

(m) Transfer the elute and the gel debris onto the top of a Spin-X filter, and spin at full speed for 2 minutes.

(n) Add 750 μL of 100% EtOH and 3μL of glycogen to the sample, and incubate at -80C for 30 minutes.

(o) Spin down at˜14K rpm for 25 minutes at 4C in a microcentrifuge.

(p) Carefully remove supernatant and wash pellet with 750 μL of room temperature 75% EtOH. Allow the RNA pellet to air dry then dissolve the RNA in total of 5.4 μL of DEPC-treated water (for mir168 control, dissolve in 4 μL).

3. 3 prime Adaptor Ligation and Purification

(a) Set up the 3 prime Adaptor ligation reaction:

Reagent	Amount
Purified 5 prime ligation product	5.4 μL
10pmole/μL 3 primeRNA Adaptor (24nt)	0.6 μL
10X Ligation Buffer	1 μL
T4 RNA Ligase (Ambion, 5U/μL)	2 μL
RNase Out	1 μL
Total reaction volume (μL)	10 μL

(b)

(c) Set up the 3 prime Adaptor ligation reaction for mir168 control:

(d)

Reagent	Amount
Purified 49nt oligo control prod	4 μL
100μM 3 primeRNA Adaptor (24nt)	2 μL
10X Ligation Buffer	1μL
T4 RNA Ligase (Ambion, 5U/μL)	2 μL
RNase Out	1 μL
Total reaction volume (μL)	10 μL

(e) Incubate at 20C for 6 hours.

(f) Stop reaction by adding 10 μL 2x Loading Dye. Heat sample/loading buffer at 65C for 5 minutes prior to loading.

(g) Prerun the 10% TBU gel for 15-30 minutes at 200V. Wash the wells with 1X TBE.

(h) Load 1μg (1μL) of 10bp DNA ladder (1μL of 10bp ladder + 1μL of 2X loading dye). Do not heat the ladder at 65C.

(i) Load the samples into other wells. Run gel at 200V for 1 hour. Stain the gel with 1X TBE / EtBr.

(j) Cut out the corresponding gel band (70-90nt) and transfer to a 0.5mL tube with 4-5 pores punctured by a 21 gauge needle on the bottom.

(k) Set this tube into a 2ml round-bottom Eppendorf tube, and spin the gel through the hole into the 2mL tube at full speed for 2 min.

(l) Add 300μL of 0.3M NaCl to the tube, and elute the DNA by rotating the tube gently at room temperature for 4 hours.

(m) Transfer the elute and the gel debris onto the top of a Spin-X filter, and spin at full speed for 2 minutes.

(n) Add 750 μL of 100% EtOH and 3μL of glycogen to the sample, and incubate at -80C for 30 minutes.

(o) Spin down at˜14K rpm for 25 minutes at 4C in a microcentrifuge.

(p) Carefully remove supernatant and wash pellet with 750 μL of room temperature 75% EtOH. Allow the RNA pellet to air dry then dissolve the RNA in total of 4.5μL of DEPC-treated water.

4. RT-PCR of small RNAs ligated with adaptors:

(a) Set up a reverse transcription reaction:

(b)

Reagent	Amount
Purified ligated RNA	4.5 μL
100uM	RT-Primer
(3 prime PCR primer)	0.5 μL

(c) Heat to 65C for 10 minutes, spin down to cool.

(d) Add following in order:

(e)

Reagent	Amount
2.0 μL	5x first strand buffer
0.5 μL	12.5mM dNTP
1 μL	100 mM DTT
0.5 μL	RNaseOut

(f) Heat to 48C for 3 min and then add 1.0 μL of Superscript II RT (200U/μL).

(g) Incubate at 44C for 1 hour.

(h) Set up 50μL SOEPCR reactions from the RT samples

(i)

Reagent	Amount
RT Reaction	10
5X PCR Buffer	10
25mM dNTP	0.5
Sbs11-p5	0.5
25 uM 5 prime PCR Primer	0.5
25 uM 3 prime PCR Primer	0.5
Phusion (NEB) high fidelity DNA pol	0.5
Water	27.5

(j) Set up 50μL SOEPCR reactions from the mir168 control

(k)

Reagent	Amount
RT Reaction	1
5X PCR Buffer	10
25mM dNTP	0.5
Sbs11-p5	0.5
25 uM 5 prime PCR Primer	0.5
25 uM 3 prime PCR Primer	0.5
Phusion (NEB)	0.5
Water	36.5

(l)

PCR temperature	time
98C	30 sec
98C	10 sec
60C	30 sec for 15 Cycles of PCR
72C	15 sec
72C	10 min

(m) Carefully load 50μl of PCR products into 2 wells of 412% TBE PAGE gel. (NOT UREA) Electrophorese 45 minutes at 200V.

(n) Pry apart cassette, and stain the gel in TE /ethidium bromide in a clean container for 2-3 minutes.

(o) Cut out ˜375 bp band with a clean razor blade, and put band into a 0.5ml Eppendorf tube whose bottom has been punctured by a 21 gauge needle.

(p) Set this tube into a 2ml round-bottom Eppendorf tube, and spin the gel through the hole into the 2ml tube (2 min spin at full speed in microfuge).

(q) Add 100 μl of 1×NEB to the gel, and elute the DNA by rotating the tube gently at room temperature for 2 hours.

(r) Transfer the eluate and the gel debris onto the top of a Spin-X filter. Spin the filter in the microfuge for 2 minutes at full speed..

(s) Add 1μl of Pellet Paint, 10μl of 3 M NaOAc and 325μl of 20C EtOH, spin at 14K for 20 mins.

(t) Wash with 500μl of RT 70% EtOH, vacuum dry and resuspend in 10μl of EB solution (10mM tris-HCl, pH 8.5)

A.2 In-situ hybridisation of *D. melanogaster* oocytes

This protocol is an adaptation of the Berkeley Drosophila Genome Project 96-well embryo in situ hybridisation protocol (BDGP, 2005).

1. Extract ovaries from female *D. melanogaster* and place in PBTween

2. Fix embryos by gently shaking in 50-50 mix of heptane and 4% formaldehyde/PBS fixative for 25 min.

3. Remove lower aqueous phase and replace with equal volume of methanol.
4. Wash 3× in methanol.
5. Store oocytes at -20 C25.
6. Rehydrate in 3:1 methanol:2.5% formaldehyde in 1× PBS for 2 min.
7. Rehydrate in 1:3 methanol:2.5% formaldehyde in 1× PBS for 5 min.
8. Post-fix in 2.5% formaldehyde in 1× PBS for 10 min.
9. Rinse 6x in PBT.
10. Add 3 ml of hybridization buffer
11. Incubate with shaking at 125 rpm on the Gyrotory shaker for at least 1 hr at room temperature to pre-hybridize oocytes.
12. During pre-hybridization put 200 μl of hybridization buffer with dextran sulfate into each well of a 96-well plate using multichannel pipette.
13. Add 2 μl of probe into each well
14. Mix thoroughly on a vortex mixer at maximum speed for 25 sec and centrifuge at 4000 rpm for one minute.
15. Add 20 μl of oocytes into each well of a 96-well filter plate (using a multichannel pipette with wide opening tips)
16. Transfer the probes from the 96-well plate into the 96-well filter plate and seal the filter plate with an aluminum foil sealer.
17. Incubate at 55C with shaking at 125 rpm on the Gyrotory shaker overnight 33.

18. Add 100 μl of room temperature wash buffer.
19. Remove the hybridization-buffer, wash-buffer mix using vacuum; once all the liquid is removed from the wells quickly turn off the vacuum.
20. Rinse 2× with wash buffer.
21. Incubate in wash buffer at 55C with shaking for 30 min with eight changes.
22. Incubate in wash buffer at 55C with shaking overnight.
23. Rinse in PBT.
24. Incubate in PBT at RT with shaking for 30 min; remove PBT.
25. Incubate in PBT, 5% goat serum, 1:2000 dilution Anti-Digoxigenin-AP Fab Fragments at RT with shaking for 2 hrs.
26. Rinse 2× with PBT.
27. Incubate in PBT at RT with shaking 9× for 10 min each.
28. Rinse 2× with AP buffer.
29. Wash in AP buffer at RT for 5 min; remove AP buffer.
30. Add developing solution
31. Incubate with shaking until desired colour development is achieved (about 75 min); remove developing solution by vacuum aspiration.
32. Rinse 3× in PBT to stop the color reaction.
33. Rinse 6x in ethanol.
34. Rinse in PBT.

35. Add 70% glycerol.

36. Store at 4C.

37. Check individual wells on the plate under a low power magnification microscope.

38. Oocytes are ready to be photographed.

Bibliography

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidén-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin, and J. C. Venter (2000). The genome sequence of drosophila melanogaster. *Science (New York,* *N.Y.) 287*(5461), 2185–2195.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research 25*(17), 3389–3402.

Ambros, V., B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl (2003). A uniform system for microrna annotation. *RNA (New York, N.Y.) 9*(3), 277–279.

Anderson, S. (1981). Shotgun dna sequencing using cloned dnase i-generated fragments. *Nucleic acids research 9*(13), 3015–3027.

BDGP (2005). 96-well rna in-situ hybridization protocol. `http://www.fruitfly.org/about/methods/RNAinsitu.html`.

Beissbarth, T. and T. P. Speed (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics 20*(9), 1464–5.

Bellman, R. E. (1957). *Dynamic Programming*. Dover Publications, Incorporated.

Benjamini, Y. and D. Yekutieli (2001). THE CONTROL OF THE FALSE DISCOVERY RATE IN MULTIPLE TESTING UNDER DEPENDENCY. *The Annals of Statistics 29*, 1165–1188.

Bennett, S. T., C. Barnes, A. Cox, L. Davies, and C. Brown (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics 6*(4), 373–382.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler (2007). Genbank. *Nucleic acids research 35*(Database issue), D21–5.

Berezikov, E., F. Thuemmler, L. W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R. H. A. Plasterk (2006). Diversity of micrornas in human and chimpanzee brain. *Nature genetics 38*(12), 1375–1377.

Berleth, T., M. Burri, G. Thoma, D. Bopp, S. Richstein, G. Frigerio, M. Noll, and C. Nusslein-Volhard (1988). The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. *EMBO 7*(6), 1748–1756.

Bernstein, E., A. A. Caudy, S. M. Hammond, and G. J. Hannon (2001). Role for a bidentate ribonuclease in the initiation step of rna interference. *Nature 409*(6818), 363–366.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science (New York, N.Y.) 299*(5611), 1391–1394.

Bor, V. V. D., E. Hartswood, C. Jones, D. Finnegan, and I. Davis (2005). gurken and the i factor retrotransposon rnas share common localization signals and machinery. *Developmental Cell 9*(1), 51–62.

Breitwieser, W., F. Markussen, H. Horstmann, and A. Ephrussi (1996). Oskar protein interaction with Vasa represents an essential step in polar granule assembly. *Genes and Development 10*(17), 2179–2188.

Brown, H., F. Sanger, and R. Kitai (1955, Aug). The structure of pig and sheep insulins. *Biochem. J. 60*, 556–565.

Bullock, S. L. and D. Ish-Horowicz (2001, Dec). Conserved signals and machinery for RNA transport in Drosophila oogenesis and embryogenesis. *Nature 414*, 611–616.

Chan, S. K., S. B. Needleman, J. W. Stewart, O. F. Walasek, and E. Margoliash (1963). *Fed. Proc 22*, 658–660.

Chang, J. S., L. Tan, and P. Schedl (1999). The Drosophila CPEB homolog, orb, is required for oskar protein expression in oocytes. *Dev Biol 215*(1), 91–106.

Cheung, H., T. Serano, and R. Cohen (1992). Evidence for a highly selective RNA transport system and its role in establishing the dorsoventral axis of the Drosophila egg. *Development 114*(3), 653–661.

Chintapalli, V. R., J. Wang, and J. A. T. Dow (2007). Using flyatlas to identify better drosophila melanogaster models of human disease. *Nature genetics 39*(6), 715–720.

Clark, A., G. Gibson, T. Kaufman, E. Myers, and P. O'Grady (2003). Proposal for drosophila as a model system for comparative genomics.

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T. C. Kaufman, M. Kellis, W. Gelbart, V. N. Iyer, D. A. Pollard, T. B. Sackton, A. M. Larracuente, N. D. Singh, J. P. Abad, D. N. Abt, B. Adryan, M. Aguade, H. Akashi, W. W. Anderson, C. F. Aquadro, D. H. Ardell, R. Arguello, C. G. Artieri, D. A. Barbash, D. Barker, P. Barsanti, P. Batterham, S. Batzoglou, D. Begun, A. Bhutkar, E. Blanco, S. A. Bosak, R. K. Bradley, A. D. Brand, M. R. Brent, A. N. Brooks, R. H. Brown, R. K. Butlin, C. Caggese, B. R. Calvi, A. Bernardo de Carvalho, A. Caspi, S. Castrezana, S. E. Celniker, J. L. Chang, C. Chapple, S. Chatterji, A. Chinwalla, A. Civetta, S. W. Clifton, J. M.

Comeron, J. C. Costello, J. A. Coyne, J. Daub, R. G. David, A. L. Delcher, K. Delehaunty, C. B. Do, H. Ebling, K. Edwards, T. Eickbush, J. D. Evans, A. Filipski, S. Findeiss, E. Freyhult, L. Fulton, R. Fulton, A. C. L. Garcia, A. Gardiner, D. A. Garfield, B. E. Garvin, G. Gibson, D. Gilbert, S. Gnerre, J. Godfrey, R. Good, V. Gotea, B. Gravely, A. J. Greenberg, S. Griffiths-Jones, S. Gross, R. Guigo, E. A. Gustafson, W. Haerty, M. W. Hahn, D. L. Halligan, A. L. Halpern, G. M. Halter, M. V. Han, A. Heger, L. Hillier, A. S. Hinrichs, I. Holmes, R. A. Hoskins, M. J. Hubisz, D. Hultmark, M. A. Huntley, D. B. Jaffe, S. Jagadeeshan, W. R. Jeck, J. Johnson, C. D. Jones, W. C. Jordan, G. H. Karpen, E. Kataoka, P. D. Keightley, P. Kheradpour, E. F. Kirkness, L. B. Koerich, K. Kristiansen, D. Kudrna, R. J. Kulathinal, S. Kumar, R. Kwok, E. Lander, C. H. Langley, R. Lapoint, B. P. Lazzaro, S.-J. Lee, L. Levesque, R. Li, C.-F. Lin, M. F. Lin, K. Lindblad-Toh, A. Llopart, M. Long, L. Low, E. Lozovsky, J. Lu, M. Luo, C. A. Machado, W. Makalowski, M. Marzo, M. Matsuda, L. Matzkin, B. McAllister, C. S. McBride, B. McKernan, K. McKernan, M. Mendez-Lago, P. Minx, M. U. Mollenhauer, K. Montooth, S. M. Mount, X. Mu, E. Myers, B. Negre, S. Newfeld, R. Nielsen, M. A. F. Noor, P. O'Grady, L. Pachter, M. Papaceit, M. J. Parisi, M. Parisi, L. Parts, J. S. Pedersen, G. Pesole, A. M. Phillippy, C. P. Ponting, M. Pop, D. Porcelli, J. R. Powell, S. Prohaska, K. Pruitt, M. Puig, H. Quesneville, K. R. Ram, D. Rand, M. D. Rasmussen, L. K. Reed, R. Reenan, A. Reily, K. A. Remington, T. T. Rieger, M. G. Ritchie, C. Robin, Y.-H. Rogers, C. Rohde, J. Rozas, M. J. Rubenfield, A. Ruiz, S. Russo, S. L. Salzberg, A. Sanchez-Gracia, D. J. Saranga, H. Sato, S. W. Schaeffer, M. C. Schatz, T. Schlenke, R. Schwartz, C. Segarra, R. S. Singh, L. Sirot, M. Sirota, N. B. Sisneros, C. D. Smith, T. F. Smith, J. Spieth, D. E. Stage, A. Stark, W. Stephan, R. L. Strausberg, S. Strempel, D. Sturgill, G. Sutton, G. G. Sutton, W. Tao, S. Teichmann, Y. N. Tobari, Y. Tomimura, J. M. Tsolas, V. L. S. Valente, E. Venter, J. C. Venter, S. Vicario, F. G. Vieira, A. J. Vilella, A. Villasante, B. Walenz, J. Wang, M. Wasserman, T. Watts, D. Wilson, R. K. Wilson, R. A. Wing, M. F. Wolfner, A. Wong, G. K.-S. Wong, C.-I. Wu, G. Wu, D. Yamamoto, H.-P. Yang, S.-P. Yang, J. A. Yorke, K. Yoshida, E. Zdobnov, P. Zhang, Y. Zhang, A. V. Zimin, J. Baldwin, A. Abdouelleil, J. Abdulkadir, A. Abebe, B. Abera, J. Abreu, S. C. Acer, L. Aftuck, A. Alexander, P. An, E. Anderson, S. Anderson, H. Arachi, M. Azer, P. Bachantsang, A. Barry, T. Bayul, A. Berlin, D. Bessette, T. Bloom, J. Blye, L. Boguslavskiy, C. Bonnet, B. Boukhgalter, I. Bourzgui, A. Brown, P. Cahill, S. Channer, Y. Cheshatsang, L. Chuda, M. Citroen, A. Collymore, P. Cooke, M. Costello, K. D'Aco, R. Daza, G. De Haan, S. DeGray, C. DeMaso, N. Dhargay, K. Dooley, E. Dooley, M. Doricent, P. Dorje, K. Dorjee, A. Dupes, R. Elong, J. Falk, A. Farina, S. Faro, D. Ferguson, S. Fisher, C. D. Foley, A. Franke, D. Friedrich, L. Gadbois, G. Gearin, C. R. Gearin, G. Giannoukos, T. Goode, J. Graham, E. Grandbois, S. Grewal, K. Gyaltsen, N. Hafez, B. Hagos, J. Hall, C. Henson, A. Hollinger, T. Honan, M. D. Huard, L. Hughes, B. Hurhula, M. E. Husby,

A. Kamat, B. Kanga, S. Kashin, D. Khazanovich, P. Kisner, K. Lance, M. Lara, W. Lee, N. Lennon, F. Letendre, R. LeVine, A. Lipovsky, X. Liu, J. Liu, S. Liu, T. Lokyitsang, Y. Lokyitsang, R. Lubonja, A. Lui, P. MacDonald, V. Magnisalis, K. Maru, C. Matthews, W. McCusker, S. McDonough, T. Mehta, J. Meldrim, L. Meneus, O. Mihai, A. Mihalev, T. Mihova, R. Mittelman, V. Mlenga, A. Montmayeur, L. Mulrain, A. Navidi, J. Naylor, T. Negash, T. Nguyen, N. Nguyen, R. Nicol, C. Norbu, N. Norbu, N. Novod, B. O'Neill, S. Osman, E. Markiewicz, O. L. Oyono, C. Patti, P. Phunkhang, F. Pierre, M. Priest, S. Raghuraman, F. Rege, R. Reyes, C. Rise, P. Rogov, K. Ross, E. Ryan, S. Settipalli, T. Shea, N. Sherpa, L. Shi, D. Shih, T. Sparrow, J. Spaulding, J. Stalker, N. Stange-Thomann, S. Stavropoulos, C. Stone, C. Strader, S. Tesfaye, T. Thomson, Y. Thoulutsang, D. Thoulutsang, K. Topham, I. Topping, T. Tsamla, H. Vassiliev, A. Vo, T. Wangchuk, T. Wangdi, M. Weiand, J. Wilkinson, A. Wilson, S. Yadav, G. Young, Q. Yu, L. Zembek, D. Zhong, (2007). Evolution of genes and genomes on the drosophila phylogeny. *Nature 450*(7167), 203–218.

Clark, D. V. and S. Henikoff (1992). Unusual organizational features of the drosophila gart locus are not conserved within diptera. *Journal of molecular evolution 35*(1), 51–59.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston (2003). Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science (New York, N.Y.) 301*(5629), 71–76.

Cohen, R. S., S. Zhang, and G. D. Dollar (2005). The positional, structural and sequence requirements of the Drosophila TLS RNA localizaiton element. *RNA 11*(7), 1017–1029.

Dienstbier, M., F. Boehl, X. Li, and S. L. Bullock (2009, Jul). Egalitarian is a selective RNA-binding protein linking mRNA localization signals to the dynein motor. *Genes Dev. 23*, 1546–1558.

Eck, R. V. and M. O. Dayhoff (1966, Apr). Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science 152*, 363–366.

Eddy, S. R. (2004). How do rna folding algorithms work? *Nature biotechnology 22*(11), 1457–1458.

Erdelyi, M., A. M. Michon, A. Guichet, J. B. Glotzer, and A. Ephrussi (1995, Oct). Requirement for Drosophila cytoplasmic tropomyosin in oskar mRNA localization. *Nature 377*, 524–527.

Ewing, B., L. Hillier, M. Wendl, and P. Green (1998). Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research 7*, 175–185.

Ferrandon, D., I. Koch, E. Westhof, and C. Nusslein-Volhard (1997). RNA-RNA interaction is required for the formation of specific bicoid mRNA 3′ UTR-STAUFEN ribonucleoprotein particles. *EMBO J 16*(7), 1751–8.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert (1976). Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature 260*(5551), 500–507.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science (New York, N.Y.) 269*(5223), 496–512.

FlyBase Consortium. (2003, Jan). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res 31*(1), 172–175.

Frederico, L. A., T. A. Kunkel, and B. R. Shaw (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry 29*(10), 2532–2537.

Gautheret, D. and A. Lambert (2001). Direct rna motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of molecular biology 313*(5), 1003–1011.

Gonzales-Reyes, A., H. Elliott, and D. St Johnston (1995). Polarization of both major body axes in Drosophila by gurken-torpedo signalling. *Nature 375*(6533), 654–658.

Haenlin, M., W. MacDonald, C. Cibert, and E. Mohier (1995). The angle of the dorsoventral axis with respect to the anteroposterior axis in the Drosophila embryo is controlled by the distribution of gurken mRNA in the oocyte. *Mechanisms of Development 49*(1-2), 97–106.

Hahn, M. W., M. V. Han, and S.-G. Han (2007). Gene family evolution across 12 drosophila genomes. *PLoS Genet 3*(11), e197.

Hales, K. H., J. E. Meredith, and R. V. Storti (1994, Oct). Transcriptional and post-transcriptional regulation of maternal and zygotic cytoskeletal tropomyosin mRNA during Drosophila development correlates with specific morphogenic events. *Dev. Biol. 165*, 639–653.

Hardison, R. C., J. Oeltjen, and W. Miller (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome research 7*(10), 959–966.

Henikoff, S., M. A. Keene, K. Fechtel, and J. W. Fristrom (1986). Gene within a gene: nested drosophila genes encode unrelated proteins on opposite dna strands. *Cell 44*(1), 33–42.

Hill, R. L., C. M. Harris, J. F. Naylor, and W. M. Sams (1969, Apr). The partial amino acid sequence of human myoglobin. *J. Biol. Chem. 244*, 2182–2194.

Hofacker, I. L. (2003). Vienna rna secondary structure server. *Nucleic acids research 31*(13), 3429–3431.

Hudson, S., M. Garrett, J. Carlson, G. Micklem, S. Celniker, E. Goldstein, and S. Newfeld (2007). Phylogenetic and genomewide analyses suggest a functional relationship between kayak, the drosophila fos homolog, and fig, a predicted protein phosphatase 2c nested within a kayak intron. *Genetics 177*(3), 1349–1361.

Illumina (2006). Solexa sequencing-by-synthesis demo. `http://www.illumina.com/media.ilmn?Title=Sequencing-By-Synthesis%20Demo%&Cap=&PageName=solexa%20technology&PageURL=203&Media=1`.

Itano, H. A., W. R. Bergren, and P. Sturgeon (1956, May). The abnormal human hemoglobins. *Medicine (Baltimore) 35*, 121–159.

Jeffery, W. R., C. R. Tomlinson, and R. D. Brodeur (1983, Oct). Localization of actin messenger RNA during early ascidian development. *Dev Biol 99*(2), 408–417.

Jongens, T. A., L. D. Ackerman, J. R. Swedlow, L. Y. Jan, and Y. N. Jan (1994, Sep). Germ cell-less encodes a cell type-specific nuclear pore-associated protein and functions early in the germ-cell specification pathway of Drosophila. *Genes Dev 8*(18), 2123–2136.

Jongens, T. A., B. Hay, L. Y. Jan, and Y. N. Jan (1992, Aug). The germ cell-less gene product: a posteriorly localized component necessary for germ cell development in Drosophila. *Cell 70*(4), 569–584.

Jordan, B. R., R. Jourdan, and B. Jacq (1976, Feb). Late steps in the maturation of Drosophila 26 S ribosomal RNA: Generation of 5.8 S and 2 S RNAs by cleavages occurring in the cytoplasm. *J Mol Biol 101*(1), 85–105.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature 423*(6937), 241–254.

Kendrew, J. C., R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore (1960, Feb). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. *Nature 185*, 422–427.

Klein, R. J. and S. R. Eddy (2003, Sep). RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics 4*(1), 44.

Kreil, G. (1965). [THE C-TERMINAL AMINO ACID SEQUENCE OF TUNA FISH CYTOCHROME C.]. *Hoppe-Seyler's Z. Physiol. Chem. 340*, 86–87.

Kulesh, D. A., D. R. Clive, D. S. Zarlenga, and J. J. Greene (1987). Identification of interferon-modulated proliferation-related cdna sequences. *Proceedings of the National Academy of Sciences of the United States of America 84*(23), 8453–8457.

Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl (2001). Identification of novel genes coding for small expressed rnas. *Science (New York, N.Y.) 294*(5543), 853–858.

Lai, E. C., P. Tomancak, R. W. Williams, and G. M. Rubin (2003). Computational identification of drosophila microrna genes. *Genome biology 4*(7), R42.

Lashkari, D. A., J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America 94*(24), 13057–13062.

Laslett, D. and B. Canback (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research 32*(1), 11–16.

Lau, N. C., L. P. Lim, E. G. Weinstein, and D. P. Bartel (2001). An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. *Science (New York, N.Y.) 294*(5543), 858–862.

Lee, M. P., S. D. Brown, A. Chen, and T. S. Hsieh (1993). Dna topoisomerase i is essential in drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America 90*(14), 6656–6660.

Lee, R. C. and V. Ambros (2001). An extensive class of small rnas in caenorhabditis elegans. *Science (New York, N.Y.) 294*(5543), 862–864.

Lee, R. C., R. L. Feinbaum, and V. Ambros (1993). The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell 75*(5), 843–854.

Lehmann, R. and C. Nusslein-Volhard (1986). Abdominal segmentation, pole cell formation, and embryonic polarity require the localized activity of oskar, a maternal gene in Drosophila. *Cell 47*(1), 141–152.

Lewis, B., I. Shih, M. Jones-Rhoades, D. Bartel, and C. Burge (2003, Dec). Prediction of mammalian microRNA targets. *Cell 115*, 787–798.

Lim, L. P., M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel (2003). Vertebrate microrna genes. *Science (New York, N.Y.) 299*(5612), 1540.

Liu, Y., X. S. Liu, L. Wei, R. B. Altman, and S. Batzoglou (2004). Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome research 14*(3), 451–458.

Lowe, T. M. and S. R. Eddy (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research 25*, 966–964.

Lu, C., K. Kulkarni, F. F. Souret, R. MuthuValliappan, S. S. Tej, R. S. Poethig, I. R. Henderson, S. E. Jacobsen, W. Wang, P. J. Green, and B. C. Meyers (2006). Micrornas and other small rnas enriched in the arabidopsis rna-dependent rna polymerase-2 mutant. *Genome research 16*(10), 1276–1288.

Lu, C., S. S. Tej, S. Luo, C. D. Haudenschild, B. C. Meyers, and P. J. Green (2005). Elucidation of the small rna component of the transcriptome. *Science (New York, N.Y.) 309*(5740), 1567–1569.

Ma, K. and D. Huen (2005). Personal communication.

MacDonald, P. M. (1990). bicoid mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. *Development 110*(1), 161–71.

Macdonald, P. M. and K. Kerr (1998). Mutational analysis of an RNA recognition element that mediates localization of bicoid mRNA. *Mol Cell Biol 18*(7), 3788–95.

Margoliash, E. (1963, Oct). PRIMARY STRUCTURE AND EVOLUTION OF CYTOCHROME C. *Proc. Natl. Acad. Sci. U.S.A. 50*, 672–679.

Margoliash, E., E. L. Smith, G. Kreil, and H. Tuppy (1961, Dec). Amino-acid sequence of horse heart cytochrome c. *Nature 192*, 1125–1127.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature 437*(7057), 376–380.

Martin, D., C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq (2004). Gotoolbox: functional analysis of gene datasets based on gene ontology. *Genome biology 5*(12), R101.

Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *PNAS 101*(19), 7287–7292.

Matsubara, H. and E. L. Smith (1962, Nov). The amino acid sequence of human heart cytochrome c. *J. Biol. Chem. 237*, 3575–3576.

McGregor, A. P. (2005). How to get ahead: the origin, evolution and function of bicoid. *BioEssays : news and reviews in molecular, cellular and developmental biology 27*(9), 904–913.

Miller, S., M. Yasuda, J. K. Coats, Y. Jones, M. E. Martone, and M. Mayford (2002). Disruption of dendritic translation of camkiialpha impairs stabilization of synaptic plasticity and memory consolidation. *Neuron 36*(3), 507–519.

Min Jou, W., G. Haegeman, M. Ysebaert, and W. Fiers (1972). Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein. *Nature 237*(5350), 82–88.

Misra, S., M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell, P. Hradecky, Y. Huang, J. S. Kaminker, G. H. Millburn, S. E. Prochnik, C. D. Smith, J. L. Tupy, E. J. Whitfied, L. Bayraktaroglu, B. P. Berman, B. R. Bettencourt, S. E. Celniker, A. D. N. J. de Grey, R. A. Drysdale, N. L. Harris, J. Richter, S. Russo, A. J. Schroeder, S. Q. Shu, M. Stapleton, C. Yamada, M. Ashburner, W. M. Gelbart, G. M. Rubin, and S. E. Lewis (2002). Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol 3*(12), RESEARCH0083.

Nakamura, A., R. Amikura, M. Mukai, S. Kobayashi, and P. F. Lasko (1996, Dec). Requirement for a noncoding RNA in Drosophila polar granules for germ cell establishment. *Science 274*(5295), 2075–2079.

Narita, K., K. Titani, Y. Yaoi, H. Murakami, M. Kimura, and J. Vanecek (1963, Aug). PEPTIDES FROM A TRYPTIC DIGEST OF BAKER'S YEAST CYTOCHROME C. *Biochim. Biophys. Acta 73*, 670–673.

Needleman, S. and C. Wunsch (1970, March). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*(3), 443–453.

Neuman-Silberberg, F. S. and T. Schupbach (1996). The Drosophila TGF-alpha-like protein Gurken: expression and cellular localization during Drosophila oogenesis. *Mech Dev 59*(2), 105–13.

Padmanabhan, R. and R. Wu (1972). Nucleotide sequence analysis of dna. ix. use of oligonucleotides of defined sequence as primers in dna sequence analysis. *Biochemical and biophysical research communications 48*(5), 1295–1302.

Pauling, L. and H. A. Itano (1949, Nov). Sickle cell anemia a molecular disease. *Science 110*, 543–548.

Perutz, M. F., M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. North (1960, Feb). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature 185*, 416–422.

Pesole, G., S. Liuni, and M. D'Souza (2000). Patsearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics (Oxford, England) 16*(5), 439–450.

Remm, M., C. E. Storm, and E. L. Sonnhammer (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology 314*(5), 1041–1052.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, O. Couronne, S. Hua, M. A. Smith, P. Zhang, J. Liu, H. J. Bussemaker, M. F. van Batenburg, S. L. Howells, S. E. Scherer, E. Sodergren, B. B. Matthews, M. A. Crosby, A. J. Schroeder, D. Ortiz-Barrientos, C. M. Rives, M. L. Metzker, D. M. Muzny, G. Scott, D. Steffen, D. A. Wheeler, K. C. Worley, P. Havlak, K. J. Durbin, A. Egan, R. Gill, J. Hume, M. B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K. P. Clerc-Blankenburg, I. Dubchak, M. A. F. Noor, W. Anderson, K. P. White, A. G. Clark, S. W. Schaeffer, W. Gelbart, G. M.

Weinstock, and R. A. Gibbs (2005). Comparative genome sequencing of drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome research 15*(1), 1–18.

Rivera-Pomar, R., D. Niessing, U. Schmidt-Ott, W. Gehring, and H. Jackle (1996). RNA binding and translational suppression by bicoid. *Nature 379*(6567), 746–749.

Rozen, S. and H. Skaletsky (2000). Primer3 on the www for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.) 132*, 365–386.

Ruby, J. G., C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel (2006). Large-scale sequencing reveals 21u-rnas and additional micrornas and endogenous sirnas in c. elegans. *Cell 127*(6), 1193–1207.

Ruby, J. G., A. Stark, W. K. Johnston, M. Kellis, D. P. Bartel, and E. C. Lai (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of drosophila micrornas. *Genome Res 17*(12), 1850–1864.

Rushlow, C. A., K. Han, J. L. Manley, and M. Levine (1989, Dec). The graded distribution of the dorsal morphogen is initiated by selective nuclear transport in Drosophila. *Cell 59*, 1165–1177.

Ryle, A. P., F. Sanger, L. F. Smith, and R. Kitai (1955, Aug). The disulphide bonds of insulin. *Biochem. J. 60*, 541–556.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith (1977). Nucleotide sequence of bacteriophage phi x174 dna. *Nature 265*(5596), 687–695.

Sanger, F. and A. R. Coulson (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology 94*(3), 441–448.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America 74*(12), 5463–5467.

Saunders, C. and R. S. Cohen (1999). The role of oocyte transcription, the 5'UTR, and translation repression and derepression in Drosophila gurken mRNA and protein localization. *Mol Cell 3*(1), 43–54.

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science (New York, N.Y.) 270*(5235), 467–470.

Schupbach, T. (1987). Germ line and soma cooperate during oogenesis to establish the dorsoventral pattern of egg shell and embryo in Drosophila melanogaster. *Cell 49*(5), 699–707.

Seeger, M. A. and T. C. Kaufman (1990). Molecular analysis of the bicoid gene from Drosophila pseudoobscura: identification of conserved domains within coding and noncoding regions of the bicoid mRNA. *EMBO J 9*(9), 2977–87.

Serano, T. L. and R. S. Cohen (1995, Nov). A small predicted stem-loop structure mediates oocyte localization of Drosophila K10 mRNA. *Development 121*(11), 3809–3818.

Shykind, B. M., J. Kim, L. Stewart, J. J. Champoux, and P. A. Sharp (1997). Topoisomerase i enhances tfiid-tfiia complex assembly during activation of transcription. *Genes and development 11*(3), 397–407.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood (1986). Fluorescence detection in automated dna sequence analysis. *Nature 321*(6071), 674–679.

Smith, T. F. and M. S. Waterman (1981). Identification of common molecular subsequences. *J Mol Biol 147*(1), 195–7.

Staden, R. (1980). A computer program to search for tRNA genes. *Nucleic Acids Research 8*(4), 817–825.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, M. Kellis, M. A. Crosby, B. B. Matthews, A. J. Schroeder, L. S. Gramates, S. E. St Pierre, M. Roark, K. L. Wiley, R. J. Kulathinal, P. Zhang, K. V. Myrick, and J. V. Antone (2007). Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature 450*(7167), 219–232.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of molecular biology 203*(2), 439–455.

Takada, S., J. T. Lis, S. Zhou, and R. Tjian (2000). A trf1:brf complex directs drosophila rna polymerase iii transcription. *Cell 101*(5), 459–469.

Tamura, K., S. Subramanian, and S. Kumar (2004). Temporal patterns of fruit fly (drosophila) evolution revealed by mutation clocks. *Molecular Biology and Evolution 21*, 36–44(9).

Tan, L., J. S. Chang, A. Costa, and P. Schedl (2001). An autoregulatory feedback loop directs the localized expression of the Drosophila CPEB protein Orb in the developing oocyte. *Development 128*(7), 1159–69.

Thio, G. L., R. P. Ray, G. Barcelo, and T. Schupbach (2000). Localization of gurken RNA in Drosophila oogenesis requires elements in the 5′ and 3′ regions of the transcript. *Dev Biol 221*(2), 435–46.

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania (2003). Panther: a library of protein families and subfamilies indexed by function. *Genome research 13*(9), 2129–2141.

Wang, C. and R. Lehmann (1991, Aug). Nanos is the localized posterior determinant in Drosophila. *Cell 66*(4), 637–647.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigó, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P.

Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S.-P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature 420*(6915), 520–562.

Watson, H. C. and J. C. Kendrew (1961, May). The amino-acid sequence of sperm whale myoglobin. Comparison between the amino-acid sequences of sperm whale myoglobin and of human hemoglobin. *Nature 190*, 670–672.

Wharton, R. P. and G. Struhl (1991, Nov). RNA regulatory elements mediate control of Drosophila body pattern by the posterior morphogen nanos. *Cell 67*(5), 955–967.

Winter, F., S. Edaye, A. Hüttenhofer, and C. Brunel (2007). Anopheles gambiae mirnas as actors of defence reaction against plasmodium invasion. *Nucleic acids research 35*(20), 6953–6962.

Wu, R. and A. D. Kaiser (1968). Structure and base sequence in the cohesive ends of bacteriophage lambda dna. *Journal of molecular biology 35*(3), 523–537.

Xu, L., L. Yang, K. Hashimoto, M. Anderson, G. Kohlhagen, Y. Pommier, and P. D'Arpa (2002). Characterization of btbd1 and btbd2, two similar btb-domain-containing kelch-like proteins that interact with topoisomerase i. *BMC genomics 3*(1), 1.

Yanai, I., H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics (Oxford, England) 21*(5), 650–659.

Yu, P., D. Ma, and M. Xu (2005). Nested genes in the human genome. *Genomics 86*, 414–422.

Zuker, M. (2000). Calculating nucleic acid secondary structure. *Current Opinions in Structural Biology 10*(3), 303–310.