"ANALYSIS OF THE *ASPERGILLUS NIDULANS* TRANSCRIPTOME USING HIGH-THROUGHPUT RNA SEQUENCING"

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

Christopher John Sibthorp

September 2012

Abstract

The filamentous fungus, *Aspergillus nidulans* is a well-characterized model organism which has been used extensively for the study of eukaryotic cell biology and genetics over the past 60 years. The *A. nidulans* genome was sequenced in 2005, and various genome annotations have been released since, the majority of which rely heavily on *in silico* gene prediction. The development of high-throughput next generation sequencing technologies has revolutionised transcriptomics by allowing RNA-analysis of whole transcriptomes through massively parallel cDNA sequencing (RNA-seq). This sequencing approach has been applied to the *A. nidulans* transcriptome, and augmented by the development of a novel strategy for selectively sequencing the 5′ ends of RNAs on the ABI SOLiD platform. This aimed to produce a more robust resource for gene interrogation and the investigation of regulatory elements which impact on the transcriptomal landscape in *A. nidulans*. Bioinformatic analysis RNA-seq data was used to define 15,375 transcription start site (TSS) regions, which have been characterised by statistical analysis of mapped 5′ end distribution**.** Motif finding within sequence regions surrounding these TSS identified 16 putative functional promoter motifs based on overrepresentation and distributional analysis within promoters, and GO annotation found significant functional enrichment amongst genes associated with two of these motifs (AARARAAA and TTTYTTY). Transcript assembly of RNA-seq data has also revealed 16065 putative transcripts, 1112 of which were mapped to regions annotated as intergenic. From these transcripts we identified 38 strong candidates for novel protein coding genes (six of which contained non-canonical translation start sites), and over 400 additional transcripts containing putative coding regions. Separation of RNA-seq data in two sets of strand specific reads was shown to greatly increase the quality of transcript assembly and facilitated the identification of 2291 occurrences of sense:antisense overlap between assembled transcripts, four of which have been proven experimentally. Finally, assembled transcripts have been

used to detect multiple transcript isoforms arising from alternative splicing events. 374 distinct loci were identified as the origins of alternatively spliced transcripts, and six of these were verified experimentally.

List of contents

List of Figures

4.2. PCR cleanups with Minelute and Purelink kits.

4.3. Effect of PCR over-amplification on SOLiD libraries.

4.4. Protocol for 5′ specific RNA-seq on the SOLID platform.

4.5. Agilent 2100 bioanalyzer plot of 5′ specific RNA-seq library.

4.6. Correction of mis-annotation.

4.7. Conversion of mapped read alignments to RH frequencies**.**

4.8. Box-plots of RH peak distribution for 4 chromosomes of *A. nidulans*.

4.9. Readhead position distribution in TSS regions on chromosomes 1-4.

4.10. Examples of readhead mapping variance in transcription start sites.

4.11. Chromosome map of identified TSS regions**.**

4.12. Motifs identified from DS1.

4.13. Classes of motif arranged from strongest to weakest.

4.14. Histograms of enrichment relative to TSS location for 13 identified motifs.

4.15. Distribution of motif classes by kernel density estimates.

4.16. Motif classes ordered by level of distributional enrichment.

4.17. Combined strong motifs from both datasets.

4.18. 16 strongest candidates for functional motifs.

4.19. Sequential alignment of motifs to promoter sequences.

4.20. Sequence logo for the region surrounding transcription start sites in *A. nidulans*.

4.21. Comparison of 5′ specific RNA-seq readhead data from libraries produced with and without alkaline phosphatase (ALP) treatment.

4.22. Comparison of Random priming P2 oligonucleotides.

4.23. Agilent bioanalyser plots for samples prepared with the barcoded and non barcoded Random Priming P2 oligonucleotide.

5. New whole transcriptome libraries

5.1. Tophat based mapping of extremely long splice junctions displayed in IGV.

5.2. Box and whisker plot of intron lengths in *A. nidulans*.

5.3. Histogram of intron lengths in *A. nidulans.*

5.4. RNA-seq and junction data for gene AN4390.4 (CADANIAG00006067).

6. Identification of putative novel and antisense genes in *Aspergillus nidulans*

6.1. Cufflinks assembly of transcripts using aligned reads from both strands.

6.2. Strand specific transcript assembly with Cufflinks.

6.3. Novel antisense regions selected for experimental investigation.

6.4. Confirmation of an antisense transcript in *meaB*.

7. Investigation of alternative splicing in *Aspergillus nidulans*

7.1. Types of alternative splicing.

7.2. First example of alternative splicing by intron retention.

7.3. Second example of alternative splicing by intron retention.

7.4. Third example of alternative splicing by intron retention.

7.5. Example of alternative splicing utilising competing 5' splice sites.

7.6. Example of alternative splicing utilising a cassette exon.

List of Tables

5.2. Read quality assessment mapping.

5.3. Mapping and splice junction identification by Tophat with a maximum intron length of 5000 bp.

5.4. Mapping and splice junction identification by Tophat with alternative settings.

5.5. Nucleotide specific alignments to the A. nidulans genome and proportion of the genome transcribed

6. **Identification of putative novel and antisense genes in *Aspergillus nidulans***

6.1. Putative novel genes reported in the S1 SOLiD library using novel software.

List of abbreviations

3'. Three Prime

5'. Five Prime

ABI. Applied Biosystems

ALP. Alkaline Phosphatase

AspGD. Aspergillus Genome Database

ATP. Adenosine triphosphate

bp. Base Pair

CADRE. Central Aspergillus Resource

cDNA. Complimentary DNA

CGR. Liverpool Centre for Genomic Research

CM. Complete Media

DNA. Deoxyribonucleic Acid

dNTPs. Deoxynucleotide Triphosphates

DPE. Downstream Promoter Element

DTT. Dithiothreitol

EDTA. Ethylene Diamine Tetra Acetic Acid

ePCR. Emulsion Polymerase Chain Reaction

g. gravity

GO. Gene Ontology

IGV. Integrative Genomics Viewer

Inr. Initiator Element

Kb. Kilobase

mg. Milligram

ml. Millilitre

µg. Microgram

µl. Microlitre

MM. Minimal Media

mRNA. Messenger Ribonucleic Acid

ncRNA. Non-Coding Ribonucleic Acid

NAT. Natural Antisense Transcript

NGS. Next Generation Sequencing

nt. Nucleotide

ORF. Open Reading Frame

PCR. Polymerase Chain Reaction

PNK. Polynucleotide Kinase

PTGS. Post-transcriptional Gene Silencing

RH. Readhead

RNA. Ribonucleic Acid

RNAi. Ribonucleic Acid Interference

RNA-seq. Ribonucleic Acid Sequencing

RO. Reverse Osmosis

rpm. Revolution per Minute

rRNA. Ribosomal Ribonucleic Acid

siRNA. Short interfering Ribonucleic Acid

snRNP. Small Nuclear Robonucleoprotein

SOLiD. Sequencing by Oligonucleotide Ligation and Detection

SREK. Small RNA Expression Kit

sRNA. Small Ribonucleic Acid

TAP. Tobacco Acid Pyrophosphatase

TF. Transcription Factor

Tris. Tris (Hydroxymethyl) Aminomethane

TSS. Transcription Start Site

UTR. Untranslated Region

UV. Ultraviolet Light

WT. Wild Type

WTP. Whole Transcriptome Pipeline

Chapter 1

1. Introduction and aims

1.1. Importance of the Aspergilli

The Aspergilli are a group of approximately 200 species of fungi, with roles in disease and food spoilage (including 20 known to be human pathogens), as well as a number of beneficial species used in the biotechnology industry and for the production of foodstuffs (Timberlake & Marshall, 1989). As a result, the Aspergilli are one of the most important and widely studied groups in the field of fungal research, and the genomes of many species have been sequenced (Ward *et al*., 2006).

Aspergillus nidulans, is a filamentous fungus which has played a central role in eukaryotic cell research for over 60 years (Martinelli & Kinghorn, *1994*). During this time, *A. nidulans* has been developed as an experimental genetics system, and played a particularly important role in the study of mitosis, being utilised for the first studies targeted specifically at nuclear division (Morris, 1976a; Morris 1976b). Morris demonstrated the utility of the sexual and parasexual cycles of *A. nidulans*, and it's homothallic nature meant that mutants could be directly tested for complementation without the problems of genetic mating types. This was aided by the fact that haploid, diploid and heterokaryotic vegative states had all been defined, along with eight well-marked linkage groups representing the eight chromosomes (Pontecorvo *et al*., 1953; Kafer, 1977). These studies helped to establish *A. nidulans* as a model genetic system for future studies, including the investigation of gene organisation and regulation (Timberlake, 1980). It is important to note the existence of multiple wild-type stains of *A. nidulans* (Jinks *et al*., 1966), however all commonly used mutant strains are derived from a single strain selected as a genetic model in 1953, often referred to as the Glasgow strain (Pontecorvo *et al*., 1953).

As a model organism, developing our understanding of the *A. nidulans* genome can provide valuable insight into other *Aspergillus* species. This work therefore has the potential to aid research not only in *A. nidulans*, but other Aspergilli of social and economic importance, such as *A. oryzae,* used in the production of soy sauce, sake and miso, and *A. fumigatus*, a human pathogen which presents an ever increasing risk to immunocompromised patients due to mutations resulting in antifungal drug resistance (Chamilos & Kontoyiannis, 2006).

1.2. The *Aspergillus nidulans* genome

The first eukaryotic organism to have its genome sequenced was *Saccharomyces cerevisiae* in 1996 (Goffeau *et al*., 1996). It was not until 2005 that the *A. nidulans* genome was sequenced and published by the Broad Institute. This genome was found to contain approximately 30 million base pairs in eight chromosomes or linkage groups (Galagan *et al*., 2005). The genome sequence was annotated with the Calhoun annotation system, utilising protein homology searches and numerous gene prediction algorithms, including FGENESH (Salamov & Solovyev, 2000), FGENESH+, and GENEWISE (Birney *et al*., 2004). *A. nidulans* EST data was not incorporated into gene predictions, but was used separately for validation, and 9,541 protein-coding genes were predicted (Galagan *et al*., 2005).

The genomes of two other *Aspergillus* species (*A. fumigatus and A. oryzae)* were also sequenced in 2005, and described in two companion papers (Nierman *et al*., 2005; Machida *et al*., 2005). All three assembled genomic sequences were produced by different sequencing centres, each using a separate and diverse annotation process. This produced inconsistencies in gene model and functional annotation, while additional attributes such as Enzyme Commission numbers (Bairoch, 2000) or Gene Ontology (GO) associations (Ashburner *et al*., 2000) were also inconsistently applied. Functional annotations for *A. fumigatus* (Af293) and

A. niger (CBS 513.88) were manually reviewed by their respective research communities, prompting a reannotation of the *A. nidulans* genome by the Eurofungbase community, published in 2008 (Wortman *et al.*, 2008).The primary goal of this reannotation was to increase the number of *A. nidulans* proteins with accurate functional assignments. The percentage of *A. nidulans* gene products with informative names was increased from 3 % to approximately 19 %, which was subsequently increased to 58 % by the transfer of information from orthologous genes annotated in *A. fumigatus* or *A. niger*. Manual curation of gene model annotations also led to an increase in the number of predicted genes, from 9,541 to over 10,500 (Wortman *et al*, 2008). While this represented an iterative improvement over previous annotations, it was by no means a finished product, and further demonstrated the high level of inaccuracy associated with *in silico* gene prediction, which was still the primary method of annotation for the majority of *A. nidulans* genes.

The sequencing of the *A. nidulans* genome allowed for major advances in our understanding of the Aspergilli through comparative genetic studies between species. Comparative study of the *A. nidulans* genome with *A. fumigatus* and *A. oryzae* led to a major discovery, in that until that point *A. nidulans* was the only one of the three with a known sexual cycle, while the other two were only known to reproduce through asexual mitotic spores. In conjunction with other work (Nierman *et al*., 2005; Paoletti *et al*., 2005), 215 genes inplicated in the mating process, meiosis, fruiting body development and pheromone response in fungi were investigated and all (with the exception of mating-type genes) genes identified in *A. nidulans* were also found in both *A. fumigatus* and *A. oryzae*. While it is possible that these genes are present due to sexual reproduction being lost only recently in these species, these data suggested that *A. fumigatus* and *A. oryzae* may be capable of sexual reproduction, which may result in a massive increase in the potential for developing genetic tools for these fungi in both medicine and industry (Gallagan *et al*., 2005). This work also showed that both large

and small scale evolutionary processes in eukaryotes are species specific and can occur at different relative rates. These findings contrasted those in vertebrate, nematode and arthropod systems, where structural and nucleotide evolution are shown to be correlated (Galagan *et al*., 2005; Coghlan & Wolfe., 2002; Burt *et al*., 1999; Sharakhov *et al*., 2002).

Since this update to the *A. nidulans* genome, annotations have been constantly evolving as work continues to more accurately identify gene models and functions. Work is currently being undertaken by the group responsible for the Aspergillus Genome Database (AspGD) (Arnaud *et al*., 2010), with the aim of reannotating the *A. nidulans* genome using high-throughput sequencing data.

1.3. Transcription of mRNA in eukaryotes

Transcription is the first step of gene expression,in which the enzyme RNA polymerase is used to create a new molecule of RNA with a sequence complimentary to that of a region of DNA. The regulation of transcription is vital for the control of gene expression, and can occur at numerous points throughout the process. Firstly, for genes to be actively transcribed, the locus needs to be accessable to the machinery of transcription, which includes various enzymes and transcription factors. In eukaryotes, transcription is constitutively repressed by nucleosomes, in which a length of approximately 146 bp of DNA is wrapped around a complex of histones H3, H4, H2A and H2B (Struhl, 1999). Activation of transcription requires decondensation of the chromatin structure around the relevent locus (Chambeyron & Bickmore, 2004; Bernstein *et al*., 2004), which in turn is controlled by modification of histones in this region (Lachner *et al*., 2003) in response to a range of regulatory proteins (Fischle *et al*., 2003) including transcription factors, histone acetyltransferases and chromatin remodeling enzymes (Swanson *et al*., 2003). These modifications clear nucleosomes from the

promoter region of the gene to be transcribed (Boeger *et al.*, 2003, Reinke & Horz, 2003). There is also evidence for a relationship between regulation of transcription and the position of genes within chromosome territories, with highly expressed genes located more commonly at the periphery (Kurz *et al*., 1996; Dietzel *et al*., 1999) or within loops which extend outside of these territories (Volpi *et al*., 2000; Williams *et al*., 2002).

In eukaryotes, a complicated system of transcription factors is required for binding of RNA polymerase II to the promoter. These transcription factors were originally in identified in *Drosophila* and *S. cervisiae* (Matsui *et al*., 1980), are collectively known as general transcription factors and have been are named TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, where TF stands for transcription factor. These factors are recruited sequentially onto the DNA promoter to form a nucleoprotein complex capable of recruiting RNA polymerase II. This process is regulated by the action of activator and repressor proteins, which respectively aid or inhibit recruitment of the general transcription factors (Orphanides *et al*., 1996). This regulation has been shown to occur at cis-regulatory elements shortly upstream of the core promoter, which are often binding sites for other regulatory elements produced *in trans* elsewhere in the genome (Gilad *et al*., 2008).

Transcription itself is initiated by the binding of RNA polymerase to a promoter sequence within DNA. The core promoter consists of a short region immediately upstream and/or downstream of the transcription start site (TSS), and contains binding sites which interact directly with components of transcriptional machinery. The vast majority of core promoters studies have been performed on promoters containing a TATA box as an essential element. This conserved motif binds TFIID (lee & Young, 2000), however promoters have been shown to possess extreme structural and functional diversity, with many lacking the TATA

motif. This indicates that other factors along with TFIID are responsible for directing and binding the machinery of transcription to the core promoter (reviewed by; Smale & Kadonaga, 2003), and this diversity makes an important contribution to the regulation of gene expression (Butler & Kadonaga, 2002; Smale, 2001). Other core promoter elements capable of binding TFIID include the motif 10 element, the TFIIB recognition elements, the X core promoter element 1 and the initiator element (Inr), which can act alongside of independantly of a TATA motif. In TATA-less promoters, an alternative core promoter motif known as the downstream promoter element (DPE) is often (but not always) found downstream of the Inr, at a precise distance of +28-32 bp. TFIID has been shown to bind cooperatively to these motifs, with the efficiency of binding and therefore transcriptional activity being highly dependant on the distance between the Inr and DPE (Juven-Gershon & Kadonaga, 2010; Kutach & Kadonaga, 2000). This is supported by the work of Jin et al., 1995 and Tan & Richmond (1998), who showed that the insertion of a single base pair between transcription factor binding motifs can cause a rotational shift of up to 35°. Some transcription factors are able to tolerate a degree of flexibility in the distance between binding sites, showing only a lowering of binding affinity, while for others this rotational shift leads to steric incompatability and abolishing cooperative binding. This demonstrates the importance of conserved functional motifs and their position relative to the TSS and each other in the initiation and regulation of transcription.

Following the successful recruitment of RNA polymerase II to the promoter and initiation of transcription, the RNA polymerase travels along the template strand of the DNA from 3'-5', using complimentary base pairing to create an mRNA copy of the coding strand. This process is known as elongation, and ends in termination of transcription and the release of the fully formed pre-RNA, which must be processed to transform it into mature mRNA. Some of this

processing occurs during elongation, such as the addition of the 5' cap structure, which consists of a 7-methylguanosine moiety attached to the 5' end of the mRNA during early elongation (Takagi *et al*., 2002). This structure has several functions, including regulation of export from the nucleus, promotion of translation and protection of the mRNA molecule from degradation by 5'-3' exonucleases (Konarska *et al*., 1984; Caponigro & Parker, 1996). Another modification vital for the stability of mRNA is the addition of a poly(A) tail at the 3' end upon termination of transcription (Guhaniyogi & Brewer, 2001). Newly synthesised mRNAs are cleaved at the 3' end and the poly(A) tail added based on the presence of a polyadenylation signal motif recognised by a complex of cleavage and polyadenlyation factors (Hunt *et al*., 2008). Genes have also been identified with multiple polyadenylation sites, suggesting that a single gene may code for several transcripts with alternative 3' ends, contributing to diversity of the transcriptome (Proudfoot *et al*., 2002; Shen *et al*. 2008). Removal of the poly(A) tail and 5' cap are a vital processes in the control of mRNA expression through transcript degradation, and are discussed further in Chapter 4.16 of this thesis.

Splicing is another mechanism which can occur during elongation of the mRNA, and which also occurs after termination on transcription. This process assembles eukaryotic mRNAs from the longer, newly synthesised precursor mRNA. Alternative splicing patterns give rise to different protein isoforms with unique chemical and biological properties based on the inclusion or exclusion of coding sequences in the mature mRNA (Grabowski & Black, 2001). This facilitates a huge amount of diversity in the proteome, as a single transcript may have multiple splice patterns, with some known examples being alternatively spliced into thousands of different mature mRNAs (Black, 2000; Graveley, 2001). The mechanism of alternate splicing is explored in greater detail in Chapter 7.1 of this thesis.

1.4. Nitrogen metabolism in *Aspergillus nidulans*

Nitrogen is a vital resource for all organisms, used in the production of proteins, nucleic acids, lipids and in the fungal cell wall. *Aspergillus nidulans* has been extensively utilised in the study of nitrogen metabolism and regulation, alongside another fungal species, *Neurospora crassa* (Caddick *et al*., 1994). The processing of various nitrogen sources has been linked to genes involed in the nitrogen utilisation pathway of these two organisms, and this work has contributed greatly to the understanding of nitrogen metabolism regulation through gene expression (Wiame *et al*., 1985).

A. nidulans is able to utilise a wide range of nitrogen sources, although not all nitrogen-containing compounds are used equally. Generally, nitrogen sources can be classified into two subgroups, distinguished on the basis of preference for their utilisation: Primary nitrogen sources (ammonia and glutamine) and secondary nitrogen sources, which includes all other compounds which can be utilised. When they are available, the uptake and metabolism of primary nitrogen sources is preferred to secondary sources, the processing of which generally requires additional cellular mechanisms. These mechanisms are tightly regulated by a number of genes associated with each metabolic pathway (Caddick *et al*., 1994), and primarily controlled by expression of the GATA transcription factor AreA (Platt *et al*., 1996). AreA is a positive regulator of secondary nitrogen metabolism, and the expression of *areA* mRNA is repressed in the presence of an abundant primary nitrogen source, effectively blocking the utilisation of secondary nitrogren sources in a process known as nitrogen metabolite repression (Arst & Cove, 1973; Wiame *et al*., 1995). In contrast, growth on secondary nitrogen sources has been shown to result in higher levels of the transcript and the effective utilisation of secondary nitrogen sources (Fraser *et al*, 2001). A second transcription factor, MeaB, has also been implicated in the regulation of nitrogen metabolism. MeaB appears to

act in opposition to AreA, being optimally expressed in response to nitrogen sufficiency and rapidly localized to the nuclei in the presence of primary nitrogen sources, with this being reversed on transfer to nitrogen limiting conditions. While MeaB has been proven capable of independantly mediating nitrogen metabolite repression, the expression of many genes regulated by MeaB are dependant on AreA, and that *areA* is able to partially regulate m*eaB* (Wagner *et al*., 2010), however the method of this regulation was not clear.

1.5. The development of high-throughput sequencing

The recent development of "next generation" high-throughput sequencing using next generation technologies such as the ABI SOLiD and Illumina Genome Analyzer II has completely transformed quantitative transcriptomics. These next generation sequencing (NGS) technologies have significantly reduced the cost-per-reaction and time required, while simultaneously increasing the amount of data produced by each experiment. This has led to a number of experimental techniques being developed to directly survey the RNA content of cells without the traditional cloning requirements of EST sequencing. Such approaches have been shown to generate quantitative expression scores that are comparable to microarrays, but allow the entire transcriptome to be surveyed with no prior knowledge of transcribed regions (Wilhelm & Landry, 2009). This provides an unprecedented potential for global transcriptome analysis and the discovery of previously unidentified transcripts.

As high-throughput sequencing became more prevalent, the data produced has provided information on non-coding transcripts, untranslated regions and gene structures, facilitating iterative improvements to existing genome annotations. Investigation of sequence reads spanning exon-intron or exon-exon junctions has also given unique insights into splicing

across introns and genes, as well as changes in splicing observed in alternative growth conditions (Wilhelm *et al*., 2008).

Cloonan *et al.,* (2008), showed that high-throughput sequencing can be used to survey the complexity, dynamics and sequence content of transcriptomes in a near-complete fashion. Using a combination of tag maps with existing genome and transcriptome annotations, they developed a protocol to study the genomic landscape of gene expression and both known and novel alternative splicing events. This clearly defined the location of exons, as well as addressing transcriptional activity, transcriptional complexity from active loci, transcriptional activity outside currently defined annotations, and allowing for the discovery of previously unidentified nonsynonymous single-nucleotide polymorphism (SNP) expression (Cloonan *et al*., 2008). This work also showed the use of tag counts to define differential expression between various states, finding good concordance with previously published expression profiles (Bruce *et al.,* 2007) and work by Wilhelm *et al*., (2008) who have also shown a strong correlation between RNA expression levels determined from sequence-read numbers and those determined from hybridization signals. This indicates that data obtained from high throughput sequencing provides quantitative analysis of transcript levels and may do so more accurately than previous techniques. Expression levels measured by Lyne *et al*., (2003) using cDNA microarrays produced measurable signals from only 80-90% of genes in proliferating cells, with the remaining 10-20% shown to be highly expressed only under specific conditions such as meiosis or response to stress (Chen *et al*., 2003; Mata & Bähler, 2003). In comparison, expression levels in proliferating cells measured by analysis of sequencing data showed an average coverage of 94.9 %, and Lyne *et al.,* (2003) suggest that the sequencing approach is sensitive enough to detect genes with low expression levels and even basal "transcriptional noise" from genes which are not actively expressed.

This unprecedented level of sensitivity and accuracy means that whole transcriptome

sequencing using high-throughput NGS technology offers a near-complete snapshop of the transcriptome, including rare and hard to detect transcripts with regulatory roles. High-throughput sequencing has therefore become the primary method for the discovery and characterization of non-coding RNAs (ncRNA). ncRNA describes a broad class of functional RNAs which are not translated into proteins. Combined with the fact that ncRNAs are often poorly conserved over evolutionary time, this made identification and sequencing of ncRNAs extremely difficult using previous technologies. Therefore, the prediction of both mature and precursor ncRNAs using *in silico* methods is of limited use, and the majority of ncRNA discovery relies on the sequencing of small RNA fragments (Mardis, 2008). Next generation high-throughput sequencing is therefore ideally suited to the global discovery of ncRNAs, and has been utilized for this purpose in a number of species ranging from *Chlamydomonas* (Zhao *et al*., 2007) to human and chimpanzee brain (Berezikov *et al*., 2006).

1.6. The challenges of high-throughput RNA-seq

While high-throughput RNA-seq has been shown to have many advantages over traditional, low-throughput EST sequencing, there are a number of complications associated with these experiments. Sequence reads produced by RNA-seq using next generation sequencing (NGS) technologies are predominantly very short. For example, reads produced by SOLiD sequencing at the start of this thesis were only 35 bp in length, and while this has been improved over time, current iterations of the SOLiD system are still limited to the production of 75 bp fragment libraries.

To achieve the short read length which can be sequenced using NGS technology, transcripts must be fragmented prior to sequencing, then reassembled from the resulting fragments (with the exception of various classes of small RNAs which are shorter than the sequenced length).

This leads to loss of information as there is no definitive way of determining the structure of the full length transcripts, which must be predicted by assembly of reads mapped to a target genome. Transcriptome assembly is extremely challenging to perform accurately as algorithms must take into account factors such as strand specificity, read and mapping quality, and the fact that transcript variants from a single gene can share much of the same exon sequence, making them difficult to resolve. Using the SOLiD NGS sequencer we have observed an additional complication in that the distribution of read mapping can be highly uneven across a transcribed sequence, and sufficient depth of sequencing must therefore be obtained to ensure complete coverage of transcribed regions and robust transcript assembly. This issue has also been reported in other NGS technologies (Hansen *et al*., 2010), so does not seem to be limited to the SOLiD, indicating the need for further development and improvement of NGS technology across all platforms.

The choice of whether to guide transcript assembly is also important, as this is largely dependent on the quality of the reference genome. In the case of *A. nidulans* supplying a reference was shown to greatly reduce the number of novel transcripts identified compared *de novo* transcript assembly. Previous studies such as the sequencing of the *Anopheles funestus* transcriptome by Crawford *et al.* (2010) have combined these two methods where the quality of reference genome was called into question, and findings from this thesis propose that a similar approach may be beneficial in *A. nidulans* (Chapter 5.7). Even when an assembly strategy has been developed, the choice of algorithm poses yet more challenges, as most transcript assembly software is developed using distinct organisms on a single NGS platform, so different algorithms can produce a variety of transcript assemblies from the same data. In some cases the choice of assembler may also be limited**,** as specific data formats (such as SOLiD colour space) are not accepted by all assemblers, and even where they can be used, support for these data types is not always robust. These problems are compounded by the fact

that there is currently no standard criteria for quality assessment of transcriptome assemblies, although the need for such criteria has been acknowledged and they are currently being developed (Salzberg *et al.,* 2005; Meader *et al*., 2010).

The result of high-throughput coupled with short read length is extremely large datasets, ranging from several gigabases to terabases of data produced from a single transcriptome. Manual data-handling is therefore impossible, and researchers must develop new skills in bioinformatics and the production of novel software to facilitate automated manipulation and interrogation of these massive datasets. Hardware limitations are also a consideration, as analysis of these data requires high-power computing systems with large memories and sufficient processing capability to run parallel algorithms.

During the course of this thesis, advances have been made in NGS technology and data analysis. Throughput, read length, mapping and assembly algorithms have all been improved, resulting in greater read coverage and higher quality transcript assembly. However, this technology is still reliant on cDNA synthesis or hybridization steps, the efficiency of which is dependent on RNA sequence and structure. This can result in spurious second-strand cDNAs through template switching (Cocquet *et al*., 2006) and primer-independent cDNA synthesis due to self-priming thought to occur through cDNA secondary structure (Ozsolak & Milos, 2011). Furthermore, reverse transcriptases have been shown to have lower fidelity than other polymerases due to their lack of proofreading mechanisms, and can add nontemplated nucleotides to cDNAs (Chen & Patton, 2001). These factors together pose limitations on the quality and accuracy of strand-specific RNA-seq using standard NGS techniques.

Current work in the field of NGS aims to address many of the issues through development of amplification-free low quantity RNA-seq (LQ-RNAseq) (Ozsolak *et al.,* 2010) and direct sequencing of RNA molecules (Lipson *et al*., 2009).

1.7. The ABI SOLiD system

Early comparisons of high-throughput NGS technologies such as those conducted by Harismendy *et al*. (2009) suggested that platforms such as the ABI SOLiD and Illumina sequencers were extremely similar in terms of accuracy and throughput. This work suggested that the major factors influencing choice of sequencing platform were therefore time, cost and availability. However, each system was shown to have minor advantages in certain applications. The ABI SOLiD system was shown to produce lower coverage variability and therefore higher accuracy at low coverage. This was of particular importance for the identification and characterization of transcripts with very low expression, and of rare intergenic transcripts such as ncRNAs (Harismendy *et al*., 2009). Combined with the in-house availability of SOLiD sequencing, this made it the system of choice for this analysis.

The applied Biosystems (ABI) SOLiD (Sequencing by Oligo Ligation and Detection) system was commercially released in 2007. This system offers a unique sequencing methodology based on sequential ligation of fluorescently labeled semi-degenerate oligonucleotide probes. Each probe queries two adjacent base positions at a time, with four fluorescent dyes used to encode for the sixteen possible di-base combinations. After the ligation step, a fluorescent readout records the colour of the dye. The fluorescent group is then removed from the ligated oligonucleotide probe by chemical cleavage, allowing a subsequent round of ligation. The use of di-base probes is referred to as "2 base encoding" and contributes to the high accuracy of the SOLiD system. Cycling of ligation, detection and cleavage determines the eventual read length. Following a series of ligation cycles, the extension product is removed and the template reset with a primer complimentary to the n-1 position for a 2^{nd} round of ligation cycles. A total of 5 cycles are performed for each sequence tag and the final sequence produced in colour space.

This method of sequencing provides numerous advantages compared to other systems:

1. Interrogation of two bases in each ligation provides increased specificity.
2. Each base in interrogated twice providing increased confidence in each call.
3. The primer is reset for five independent rounds of extension, improving signal to noise ratios.
4. The design of the four dyes encoding sixteen possible two base combinations enables built in error checking.

The sequencing process and colour space format are detailed in Fig 1.1. An overview of the SOLiD workflow is presented at:

http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/

flash/102207/solid.html

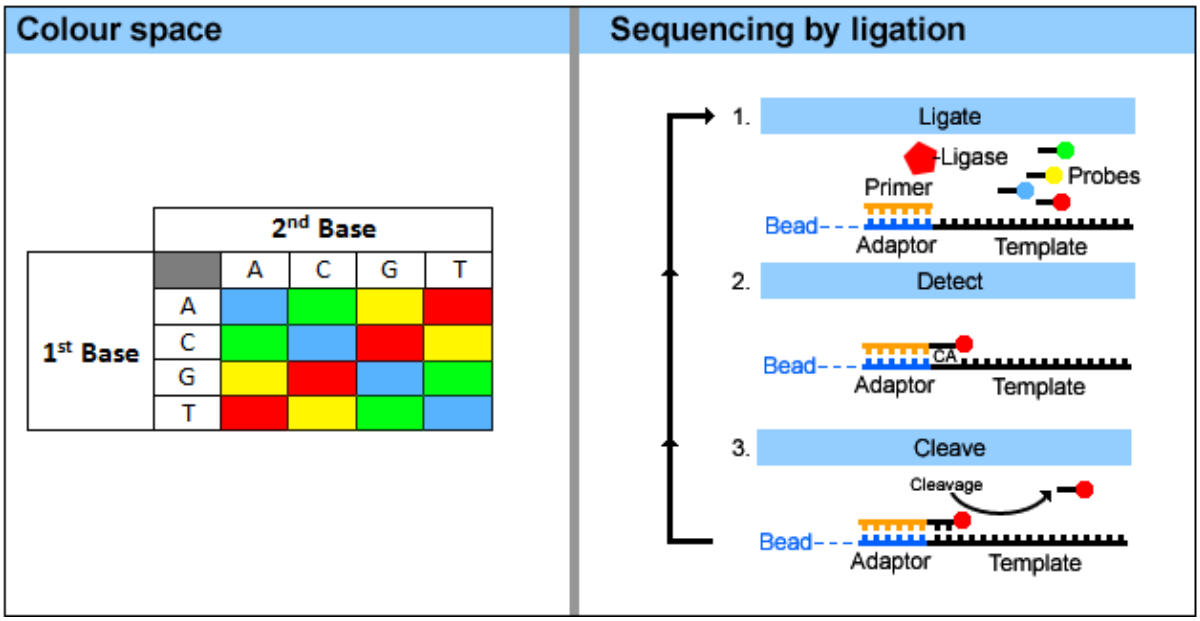


Figure 1.1. The colour space format and SOLiD sequencing by ligation. The SOLiD system uses a set of semi-degenerate and fluorescently labelled di-base probes to create sequences in colour space. A key for converting di-base sequences to colour space is illustrated in the left pannel. The SOLiD sequencing process is shown in the right panel, and begins with hybridization of primers to the P1 adapter sequence within the library template. The four probes then compete for ligation to the sequencing primer. The fluorescent label attached to the ligated probe is detected by the SOLiD, the probe is then removed by cleavage and the process repeated for the next base in the template. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction.

RNA samples can be prepared for sequencing on the ABI SOLiD platform in a number of different ways, depending on the application and desired outcome. SOLiD sequencing of the whole transcriptome utilizes fragment libraries created from mRNAs, which must be isolated from total RNA by poly(A) selection or ribosomal RNA depletion. Prior to fragmention with RNase III, the mRNA library is treated with tobacco acid pyrophosphatase (TAP) to cleave the cap structure present at the 5' end of the mRNA molecules, leaving a 5'-monophosphate and a 3'-hydroxl. These structures are characteristic of RNase III cleavage products and are required to make the mRNA competent for adaptor ligation by RNA ligase which selectively ligates synthetic oligoribonucleotides to the phosphate at the 5' end. The adaptors used are sets of RNA/DNA oligonucleotides with a single-stranded degenerate sequence at one end and a defined sequence required for SOLiD sequencing at the other. Preparation of libraries for SOLiD sequencing is shown in Fig 1.2.

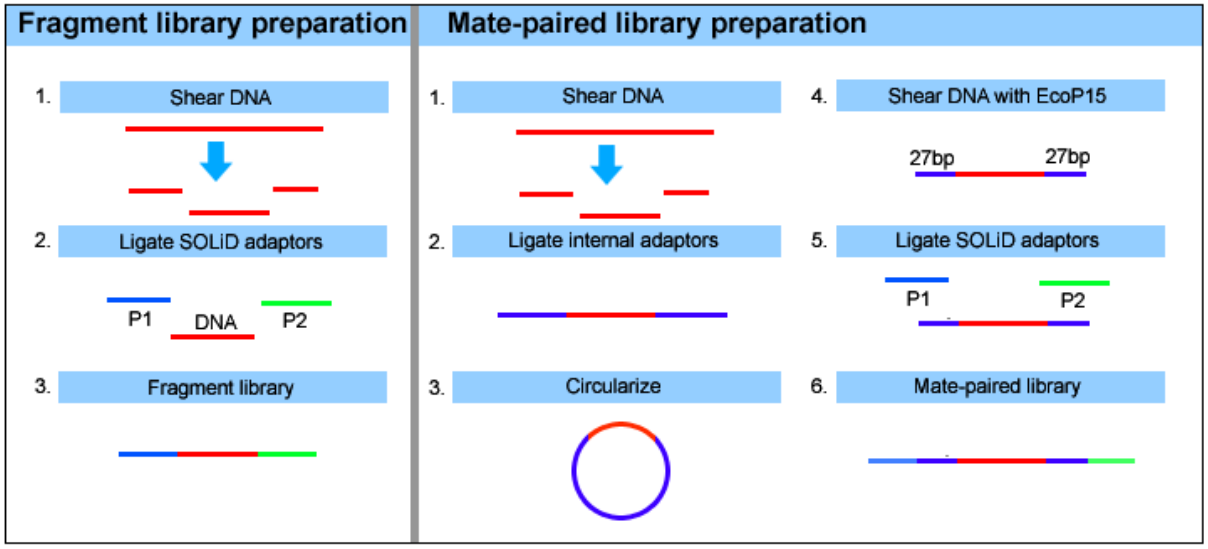


Figure 1.2. Creation of SOLiD fragment libraries. The SOLiD system can use two types of libraries, fragmented (left panel) or mate-paired (right panel). The fragment libraries utilized in this thesis are prepared by fragmenting mRNAs with RNase III, followed by ligation of the SOLiD 5' (P1) and 3' (P2) adaptors to each fragment.

Following the ligation of the adaptor to the fragmented RNA, reverse transcription is performed to synthesise a cDNA library of fragments containing the P1 and P2 adaptors required for sequencing. The RNA/cDNA duplex is then denatured and the library is selected by size for fragments with sequence inserts of sufficient length to generate full-length SOLiD RNA-seq reads. The products of size-selection are then amplified by PCR to increase quantity of fragments, using the lowest possible number of cycles to maintain an accurate expression profile. PCR products are then purified and subjected to quality control (QC) checks, ensuring that the library is of sufficient quality and concentration. Libraries which pass QC are prepared for SOLiD sequencing by emulsion PCR (ePCR) of the template with P1 coupled beads. In this process, the template is annealed to the P1-coupled beads, and a polymerase extends from the P1 adapter to generate a complimentary sequence extending from the beads surface. Following ePCR, templates are denatured and bead enrichment performed using a glycerol gradient to separate beads with extended templates from the undesirable non-templated beads. Templated beads are 3' modified to facilitate covalent linkage to the SOLiD slide, where they are deposited in a random array (Fig. 1.3). Deposition chambers offer the ability to segment the slide into multiple chambers during the loading process, allowing increased densities of beads per slide and resulting in a higher level of throughput.

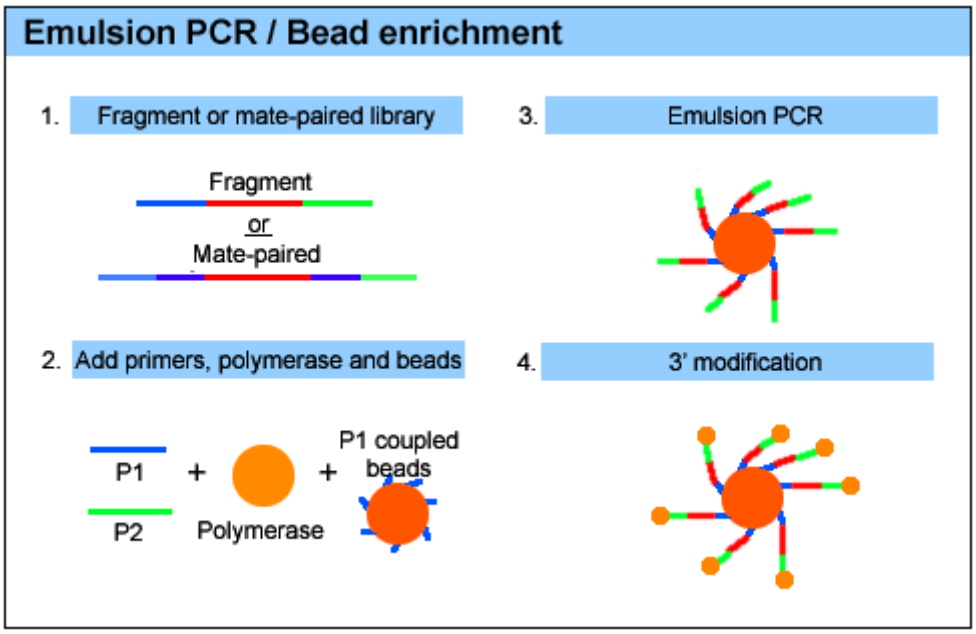


Figure 1.3. **Emulsion PCR and bead enrichment for SOLiD sequencing**. Emulsion PCR is used to capture the templates on P1-coupled beads. The templates are then denatured and bead enrichment is performed to separate beads with extended templates from undesirable beads. Templated beads are then 3' modified to facilitate covalent linkage to the SOLiD slide,

1.8. Aims

The aim of this thesis was to develop modern, high-throughput strategies to investigate the *A. nidulans* transcriptome. These strategies aimed to identify key sequence elements within the genome which may be used to predict and/or regulate gene expression. This included *in silico* prediction of promoters as well as the identification of specific DNA or RNA motifs associated with regulatory proteins.

Through interrogation of high-throughput sequencing data, this thesis also aimed to assess the use of high-throughput sequencing for the identification and characterization of novel genes, intergenic non-coding RNAs, antisense genes, and to provide new insights into the integration of different biological processes. While the aim was not to reannotate the *A. nidulans* genome, a priority was placed on collaboration with CADRE and AspGD to make the resultant high-throughput sequencing data publicly available as a community resource to assist with gene interrogation and genome reannotation.

Chapter 2

2. Materials and Methods

2.1. Synthetic oligonucleotides

Table 2.1. List of custom oligonucleotide names and sequences.

Name	Sequence (5′ - 3′)
P1 (comp)	Phosphate- ATCACCGACTGCCCATAGAGAGGAAAGCGGAG GCGTAGTGGTT -biotin TEG
P1 random (6)	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT NNNNNN -phosphate
P1 random (10)	CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT NNNNNNNNNN -phosphate
Random priming P2 (barcoded)	CTGCCCCGGGTTCCTCATTCTCTAAGCCCCTGCTGTACGGC CAAGGCGNNNNNC
Random priming P2 (non-barcoded)	CTGCTGTACGGCCAAGGCGNNNNNC
P1 PCR	CCACTACGCCTCCGCTTTCCTCTCTATG
P2 PCR	CTGCCCCGGGTTCCTCATTCT
Antisense A (AN8048.4)	GTCCGCCAGATATACTA
Antisense B (AN8040.4)	GAACTGGCTCTTAATG
Antisense C (AN4023.4)	ATCACCGAACTGAGACT
Antisense D (AN4058)	CCACCGTATATCATCAG
Splicing A forward 1	GAGCATCGCTACGCTGTT
Splicing A forward 2	CACCTCAAGGAAGCTACA
Splicing A reverse	CGATACGACGCCTCTTCT
Splicing B forward	TAGAGGTCGGGAGTGATG
Splicing B reverse	TTTGTAGAAGGCGGGCTC
Splicing C forward	CAGCCAGAGGAGATCAAG
Splicing C reverse	GGGGACTCTGTCGAATCT
Splicing D forward	AAGGTCGTCGGTTAAGCG
Splicing D reverse	AACCCGGTTCTTTCTCCG
Splicing E forward	GCCAGATATATTATCGGC

Splicing E reverse	TGCTAGAACGATTTGACC
meaB F1	ATCTTGCGTCAAGACCT
meaB F2	CTTCCGCTTCTGTGTCT
Meab R	CCGGTTCTCTAACTGTC

2.2. *Aspergillus nidulans* strains and maintenance.

2.2.1. *Aspergillus nidulans* strains

Aspergillus nidulans wild-type strain is designated G00 in the Glasgow collection and is a natural isolate and therefore has no explicit genotype (Pontecorvo *et al*., 1953).

2.2.2. Growth media

Growth media were as described by Cove (1966) and are listed in Appendix 1.2.1.

2.2.3. *Aspergillus nidulans* solutions and media

For a list of solutions and media used throughout this thesis, as well as their compositions. See Appendix 1.2.1.

2.2.4. Maintenance of *Aspergillus nidulans* cultures

A. nidulans stock cultures were kept as conidia at -80 °C. Strains were grown on minimal media (MM) containing 3 % agar (w/v), with appropriate supplements for 2-3 days at 37 °C. Conidial suspensions were created by scraping conidia from confluent plates and resuspending in 20 ml of 0.1% Tween. The approximate concentrations of the resultant suspensions were determined using a hemocytometer, and 10^8-10^9 spores used to inoculate 1 l flasks containing appropriate media and supplements for each growth condition described in Table 2.2. Incubations were at 37 °C in an orbital incubator at 200 rpm. Mycelia were harvested by filtration through Miracloth [Calbiochem Corp.], washed with fresh media, then

washed again with RO filtered water and dried by blotting on paper towels. Dried mycelia were snap frozen in liquid nitrogen and stored at -80 °C.

Table 2.2. Large scale cultures. This table describes the growth conditions for large scale liquid *A. nidulans* cultures used throughout this thesis.

Growth condition	Media and Supplements	Incubation
MM + nitrate	250 ml MM + sodium nitrate (final concentration 10 mM NO_3^-)	16 hours
MM + ammonium	250 ml MM + Ammonium D tartrate (final concentration 10 mM NH_4^+)	16 hours
Complete medium	250 ml CM	16 hours
4 hour nitrogen starvation	250 ml MM + sodium nitrate (final concentration 10 mM NO_3^-)then transfer to 250 ml MM	16 hours then transfer to unsupplemented MM and incubate for a further 4 hours
72 hour nitrogen starvation	250 ml MM + sodium nitrate (final concentration 10 mM NO_3^-) then transfer to 250 ml MM	16 hours then transfer to unsupplemented MM and incubate for a further 72 hours
Oat	1 % ground oat in 500 ml water	72 hours

2.3. Molecular techniques for the manipulation of nucleic acids

2.3.1. Extraction of RNA from *Aspergillus nidulans*

Mycelia harvested from liquid culture and stored at -80 °C was transferred to liquid nitrogen. Mycelia were ground to fine powder using a pestle and mortar with the addition of liquid nitrogen to maintain a fully frozen state. Approximately 1 g of dry, powdered mycelium was transferred to a 30 ml centrifuge tube containing 1 ml phenol (pH 5.0) and 1.5 ml lysis buffer (100 mM Tris-HCL pH 8.0, 300 mM NaCl, 10 mM EDTA, 3 % SDS). Tubes were vortexed

to mix and centrifuged at 4500 x g for 15 minutes. 0.75 ml supernatant was transferred into a fresh 1.5 ml Eppendorf tube, to which was added 0.5 ml phenol (pH 5.0). Tubes were vortexed prior to centrifugation for 15 minutes at 15,000 x g. Approximately 0.5 ml of supernatant was collected and transferred to a fresh 1.5 ml Eppendorf tube. An equal volume of 5 M LiCl was deed and tubes left overnight at 4 °C. Precipitated RNA was pelleted by centrifugation at 4 °C for 60 minutes at 15,000 x g. Pelleted RNA was washed twice with 180 µl 70 % ethanol, dried for 10 minutes in a fume cupboard to prevent contamination, then dissolved in 400 µl SDW. Ethanol precipitation was performed by the addition of 40 µl 3M sodium acetate and 1 ml 70 % ethanol, then incubating for at least 1 hour at -20 °C. Precipitated RNA was pelleted by centrifugation for 15 minutes at 15,000 x g. The supernatant was removed and pellets washed with 180 µl 70 % ethanol, then re-centrifuged for 5 minutes at 13,000 x g to ensure retention of the pellet. Pellets were dried for 10 minutes in a fume cupboard, then dissolved in 50 µl RNA storage solution (5 mM EDTA, 5 % SDS) and stored at -20 °C, or 50 µl nuclease free water for immediate use.

2.3.2. Nucleic acid quantification

Concentrations of DNA and RNA solutions were measured by Nanodrop [Thermo Scientific] spectrophotometer, using 2 µl of sample per measurement. Where sample concentration was expected to be greater than 5 µg/ml samples were diluted by a factor of 10 to achieve greater accuracy.

2.3.3. Agarose gel electrophoresis of DNA

Agarose gel electrophoresis of DNA was performed using horizontal electrophoresis gel tanks [Fisher Scientific] with 1 X TAE (0.4 M Tris-acetate; 1 mM EDTA) buffer. Gels were composed of 1.0 to 2.0 % (w/v) agarose in 1 X TAE buffer. Where required, 1 in 5 volumes of loading buffer was added to DNA samples prior to loading of the gel. Gels were run at 50 to 200 V depending on the size of apparatus and required resolution of DNA size bands, until the bromophenol dye had travelled 3/4 of the gel length. DNA was stained by the addition of ethidium bromide to the molten agarose gel at a final concentration of 0.5 µg/ml. Stained DNA was visualised under ultraviolet (UV) light.

2.3.4. Ethanol precipitation

Nucleic acids were purified and concentrated from solution by ethanol precipitation in the presence of sodium ions. To improve the efficiency of nucleic acid recovery, DNA or RNA samples under < 100 µl were diluted to 100 µl volume using nuclease-free water. Sodium ions were introduced by the addition of 1/10 volumes of sodium acetate solution (3M, pH 5.2), followed by 3 volumes of 100% ethanol. Precipitation reactions were incubated for at least 2 hours at -20 °C. Samples containing small nucleic acid fragments were incubated overnight to give the highest possible yield (Zeugin and Hartley, 1985). Precipitate was collected by centrifugation at 14,000 x g at 4 °C for 30 minutes. The supernatant was discarded and the pellet washed with 200 µl of 70 % ethanol. A second centrifugation at 14,000 x g for 5 minutes ensured that the pellet was at the bottom of the tube before the supernatant was discarded. Pellets were dried for 10 minutes in a laminar flow cabinet to prevent contamination. Dried pellets were dissolved in the appropriate volume of nuclease free water or the desired buffer for the next reaction.

2.3.5. Phenol:chloroform extraction of nucleic acids

Phenol:chloroform extraction was used to isolated polynucleotides from aqueous solutions containing proteins. This method relies on phase separation by centrifugation of a mix of the aqueous sample and a solution containing water-saturated phenol, chloroform and a chaotropic denaturing solution (guanidinium thiocyanate) (Chomczynski & Sacchi, 1987).

Equal volumes of phenol:chloroform (1:1; v/v) and an aqueous solution containing polynucleotides to be isolated, were mixed vigorously by vortexing for 20 seconds. Phase separation was achieved by centrifugation at 13,000 x g for 5 minutes. This resulted in an upper aqueous phase (containing polynucleotides) and a lower phase (containing proteins dissolved in phenol and lipids dissolved in chloroform). The aqueous phase was extracted and ethanol precipitation performed to purify the polynucleotides for downstream applications.

2.3.6. Poly(A) selection using oligo(dT)

Enrichment for mRNA from total RNA samples was performed by Poly(A) selection using an Oligotex direct mRNA midi kit (QIAGEN) following the manufacturer's protocols for purification of mRNA from samples containing starting material of ~500 µg total RNA.

2.3.7. DNA depletion with DNase I

DNase I is an endonuclease that nonspecifically cleaves DNA to release di-, tri- and oligonucleotide products (Kunitz, 1950; Vanecko and Laskowski, 1961). This enzyme was used for the depletion of contaminating DNA in total RNA samples extracted from cell cultures, with RNase inhibitor added to the reaction mix to protect the total RNA from

degradation. The following components were combined and incubated at 37 °C for 1 hour: 2 µl 10X DNase I buffer (Invitrogen), 0.5 µl RNasein RNase inhibitor (Promega), 0.5 µl DNase I (Invitrogen) and 17 µl Total RNA sample.

Reactions were stopped by phenol:chloroform extraction and ethanol precipitation.

2.3.8. 5′ decapping of RNA with tobacco acid pyrophosphatase

The 5′ cap structure was removed from mRNAs by treatment with Tobacco Acid Pyrophosphatase (TAP). The following components were combined and incubated at 37 °C for 1 hour: 2 µl 10X TAP buffer (Epicentre), 0.5 µl RNasein RNase inhibitor (Promega), 15.75 µl RNA in nuclease free water and 0.25 µl TAP (10 u/µl) (Epicentre).

Reactions were stopped with phenol:chloroform extraction and ethanol precipitation of the RNA.

2.3.9. RNA fragmentation with zinc acetate

Poly(A) selected RNA was fragmented using the RNA Fragmentation Reagents kit (Ambion) to achieve fragment sizes of approximately 50-200 nt. This kit uses a form of heat fragmentation catalyzed by metal ions in the form of Zn^{2+}. 0.9 µl of 10X fragmentation reagent (zinc acetate solution) was added to 300 µg of poly(A) selected RNA dissolved in 8 µl of nuclease-free water and incubated for 5 minutes at 70 °C. The reaction was terminated by the addition of 0.9 µl of stop solution (metal chelating agent) and chilling the solution on ice. The stopped reaction mix was diluted to 500 µl with nuclease-free water and purified by either ethanol precipitation or ultra-filtration using a YM-30 column (Millipore).

2.3.10. End repair with T4 polynucleotide kinase

To facilitate SOLiD library preparation, T4 polynucleotide kinase (T4 PNK) end repair was conducted for RNA fragments produced by chemical shearing. The dual functionality of T4 PNK allows it to act as both a 5′-kinase and 3′phosphatase, producing both the 5′ phosphate group and 3′ hydroxyl group required for SOLiD adaptor ligation. The following reaction mix was incubated for 10 minutes at 37 °C: 12.5 µl RNA sample in nuclease free water, 2 µl 10X T4 PNK buffer (Fermentas), 0.5 µl RNasein RNase inhibitor (Promega), 3 µl T4 PNK (10 u/µl) (Fermentas) and 2 µl 5 mM ATP.

Reactions were stopped by the addition of 4 µl of 0.5 M EDTA. RNA was purified by phenol-chloroform extraction and isolated by ethanol precipitation. The resultant pellet was dissolved in 8 µl of nuclease-free water for use in SOLiD library preparation.

2.3.11. RNase H degradation of RNA in RNA/DNA duplexes – manufacturer's protocol

RNase H (Fermentas) was used to selectively degrade the RNA strand in the RNA/cDNA duplexes created by first strand cDNA synthesis (Vorobjev & Zarytova, 2000) in modified SOLiD library preparation protocols.

RNA/cDNA hybrids were purified from first strand synthesis reactions by ethanol precipitation and pellets dissolved in 8 µl of nuclease free water. The following reagents were mixed and incubated at 37 °C for 20 minutes: 8 µl RNA/cDNA hybrid solution, 1 µl 10X RNase H reaction buffer (Fermentas) and 1 µl RNase H (Fermentas),

Reactions were stopped by heat inactivation of the enzyme at 65 °C for 10 minutes.

2.3.12. RNase H degradation of single stranded RNA in RNA/DNA duplex - modified protocol

1 µl of RNase H (Fermentas) was added to 10 µl of first strand synthesis solution containing RNA/DNA hybrids and the reaction mix incubated at 37 °C for 20 minutes. Stopping of the reaction was not required for downstream applications.

2.3.13. First strand cDNA synthesis with random priming P2

Hybridisation of the degenerate N6 sequence of random primer P2 was performed by combining the following reagents and incubating the reaction at 65 °C for 5 minutes then snap cooling on ice for 2 minutes: 10 µl (~1 µg) Fragmented RNA sample, 5 µl Random primer P2 (10 nm/µ1) and 2 µl dNTPs (10 µM),

First strand synthesis was performed using Superscript III reverse transcriptase (Invitrogen). An RT master mix was created with the reagents in Table 2.3. 23 µ1 of RT master mix was added to each hybridisation reaction and RT incubation performed in a thermal cycler as shown in Table 2.4.

Table 2.3. Reaction components for reverse transcription.

Volume	Component
9 µl	Nuclease-free water
8 µl	5X First Strand buffer
2 µl	DTT (0.1 M)
2 µl	RNasin
2 µl	SuperScript II Reverse Transcriptase

Table 2.4. Incubation temperatures and times for reverse transcription.

Temperature	Time
12°C	1 hour
25°C	5 min
42°C	10 min
50°C	30 min
55°C	30 min
70°C	15 min
4°C	hold

2.3.14. P1 adaptor ligation

Ligation of the P1 adaptor to cDNA was performed using DNA ligase (Promega). The following reagents were combined and incubated for 16 hours at 16 °C in a thermal cycler: 7 µl cDNA, 2 µl Adaptor N6 **or** N10 (10 nm/µ1), 10 µl 2X ligation buffer and 1 µl DNA ligase (Promega).

2.3.15. Second strand synthesis with proofreading TAQ

Second strand synthesis of cDNA ligated to the P1 adaptor using KOD hot start DNA polymerase (Novagen). The components in Table 2.5 were combined and incubated at 95 °C for 4 minutes to activate the enyzme.

Table 2.5. Reaction components for second strand synthesis.

Volume	Component
6 µl	$MgSO_4$ (25 mM)
10 µl	DNTPs (2 mM)
10 µl	10 X buffer
2 µl	KOD hot start polymerase
52 µl	Nuclease-free water

The 20 µl P1 adaptor ligation mix was then added and the reaction incubated at 70 °C for 10 minutes in a thermal cycler.

2.3.16. Chemically fragmented library amplification with primers P1 and P2

PCR amplification of libraries was conducted using KOD hot start polymerase (Novagen).

KOD hot start polymerase is a premixed complex of KOD DNA Polymerase and two monoclonal antibodies that inhibit the DNA polymerase and 3′→ 5′ exonuclease activities at ambient temperatures (Mizuguchi *et al*., 1999). This helped to avoid non-specific amplification due to mispriming events that can occur during setup and initial temperature increase. Primer degradation during setup at room temperature due to exonuclease activity is also effectively inhibited.

PCRs were set up as shown in Table 2.6. Reactions were incubated at 95 °C for 5 minutes to activate the KOD hot start polymerase, then subjected to 10-18 cycles of PCR as described in Table 2.7. A final extension step was conducted for 5 minutes at 72 °C to promote complete synthesis of all PCR products (Sullivan *et al*., 2006).

Table 2.6. Reaction components for PCR amplification of chemically fragmented libraries.

Volume	Component
10 µl	cDNA
60µl	Nuclease-free water
6 µl	$MgSO_4$ (25 mM)
10 µl	dNTPs (2 mM)
10 µl	10X KOD Hot Start Polymerase buffer
1 µl	P1 primer (10 nm/ µl)
1 µl	P2 primer (10 nm/ µl)
2 µl	KOD Hot Start Polymerase

Table 2.7. PCR cycle for amplification of chemically fragmented libraries. Reactions were incubated at 95 °C for 5 minutes to activate the KOD hot start polymerase, followed by 10-18 cycles of PCR.

Temperature	Time	Cycles
95 °C	5 min	-
95 °C	15 sec	10-18
62 °C	15 sec	
72 °C	1 min	
72 °C	5 min	-
4 °C	hold	-

2.3.17. Size selection of RNA fragments by FlashPAGE fractionator

Size selection of RNA fragments using the FlashPAGE fractionator (Ambion) was performed as an alternative to agarose gel size selection of cDNA libraries prior to PCR amplification in SOLiD library preparation. During FlashPAGE electrophoresis, RNA and DNA molecules were separated from longer species and collected in the lower running buffer chamber. This removed the requirement for libraries to be purified from gels, a step that was found to be unreliable when performed with previous protocols supplied with SOLiD Small RNA Expression Kit (SREK) protocols. The FlashPAGE fractionator was set up using FlashPAGE pre-cast gels (Ambion) and electrophoresis performed following the manufacturer's protocol for the recovery of polynucleotide fragments 50-150 bp in size.

RNA fragments were purified and concentrated from FlashPAGE running buffer using a FlashPAGE reaction cleanup kit (Ambion) following the manufacturer's instructions.

2.3.18. TA cloning

PCR products were run on 1.5 % agarose gels to identify products of the appropriate 200 bp length. These regions of the gel were excised, and DNA isolated using a Qiaquick gel extraction kit following the manufacturer's protocols. DNA was transformed into highly competant E.coli cells using a pGEM-T Easy Vector System (promega) according to the manufacturer's protocols. Transformants were grown on selective media containing ampicilin and blue-white scoring used to identify successful transformations. Plasmids were extracted from successful transformants using a plasmid extraction kit (QIAGEN) according to manufacturer's protocols, and digested with EcoR1 prior to running on a 1.5 % agarose gel to identify cloned fragments.

2.4. Bioinformatics and computational biology

2.4.1. Motif finding with YMF and MEME

To ensure robust analysis, *de novo* motif identification was performed using two different motif finding algorithms; YMF 3.0 (Sinha & Tompa, 2000; Blanchette & Sinha, 2001; Sinha & Tompa, 2002; Sinha & Tompa, 2003) and MEME (Bailey & Elkan, 1994). Both programs can be run via webservers, but due to the large number of sequences being searched, YMF software and the MEME Suite (Bailey *et al*., 2009) were downloaded and run locally.

For motif finding with YMF, it was necessary to construct a background model for *A. nidulans* transcription start sites. The background sequences supplied to YMF consisted of 1000 bp upstream and 199 bp downstream (1200 bp total) from each putative TSS identified by 5' RNA-seq analysis.

To capture the promoter, a sequence 200 bp upstream and 24 bp downstream was defined for each TSS. YMF motif finding was performed within these regions for motif lengths of 6 bp, 7

bp, 8 bp, 9 bp and 10 bp. Each run produced 1000 predicted motifs ordered by Z-score, indicating overrepresentation of the patterns against background sequences.

To reduce redundancy in the raw YMF output, the software MATLIGN (Kankainen & Löytynoja, 2007**)** was used to group the motifs into subsets. MATLIGN is extremely inefficient when processing very large datasets, so only the top 500 motifs (by Z-score) of YMF output were selected. The Z-scores of motifs after the 500 cut-off were relatively low (5 times less than the top motifs) so we were confident that no significant motifs were discounted.

Motifs found by YMF were verified by a second round of motif finding using MEME, again repeated 5 times for motif lengths of 6-10 bp. The output of MEME provided several statistics relating to each motif. Sites is the number of occurrences; llr is the log likelihood ratio which measures difference from the background model; and E-value indicates the statistical significance of llr value. Lower E-values indicate stronger putative motifs.

2.4.2. Building a Bowtie reference index

The Bowtie short read aligner uses a Burrows-Wheeler Transform to carry out rapid alignment of sequence reads. In order to do this, the reference genome must be represented as an 'index'. The function "Bowtie-build" was used to create a Bowtie reference index from a set of DNA sequences. To obtain these sequences, the latest *A. nidulans* genome reference was downloaded in FASTA format from the Ensembl Genomes ftp server (ftp://ftp.ensemblgenomes.org/pub/fungi). The resulting output was a set of 6 files with suffixes .1.ebwt, .2.ebwt, .3.ebwt, .4.ebwt, .rev.1.ebwt, and .rev.2.ebwt, which together

comprise the Bowtie index. The original sequence file was no longer needed by Bowtie, as all alignments are made against the index files.

The only options supplied were "-C" to generate a colour space reference, and "-f", which told Bowtie that the reference genome was in FASTA format. All other settings were run as default. <reference_in> corresponded to the reference FASTA, while <ebwt_base> defined the prefix of the Bowtie index files.

Bowtie-build usage:

bowtie-build [options] <reference_in> <ebwt_base>*

2.4.3. Mapping of SOLiD RNA-seq reads with Tophat

Alignment of reads to the assembled Bowtie reference index and subsequent junction finding was performed with Tophat version 1.3.1. Tophat is an aligner for mapping RNA-seq data. It uses Bowtie to align reads to a reference index. It enables mapping of spliced reads by splitting the reads into segments, mapping each segment and defining instances where mapped segments are separated by a putative intron. For all Tophat runs, reads produced by SOLiD sequencing were supplied in ".csfasta" format (<reads>), with their associated quality values in ".qual" format (<quals>). The folder containing the six Bowtie index files was given as the reference against which reads would be aligned (<index>).

The option "-C" was used to indicate that the input reads were in colour space, while "-Q" told Tophat to utilise the separate .qual files associated with colour space reads. Finally, the library type was defined using "--library-type fr-secondstrand". This told Tophat that reads were strand specific, in SOLiD format, and to enforce the rule that the left-most end of the

fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Output directories for each run were defined using the option "-o". Additional options were also used depending on the required outcome of the mapping run. Where required, maximum intron length was set using the option "-I" followed by the length required. If a reference gene model annotation was required, this was supplied using the option "-G" followed by the name of the reference file. The CADRE 2.5 gene model annotation was used throughout this analysis (A_nidulans.CADRE2.5.gtf).

Tophat usage:

Tophat [options] -C –Q --library-type fr-secondstrand -o outputdirectory <index> <reads1> <quals>*

Where multiple inputs from more than one SOLiD library were being aligned in a single Tophat run, the read and quality files were supplied in comma separated lists, e.g. <reads_1,reads_2,…,reads_N>.

Mapped reads were output in binary alignment/map (BAM) format, a compressed, binary form of the sequence alignment/map (SAM) format (Li *et al.,* 2009). SAM/BAM format is a popular way to represent large scale alignments. Optional header lines contain information (metadata) on the reference sequence, library, sequencing platform etc., while all other lines consist of a set number of tab-separated fields describing each sequence read (mapped position, orientation, number of mismatches, uniqueness of mapping etc.). Predicted splice junctions were separately annotated in a tab delimited BED file, the format of which is described on the UCSC Genome Bioinformatics web site (http://genome.ucsc.edu/FAQ/FAQformat).

2.4.4. Indexing and sorting of BAM files with SAMtools

BAM format is the recommended input for Broad Institutes Integrative Genomics Viewer (IGV). IGV requires that BAM files be sorted and indexed, and that the BAM index file follows a specific naming convention. Specifically, BAM index files should be named with suffix ".bai" appended to the name of the original BAM file (The SAM Format Specification Working Group, 2011). Sorting a BAM file places reads in order of leftmost coordinates and was performed with the "sort" command in SAMtools, defining a BAM input (aln.bam) and sorted BAM output "aln.sorted". All options were as default.

SAMtools sort usage:

samtools sort aln.bam aln.sorted

Once the alignments had been sorted, they were indexed to facilitate fast random access. This enables an alignment viewing program to access a portion of this very large file at a time, reducing the amount of computer memory required. The sorted BAM file (<aln.sorted.bam>) was indexed using the "index" command in SAMtools.

Index usage:

SAMtools index <aln.sorted.bam>

This command created the index file <aln.sorted.bam>.bai, which was renamed to aln.bam.bai for use with IGV.

2.4.5. SamTools pileup from BAM read alignment data

Pileup is a text-based format for summarizing the base calls of aligned reads to a reference sequence, and was originally developed by Tony Cox and Zemin Ning at the Wellcome Trust

Sanger Institute before becoming widely known within the SAMtools software suite (Li *et al.*, 2009).

The mpileup feature of SAMtools was used to convert read alignment data from BAM files (<aln.bam>) into a pileup format. A number of options were used to generate the optimal pileup for downstream analysis of read alignment. The option "-D" produces an output which gives output of read depth at each position. "-B" disabled probabilistic realignment for the computation of base alignment quality (BAQ). BAQ is the Phred-scaled probability of a read base being misaligned. Applying this option greatly helps to reduce false SNPs caused by misalignments. "-Q" was used to set the minimum mapping quality to 0, providing the most comprehensive pileup. "-d" was used to set the maximum depth of coverage to an arbitrary number which was extremely high, to ensure that all aligned reads would be counted. Finally, "-f" was used to supply mpileup with a reference genome: the same FASTA format genome used to build the Bowtie reference index against which reads had been aligned (<AN_bowtie.fa>) (Chapter 2.4.2). Output was redirected to a text file (<pileup.txt>) for downstream analysis.

SAMtools mpileup usage:

samtools mpileup -D -B -Q 0 -d 1000000000 -f <AN_bowtie.fa> <aln.bam> > <pileup.txt>

2.4.6. Combining alignment coverage for multiple libraries

BAM files containing read alignments from multiple libraries were combined using the merge function in SAMtools. The output was a single BAM file containing all the reads from each input BAM. As this format of compiled data was being used only for the purpose of

producing a pileup of read coverage from multiple libraries, the merge was performed using default settings.

SAMtools merge usage:

samtools merge <out.bam> <in1.bam> <in2.bam> [...]

2.4.7. Quantification of uniquely mapped reads from Tophat alignment data

SOLiD read IDs were extracted from the first column of Tophat read alignment data which had been converted from BAM to SAM format as previously described. These IDs were sorted and made unique. The number of lines with unique IDs was then counted. This process was performed using the following command:

cut -f 1 <accepted_hits.sam> | sort | uniq -c | wc -l

It is important to note that SAM files may contain a header conserved from the original BAM file, which does not represent read alignment data. Where this header was found, the number of unique read alignments was reduced by 3, representing the number of lines in the header with unique values in column 1.

2.4.8. Cufflinks assembly of transcripts

Cufflinks was used to assemble mapped RNA-seq reads into a list of predicted transcript structures, taking a BAM format alignment file as input. To ensure accurate assembly, the SOLiD library format was specified using the command "--library-type fr-secondstrand".

Cufflinks usage:

Cufflinks [options]* <aligned_reads.bam>

The cufflinks software is available online at: http://bio.math.berkeley.edu/cufflinks.

2.4.9. Counting unique genetic loci and numbers of alternative isoforms predicted by Cufflinks assembly

Cufflinks is able to predict multiple isoforms of transcripts originating from the same genetic locus. In these cases, additional isoforms can be identified as having the same basic Cufflinks CUFF ID, but with a unique number after the final stop character (e.g. CUFF00001.1, CUFF00001.2 etc). The number of unique genetic loci at which transcripts have been assembled can therefore be obtained using the grep function of unix to search for all CUFF IDs ending in ".1", sorting them, then finding the number of unique IDs and counting the lines on which they occur.

Grep usage to find the number of unique loci:

```
grep -o '\"CUFF\.[1-9]*\.[1]' transcripts.gtf |sort |uniq | wc -l
```

Some transcripts may have several isoforms, to determine the number of transcripts with 2 or more isoforms, the above command was altered to search for CUFF IDs ending in ".2". This number (N) was then changed to find those transcripts with increasing numbers of isoforms.

Grep usage to find the number of loci with N transcript isoforms:

```
grep -o '\"CUFF\.[1-9]*\.[N]' transcripts.gtf |sort |uniq | wc -l
```

2.4.10. Converting BAM to SAM format

BAM format files were converted to SAM format using the view function of SamTools. The –h option was used to maintain the BAM headers, and –o to define the output file.

Samtools usage:

samtools view -h -o out.sam in.bam

2.4.11. Splitting BAM files by strand

To improve the accuracy of Cufflinks transcript assembly, read alignments in BAM format were split by strand. The "view" feature of Samtools was used (Li *et al*. 2009) to decompress the BAM file (binary) to a SAM file (a text flatfile) and search the data stored for each read. The hexadecimal code 0x10 (16) in the flag column indicated mapping to the minus strand. Lines containing this code were found using the "–f" command and output to a reverse strand specific BAM file. The command was then modified to "-F" to identify lines which do not match the 0x10 search function, and these lines were output to a forward strand specific BAM file (note that the Tophat output contains only mapped reads with bitwise flags of 0 for forward and 16 for reverse, some mappers also output unmapped reads with the 'unmapped' flag 4). Two additional options were used in the command line; "-b" instructed Samtools to produce an output in BAM format, and "-s" was used to maintain the headers in the output file.

Samtools usage to isolate forward strand alignments:

samtools view -b -h -F 0x10 input.bam > output.forward.bam

Samtools usage to isolate reverse strand alignments:

samtools view -b -h -f 0x10 input.bam > output.reverse.bam

2.4.12. Cuffcompare of transcripts to reference

Cuffcompare takes a cufflinks GTF output as input, comparing it to other cufflinks GTFs and/or to a reference annotation. The reference annotation file is provided in GTF format using the option "-r" followed by the reference file location.

Cuffcompare usage:

cuffcompare [options]* <cuff1.gtf> [cuff2.gtf] ... [cuffN.gtf]

2.4.13. Alignment of ORFs to the Pfam protein database

Lists of putative ORFs produced by the software "getorf" and processed to remove duplicate sequences and unwanted newline characters were aligned to the Pfam protein database using the online batch sequence search facility (Punta *et al*., 2012) available at: http://pfam.sanger.ac.uk/search#tabview=tab1. Standard settings were used.

2.4.14. Alignment of ORFs to the Kyoto Encyclopaedia of Genes and Genomes

Alignment of ORFs to the Kyoto Encyclopedia of Genes and Genomes (KEGG) was performed using the online KEGG Automatic Annotation Server (KAAS). As a limited number of putative genes were being aligned, the single-directional best hit (SBH) method was used and ORFs uploaded in multi-FASTA format. Gene data sets from all available Eurotiomycetes were used for KAAS alignment, as well as other well-characterised fungal genomes including *Saccharomyces cerevisiae* (budding yeast), *Neurospora crassa* and *Schizosaccharomyces pombe* (fission yeast). All settings were as standard.

2.5. Novel Software

2.5.1. Use of Perl

All novel software was created using the Perl programming language (Wall, 1999).

2.5.2. Pileup software for Corona lite aligned RNA-seq reads

Software designed and produced by Dr Kevin Ashelford for the creation of pileup data from Corona lite alignments of SOLiD RNA-seq reads and an example shell script for running this software can be found in Appendix2.2. The software is not described in further detail as its creation was not part of this thesis.

2.5.3. Gene expression using SOLiD version 2 libraries

Running this software requires three additional options set in the command line. The option "–f" is used to define the file location containing mapped SOLiD read data in double stranded pileup format. The chromosome to be processed is defined using the "-c" option, followed by the chromosome number. The final option "-m" is used to set the minimum read coverage per exon base for a gene to be reported as expressed. Three text files are produced as output, "expressed.txt" contains a list of genes found to be expressed, "nonexpressed.txt" containing a list of genes which were not expressed, and "Totals.txt" containing a numerical summary of these data. All output file names are prefixed with the appropriate chromosome number to facilitate batch processing using shell scripts.

This software can be found in Appendix 2.1.1.

2.5.4. Antisense estimation

Software was developed to assess antisense transcription. This can be done by comparing read coverage on forward and reverse strands of each chromosome to annotated genes to quantify the amounts of sense and antisense RNA across each gene. To do this, the software took chromosome specific *A. nidulans* reference annotation files (in .gbk format) and double stranded pileup data in .csv format as input. The minimum overlap percentage, minimum overlap bases and minimum coverage for an overlap base to be reported can be set within the script. The default values were all set to 1, reporting all putative antisense occurrences.

This software read the read alignment frequencies of each base position from the 2 columns of the pileup data, placing them into 2 arrays; coverage of the forward strand and coverage of the complementary strand.

Gene model positions were then extracted from the .gbk format reference using regular expressions to identify lines corresponding to mRNA annotations, then splitting these lines by the comma character used to separate position values, which were then stored in an array. Strand specificity is achieved by querying each line for the “complement” marker uses to denote a gene on the complement strand, and genes for each strand are processed separately.

Genes were processed individually, with positions for each gene being added to the array until the start of the next gene annotation was reached. At this point, the gene positions were sorted by size from smallest to largest, and the length of the gene calculated. This sorting step was required because the .gbk annotation format lists compliment gene positions in relation to the forward strand, listing them from highest to lowest.

Gene positions were then used to interrogate the pileup array corresponding to the opposite DNA strand. Within the compliment region, the number of bases shown to have a read mapping frequency above the minimum value was counted and used to calculate the percentage of the gene covered by antisense transcription. Genes with antisense coverage greater than the minimum values were reported as having a putative antisense transcript, and recorded in output files. A summary file was also created, listing the total number of antisense occurrences, and a breakdown of this total by coverage percentage.

This software can be found in Appendix 2.1.2.

2.5.5. Masking positions with low readhead coverage

This software requires readhead pileup data in single stranded format as input, with strand defined as “p” or “m” for the forward and complement strands respectively. Input files should follow the naming convention: <chromosome number>_<strand>.csv

Upon running the software, the user was required to input the location of the folder containing the input files, the number of chromosomes to be processed and the maximum coverage for a RH position to be masked. The software then read through the pileup data and set the coverage of any positions where RH frequency was less than or equal to this number as "0". Modified pileups were output to a new .csv file.

This software can be found in Appendix 2.1.3.

2.5.6. Association of TSS and motifs to gene models

This software required several additional options set within the command line. " –s <sample name>" specified the sample name, corresponding to the folder in which input files were located. "-c <[1-8]>" specified the chromosome to be processed, while "-d [+-]" specified the positive or negative strand of the chromosome. The facility to produce an output file for use with the SEED database (Overbeak *et al.* 2005) was built into the software, and the option "–o [yn]" used to determine if this output was produced, however this was not used in this thesis and this option was therefore set to "n".

A number of input files were required, including a set of motifs (in IUPAC code) in a newline delimited text file, lists of TSS region locations and their associated CI values, and the original RH data from which these TSS regions had been located. The required file names and locations required were as follows:

Motifs: <sample name>/motifs.txt

TSS regions: <sample name>/TSS_locations/out_<[1-8]><[+-]> .txt

RH pileups: <sample name>/readheads/c[1-8]_[+-].csv

The *A. nidulans* genome assembly in chromosome specific .gbk format was also required in a subfolder (named "references") of the directory from which the software was run.

Further options could be set within the script, allowing the user to define the promoter length upstream and downstream of the TSS (default: 200 and 25 respectively), as well and the lower and upper CIL values to use when selecting TSS (default 0 and 100 respectively). The furthest distance from an annotated gene start site for a TSS to be associated with that gene could also be defined, as well as the closest distance upstream of a gene stop site (defaults 500 and 0 nt respectively).

This software can be found in Appendix 2.1.4.

2.5.7. Generating statistics for production of consensus sequences surrounding transcription start sites

Running this software requires an additional option set within the command line; "-s" followed by the sample ID. This ID should be the name of the folder containing the TSS list files used as input, and the prefix of each input file, the names of which should be in the format "<Sample ID>_out_<chromosomenumber><strand p or m>.txt", e.g. "Sample1_out_1p.txt". Chromosome specific reference files in .gbk format should be present in a subdirectory titled "references", with file names corresponding to chromosome number (e.g. "1.gbk").

This software generates counts of nucleic acid frequency at each base position within a set of 21 nt sequences centred on the TSS locations defined by the input files. The percentage occurrence of each nucleic acid at each base position is calculated, and the "Spreadsheet::WriteExcel" module of Perl used to write these data to an Excel spreadsheet.

The software can be found in Appendix 2.1.6.

2.5.8. Selection of uniquely mapping reads with complete 5' end alignments

Novel software was developed using the Perl programming language to filter reads based on these criteria. Using the "output.csfasta.ma" files produced by Bioscope mapping of SOLiD 5' specific RNA-seq reads as input, this software identified csfasta headers and associated read data. The position in the read identified as the start of the alignment was isolated from the data in the csfasta header, and reads with alignment start data matching "0" (i.e. alignment starts at the 0 position in the read, representing the first base at the 5' end) were output to a new file in ".csfasta.ma" format. Where reads had been aligned to multiple positions, multiple alignment start positions were listed. These positions therefore did not match to the single "0" required for output, and the corresponding reads were excluded from the output file.

The software can be found in Appendix 2.1.7.

2.5.9. Generating readhead pileups from selected read data

To convert mapping data from selected reads to a pileup of read 5' mapping positions or readheads (RH). This software generated a series of arrays, representing the forward and reverse strands of each *A. nidulans* chromosome. Each base was represented by a single array element, which would contain the RH mapping frequency at a given position. The length of each array was determined using the number of base positions indicated by *A. nidulans* genome reference files in FASTA format, obtained from CADRE (www.cadregenomes.org).

Reads which had previously been selected for full length 5' end mapping and unique genome position (Chapter 2.5.8) were taken as input. Reads were processed individually, firstly extracting the chromosome number and the position at which mapping initiated (i.e. the readhead). Positions with a negative value indicated that the read was mapped to the reverse strand. These data were used to define the array representing the chromosome and strand to which the read had been mapped, and the element of the array corresponding to the RH position was incremented by 1 to record a single readhead mapping to this position. Processing of each read in this way created a pileup of RH mapping frequency at each position throughout the *A. nidulans* genome. Once all reads had been processed, arrays were output to text files with each element on a newline, creating a pileup format which could be viewed in the Artemis genome browser (Rutherford *et al*., 2000; Carver *et al*., 2012).

The software can be found in Appendix 2.1.8.

2.5.10. Comparison of read head mapping positions

When run, this software required user input to define folders containing two sets of chromosome and strand specific pileups of read head (RH) data to be compared. Each chromosome strand is processed individually. RH frequencies from the first set of pileup data is read into an array. The software then moves through this array and the pileup data from the second dataset simultaneously, comparing the RH frequencies at each base position. The number of bases where RH mapping was observed simultaneously in both datasets was recorded, as well as the total number of bases with RH mapping. Once both strands of each chromosome had been processed, the percentage of matching RH positions was calculated and output to "comparison_report.txt".

This software can be found in Appendix 2.1.9.

2.5.11. Analysis of intron lengths

Software was developed to extract intron lengths from the *A. nidulans* gene model annotation.

The workflow for this software is illustrated in Fig. 2.1. Taking chromosome specific gene annotation files as input, the software searched through the annotation line by line to identify exon start and stop positions for each gene. These positions were stored in a data array, and processed when the software reached a line in the annotation corresponding to a new gene. The first element of the array was removed as this indicates the start of the first exon, and gives no information on intron positions. The length of the first intron was then calculated from the difference between the first two elements of the array, which corresponded to the first exon stop and the second exon start within the gene in question. These two elements were then removed from the array, the next two processed in the same way, and this cycle repeated until all introns in the gene had been identified. Once calculated, individual intron lengths were stored in a separate data array for future use.

The exon position array was then cleared, and the software continued to read through the reference annotation, gathering the exon positions for the next gene. Once all genes had been processed in this way, each element of the intron length array was written to a new line of a text file, producing a list of all intron lengths in *A. nidulans*.

The software also recorded the ID of each gene as it was being processed. A scalar variable was added to record the largest intron found. The size of each newly identified intron was compared to the length stored in the variable. If the new intron was found to be larger, the variable was overwritten with this new value. Two more scalars were added to store the gene

ID associated with largest intron and the chromosome number on which the gene was located. After processing all genes in the reference, the value in each of these scalars was printed to the console, giving the largest intron and the gene in which it was found. By providing these data directly from the software output, the gene in question was quickly identified. Gene structure analysis could then be conducted using mapped RNA-seq reads displayed against the reference genome in the Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdottir *et al*., 2012).

A copy of this software can be found in Appendix 2.1.10.

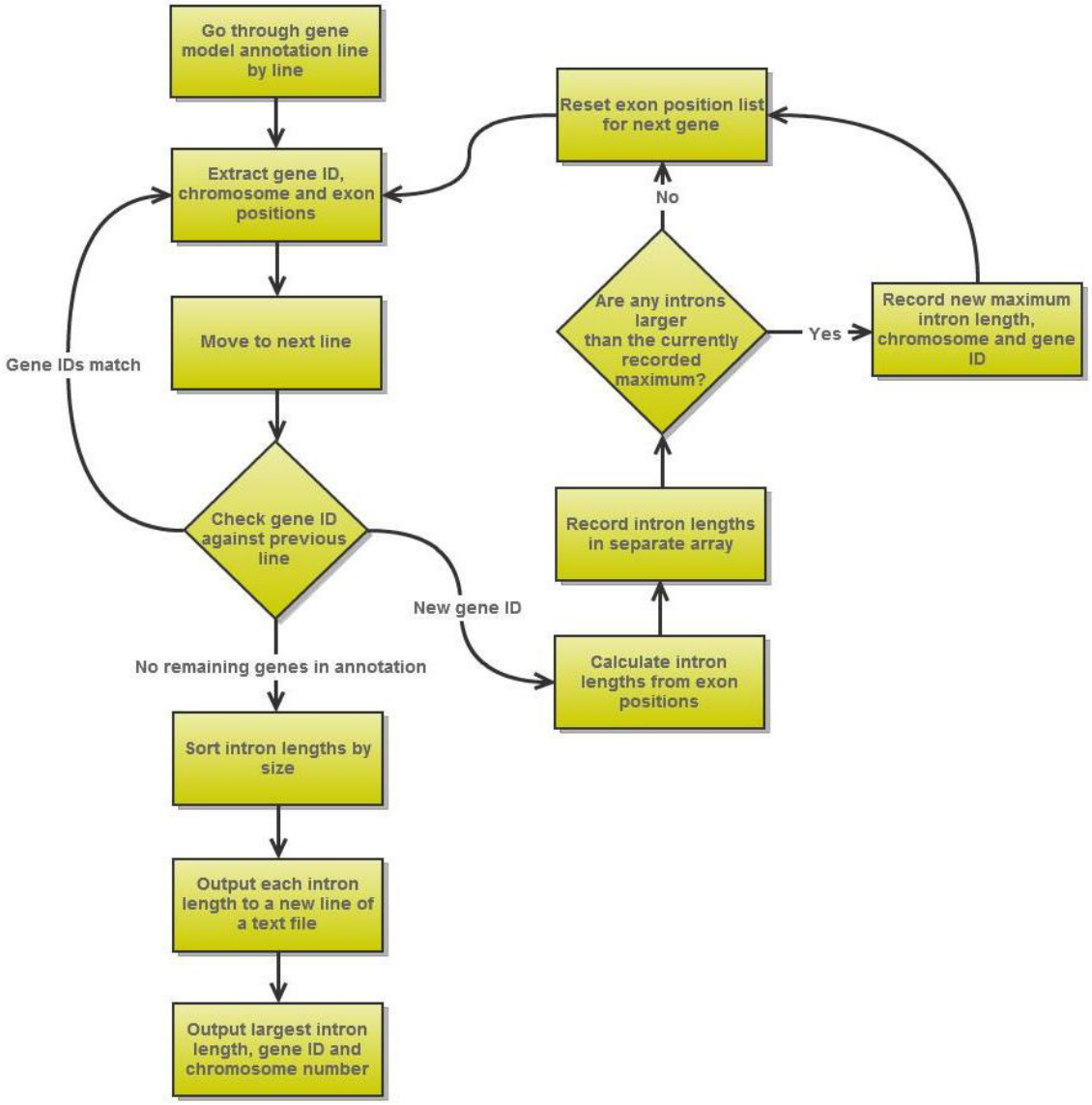


Figure 2.1. Process for extraction of intron lengths and largest intron data. Software read through the reference gene model annotation line by line. Exon positions were extracted and used to calculate intron lengths for each gene. Intron lengths for all genes were recorded, sorted in order of size and output as a newline separated list in .txt format. The largest intron found was also recorded and printed alongside its associated gene ID and chromosome number into an output file.

2.5.12. Comparison of Tophat splice junctions

Novel software was created to compare the junction locations from two junction.bed files, produced by Tophat mapping of RNA-seq reads. One of these junction files contained the results of *de novo* junction finding, while the other contained no novel junctions, thereby facilitating the identification of novel junctions.

This software first interrogated the file containing no novel junctions. Each line of a .bed format file contains the data for a single junction feature in a TAB delineated format. These data were converted into elements of an array by splitting it based on whitespace characters (i.e. TAB). Feature chromosome location, start and stop positions were then extracted from the array, along with the values by which the start and stop should be modified to locate the true start and stop positions of the respective junction. Three new arrays were created to contain chromosome locations, true junction starts and true junction stops. A given position in all three arrays therefore corresponded to the three data for a single junction.

Each line of the second file was then interrogated in a similar fashion, but chromosome, junction start and junction stop positions were not placed into arrays. Instead, these data were compared to all the junctions stored in the three arrays. Where no match was found, the current line of the second file (representing a novel junction) was output to a new gtf file.

This software can be found in Appendix 2.1.11.

2.5.13. Remove RH pileup positions covered by annotated gene models

This software required additional options to be set in the command line. "-s <sample name>" was used to define the file containing mapped read coverage data in double stranded pileup format (.csv). A second option "-c [1-8]" was used to set the chromosome number to be

processed. The *A. nidulans* genome assembly in chromosome specific .gbk format was also required in a subfolder (named “references”) of the directory from which the software was run.

This software can be found in Appendix 2.1.1.

2.1.14. Software to obtain transcript positions

Software was developed to take a list of transcript IDs as input, and compare these to the transcript IDs of the corresponding Cufflinks “transcripts.gtf” output file. The ID of each transcript in the input list was read into an array, and processed to remove any surrounding information. For each transcript ID in the array, the transcripts.gtf file was searched for lines with matching IDs. As only the start and stop position of each transcript was required, only the first line of information for each transcript was interrogated, as this gave the full length of the transcript, with subsequent lines giving information on each predicted exon. Start and stop positions were placed into an array. As the GTF format lists compliment strand sequence positions from 5' to 3', elements in the array were sorted by position on the chromosome to produce a consistent output format. Transcript positions, chromosome and strand were printed to a text file, and the entire line of the GTF file printed to a new GTF which would contain only identified intergenic transcripts for visualisation in genome browsing software.

This software can be found in Appendix 2.1.15

2.1.15. Software to obtain transcript sequences

Software was designed to take a reference sequence in FASTA format and a list of transcript positions which had been created as described in Chapter 2.5.14 as input. Using the FASTA reference, arrays were produced for each chromosome, with individual elements corresponding to a single nucleotide position and containing the appropriate nucleotide sequence. To obtain the correct sequences for transcripts on the reverse strand, a second set of sequence arrays was created, and sequences converted to their compliment. The compliment sequences did not need to be reversed as all positions given in Cufflinks output GTF files were in relation to the positive strand.

For each line of the transcript position list, the information was split to allow the isolation of chromosome number, start and stop position. The chromosome number and strand information were used to define the correct sequence array to use for each transcript. Sequences in the array elements corresponding to nucleotide positions between the transcript start and stop locations were extracted and printed to an output file in multi-FASTA format, with headers consisting of the transcript ID, chromosome, strand and positions.

This software can be found in Appendix 2.1.16

2.5.16. Counting of ORFs and their associated transcripts

To obtain statistics regarding ORFs identified by the ORF finding software "getorf" (http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html), novel software was developed to process the getorf output file. This software requires the user to define a getorf output file in FASTA format. The input file is processed line by line, counting the number of FASTA headers and interrogating them for the cufflinks ID of the transcript in which the

ORF was identified. These cufflinks IDs are listed as the first value of each FASTA header, with a "_" between each subsequent value. Headers were split into arrays using "_", resulting in the cufflinks ID being placed into the first element of the array. Each ID was pushed into a new array, which by the end of the file contained a list of the transcripts associated with each ORF. The elements in the array were processed using the "uniq" function to remove repeats, so each unique transcript ID was listed once. The length of the array was then found to determine the number of unique transcripts in which ORFs had been found by getorf.

This software can be found in Appendix 2.1.17.

2.5.17. Remove duplicates and newline characters from multi-FASTA sequence lists

Lists of ORFs in FASTA format were produced by the software "getorf". These lists were not immediately suitable for input to the Pfam batch sequence search as they contained a large number of duplicate sequences and newline characters. Software was developed to remove these by reading through the FASTA file line by line and removing newline characters. Duplicate sequences were removed by comparing each header to all previously processed headers. The first header in the file and its accompanying sequence were printed to a new FASTA format file. The header was then stored in an array. The next header was then compared to the header stored in the array, if a match was found, the new header and its sequence were skipped, otherwise they were printed to the output file and this new header stored in the array in addition to the previous one. Each new header was compared to headers stored in this array, with matches identifying repeat sequences. These repeat sequences were discarded, while previously unprocessed sequences were printed to the output file and their headers stored in the array to interrogate future sequences. This process produced a multi-

FASTA file with unique sequence headers in the correct FASTA format for input into downstream applications such as Pfam.

This software can be found in Appendix 2.1.18

2.5.18. Comparison of Pfam and KEGG protein database alignments

Software was developed to compare the results of ORF alignments with the protein domain databases Pfam and KEGG. This software took the output file from Pfam and a KEGG Ontology (KO) list as input. The Cufflinks transcript ID associated with each ORF had been used as the start of each ORF ID. ORFs with matches in the Pfam output file were therefore identified by lines starting with a "C", indicating the start of the associated Cufflinks ID, and were recorded in an array. The KAAS KO list contained all ORF IDs in a single column. A second column contained a list of KO terms alongside ORFs where an alignment had been made. Lines with values in both columns therefore indicated aligned ORFs, and the ORF IDs from these lines were extracted and placed into a second array. Elements in each ORF ID array were made unique to remove any duplicates of ORF IDs which were present due to multiple protein database alignments to a single ORF. The elements of each array were then compared. Matches were recorded and counted. The number of matches between the two datasets were then printed to an output file, along with lists of matched and unmatched ORFs.

This software can be found in Appendix 2.1.19.

2.5.19. Identification and classification of antisense transcripts

This software required the user to define 2 Cufflinks transcript files in gtf format, representing forward and complement strand transcripts produced from alignment data which had been split by strand.

The software used the first line of the GTF file to determine which transcripts belong to each chromosome, and processed each chromosome individually. Transcript start and stop positions were compared in three ways, looking for overlaps between the 5' or 3' ends of transcripts, or for instances of one transcript being totally covered by a transcript on the opposite strand. The number of matches in each category was counted as they were identified and these counts were printed to a text file. The software also produced a list of overlapping transcripts, with each pair given a header listing the chromosome, an ID number for that pair, the class of antisense relationship and the length of the overlap between the two transcripts (Format: <chromosome>_AS_<ID>_<Class>_<Overlap>)

Antisense classes were as follows:

- strand transcript entirely within + strand transcript (class 1)

+ strand transcript entirely within - strand transcript (class 2)

Overlap at 3' end (class 3) = $prime3

Overlap at 5\' end (class 4)

A minimum overlap of 20 bp was used for this analysis, but this can be set within the script if a different overlap length is required.

This software can be found in Appendix 2.1.20.

Chapter 3

3. High-throughput whole transcriptome RNA-seq in *Aspergillus nidulans* on the ABI SOLiD version 2 platform

3.1. Sequencing of SOLiD Total-RNA-seq libraries

To accurately define the *Aspergillus nidulans* transcriptome on a global scale, high throughput total RNA-seq was performed on the ABI SOLiD platform.

Initial attempts to create total RNA-seq libraries for SOLiD sequencing using the SOLiD Whole Transcriptome Analysis Kit (ABI) failed repeatedly. The SOLiD system was in its infancy and library preparation protocols and kits were not yet perfected. Library preparation using the SOLiD Small RNA Expression Kit (SREK) (ABI) was explored as an alternative to the Whole Transcriptome Analysis Kit. SREK provided a more up to date, rapid and robust protocol for library preparation, and allowed each sample to be tagged with one of ten barcode sequences. Barcoding allowed libraries to be pooled together into a single sample for ePCR and sequenced simultaneously on a single slide. This multiplexing capability unlocked the full sequencing capacity of the SOLiD system, facilitating rapid data production and greatly reducing the per-library cost of SOLiD sequencing.

The more robust SREK protocol was successful in producing sample libraries which passed quality control criteria for SOLiD sequencing.

3.2. Creation of two SOLiD libraries with SREK

Two whole-transcriptome libraries were created using SREK and successfully sequenced on the SOLiD system version 2, producing reads 35 bp in length. Two additional libraries were also sequenced, but gave no meaningful results.

RNA used in the creation of the two sequenced libraries was extracted from G00 wild type *A. nidulans* cultures grown under standard conditions, one on complete media (library ID: S1) and one on minimal media (library ID: S3) supplemented with nitrate as sole nitrogen source (see Chapter 2.2.4 for detailed growth conditions). Contaminating DNA was removed from total RNA samples by treatment with DNase I.

The resultant total RNA samples represented all RNAs in the *A. nidulans* cells at the time of extraction. Large ribosomal RNA (rRNA) constitutes approximately 90% RNA species in total RNA. The whole transcriptome must therefore be isolated prior to RNA-seq (Chen & Duan, 2011), by either depletion of rRNA or enrichment for mRNA. Eukaryotic mRNA utilise two cotranscriptional modifications to determine the stability of the molecule; the 5′ cap and the 3′ poly(A) tail. Shortening of the poly(A) tail is generally the rate-limiting step, resulting in either transcript decapping and subsequent 5′-3′degradation (Tucker & Parker, 2000) or exosome-dependent 3′-5′ decay (Butler, 2002; Mitchell & Tollervey, 2000). Decapping has been shown to occur when the poly(A) tail is shortened to around 15 residues (Couttet *et al*., 1997; Decker & Parker, 1993; Muhlrad *et al*., 1994). Full length, capped mRNAs can therefore be isolated from total RNA by complimentary binding of the poly(A) tail to immobilised oligo(dT). Enrichment for mRNAs from total RNA samples was therefore performed using an Oligotex oligo(dT) mRNA purification kit (QIAGEN).

mRNAs were decapped with Tobacco Acid Pyrophosphatase (TAP). Decapping of mRNAs is vital to SOLiD library preparation as the 5′cap structure inhibits the ligation of the 5′ adaptor. TAP hydrolyzes the phosphoric acid anhydride bonds in the triphosphate bridge of the 5′ cap structure found in most eukaryotic mRNA. This releases the cap nucleoside and generates a 5´-monophosphorylated terminus on the RNA molecule (Lockard *et al*., 1981), which is required for SOLiD adaptor ligation.

SOLiD v2 sequencing of S1 and S3 produced 52,626,177 and 46,271,839 reads respectively. A current assembly of the *A. nidulans* genome was downloaded from CADRE (annotation version 14/05/2009) and used to assemble a reference genome for the SOLiD System Analysis Pipeline Tool; Corona.

Initial read mapping with Corona Lite utilised a 35_3 schema, allowing 3 colour space mismatches across each 35 bp read length, and resulting in 13.24% (S1) and 10.29% (S3) of reads mapping to the reference genome. The numbers of uniquely mapping reads were very similar to the total mapping, at 12.75% (S1) and 9.84% (S3), reflecting the low number of repeat regions in the transcriptome.

Following updates to the SOLiD software, mapping was repeating using the same reference genome with the Whole Transcriptome Analysis Pipeline (WTP) version 1.1 (ABI). The more sensitive WTP increased the number of mappable reads to 16.4% (S1) and 13.4% (S3), with unique mapping rates of 14.6% and 11.7% respectively. While a notable improvement over previous mapping figures, these percentages were still far lower than mapping figures achieved by a library created from SREK control RNA (human). 29.8% of sequenced reads from a library prepared with this control RNA mapped to the human genome, with 16.7% mapping uniquely.

Although we had successfully sequenced and mapped two whole transcriptome libraries in *A. nidulans*, the two unsuccessful libraries and low mapping percentages of S1 and S3 required further investigation into quality control before the resultant data were considered fit for utilisation in downstream analysis.

3.3. Investigation of mapping quality in SOLiD libraries

The high proportion of well mapped and uniquely mapping reads was an encouraging indicator of quality in both S1 and S3 libraries. However, the percentage of mappable reads in both cases was still low in comparison to libraries created with SOLiD control RNA (human) and mapped to the appropriate genome. The quality of the *A. nidulans* libraries was also called into question by two out of the four sequenced libraries producing data which could not be mapped to the *A. nidulans* genome. This prompted an investigation into possible contaminants which could negatively influence the integrity of SOLiD libraries.

To better assess the quality of these libraries, reads were mapped against human genome to test for possible contamination. Due to the significant size of the human genome, a certain level of background mapping was inevitable. To facilitate an accurate assessment of this mapping, a second mammalian genome (bovine) was used as a control. A summary of mapping results is shown in Table 3.1.

Table 3.1. Summary of whole transcriptome pipeline mapping results. Total read mapping percentages for both *A. nidulans* libraries are low compared the control library when mapped to their respective genomes. Both libraries also map the human and bovine genomes at nearly the same rate as the *A. nidulans* genome. While this may indicate contamination of the fragment libraries, the low percentage of reads mapped with high quality (align score >= 24) and similar number of hits in both human and bovine genomes suggests that this not be the case.

		Genome		
	Mapping	***A. nidulans***	**Human**	**Bovine**
S1	Total	16.4 %	16.7 %	15.3 %
	High quality	14.7 %	0.6 %	6.1 %
	Unique	14.6 %	0.2 %	1.7 %
S3	Total	13.4 %	10.1 %	14.2 %
	High quality	11.8 %	0.6 %	5.1 %
	Unique	11.7 %	0.2 %	1.4 %
Human (control)	Total	0.4 %	29.8 %	21.2 %
	High quality	0.1 %	20.1 %	8.3 %
	Unique	0.1 %	16.7 %	4.2 %

Contamination of the fragment library was the primary concern, as S1 and S3 mapped to both human and bovine genomes at a similar rate as to *A. nidulans.* However, the extremely low numbers of reads mapping with high quality, and/or uniquely mapping, indicates very poor mapping quality in each case. This low quality suggests that reads mapped to human and bovine are likely due to these genomes being sufficiently large that a high number of false, low quality read hits are inevitable. Mapping of the control library appears to support this hypothesis, as it also maps to the large bovine genome with relatively high frequency and quality, but not to the much smaller *A. nidulans* genome. It is also interesting that there is a much higher discrepancy between total mapped reads, high quality reads and uniquely mapping reads in the control. The latter figure being much closer to the percentages we achieve with our own libraries again indicates that false hits due to the size of the human genome play a major role in mapping profile.

The nature of colour space data and the mapping used make it difficult to ascertain exactly why a relatively low number of reads were mapped to the *A. nidulans* genome. While contamination cannot be completely ruled out, the data indicate that it most likely did not play a major role in affecting mapping quality, and that the low mapping frequency is simply due to the relatively small size of the *A. nidulans* genome. While it is vital to always strive for improved quality in RNA-seq libraries, S1 and S3 appear to be valid and of reasonable quality, despite showing relatively low rates of total mapping.

3.4. Analysis of RNA-seq data

Mapped RNA-seq data were visualised using the Artemis genome browser (Rutherford *et al*., 2000; Carver *et al*., 2012). To ensure data compatibility with Artemis and facilitate the analysis of coverage frequency at the nucleotide level, the mapped read data was converted to strand specific coverage plots by Dr Kevin Ashelford, using software of his own design (Chapter 2.5.2; Appendix 2.2). Visual analysis of the mapped data in Artemis reveals a number of issues which we aimed to address. As previously discussed, the majority of *A. nidulans* genes in this annotation were predicted *in silico*. Our data indicated that this approach was not comprehensive, with a significant number of genes excluded or incorrectly annotated. A large number of potentially novel genes was indicated by significant read mapping throughout intergenic regions, an example of which can be seen in Fig. 3.1. Differential expression between growth conditions was also observed (Fig. 3.2). Many genes appeared poorly annotated and/or overlapping in antisense configuration between the two strands. An example of this can be seen in genes AN0104.4 and AN0105.4 (Fig. 3.3). Transcription of these genes was shown to terminate far beyond the annotated 3′ ends, indicating both poor quality annotation and a previously unidentified antisense overlap between the two genes. As antisense pairs are often related, a BLAST search was performed for both genes. AN0104.4 demonstrates high sequence homology with yeast ALG3, which encodes Dol-P-Man:Man(5)GlcNAc(2)- PP-dolichyl alpha-1,3-mannosyltransferase, while AN0105.4 encodes a eukaryotic translation initiation factor eIF1a-like protein. A direct functional association between these two genes could therefore not be ascertained by these data alone, and while RNA-seq data provides a resource for the identification of antisense transcripts, such relationships must be proven experimentally.

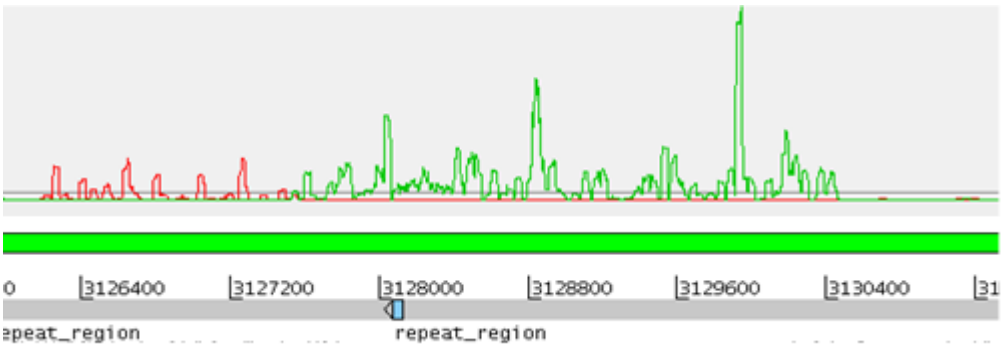


Figure 3.1. Observation of potentially novel genes. Mapped RNA-seq coverage for the forward (green graph) and reverse (red graph) strands are displayed in Artemis against the forward (green track) and reverse (grey track) strands of the *A.nidulans* genome assembly. Regions of coverage on each strand appears to indicate the presence of two genes in the region shown. The region appears to be intergenic in this version of the *A. nidulans* annotation, as no genes are displayed on either strand. These reads therefore represent potentially novel transcripts originating from unannotated genes in this area.

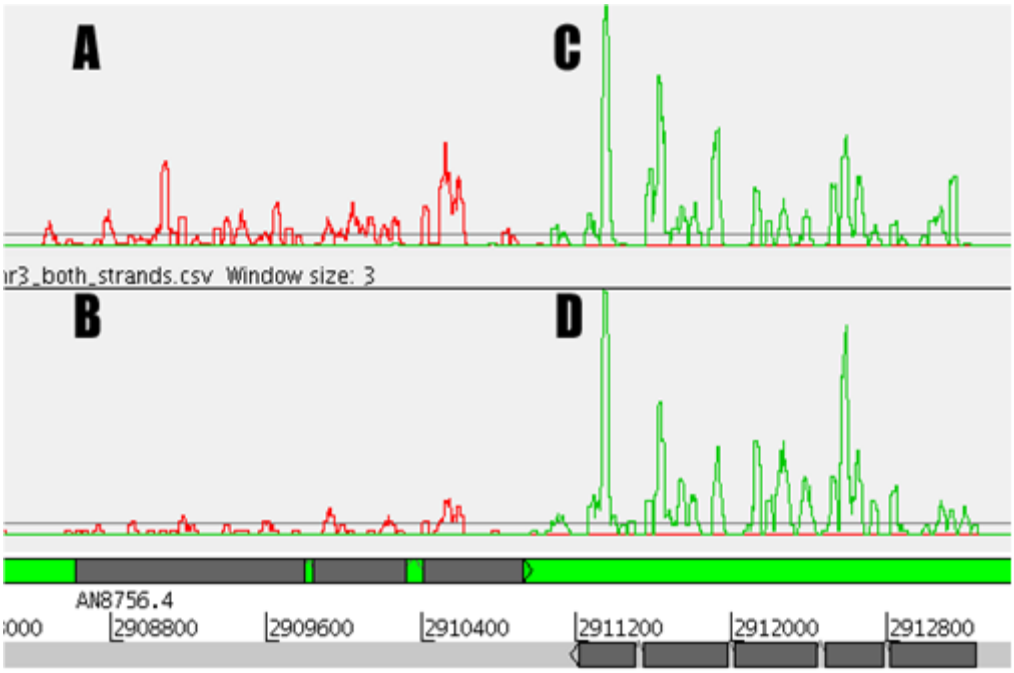


Figure 3.2. Artemis display of reads from two strains of *A. nidulans* grown in different conditions. Mapped RNA-seq coverage for the forward (green graph) and reverse (red graph) strands are displayed in Artemis against the forward (green track) and reverse (grey track) strands of the *A.nidulans* genome assembly, with annotated genes shown in dark grey. The top graph (A+C) represents the wild type strain grown overnight on complete media, while the bottom graph (B+D) was grown on minimal media supplemented with NO_3^-. Two genes are shown, one of which (AN8756.4) appears to be differentially expressed under these two conditions, while the second appears reasonably constant in both cases.

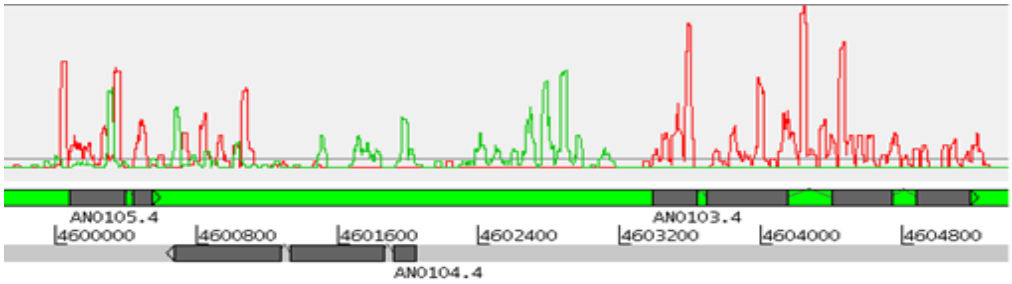


Figure 3.3. Demonstration of sense-antisense overlap. Mapped RNA-seq coverage for the forward (green graph) and reverse (red graph) strands are displayed in Artemis against the forward (green track) and reverse (grey track) strands of the *A.nidulans* genome assembly, with annotated genes shown in dark grey. Strand specific RNA-seq data indicates that transcription of AN0104.4 and AN0105.4 terminates significantly downstream from the annotated 3′ ends. The two genes are shown to exist in a sense-antisense overlap configuration, indicating a possible functional or regulatory relationship between these genes.

From visual analysis in Artemis, it was clear that reads were not mapping uniformly across their associated genes. These data instead presented as jagged, asymmetrical peaks, with extreme variation in depth of coverage across a given locus, making the precise level of individual gene expression difficult to determine. To address this issue, novel software was developed to determine the average depth of coverage across each gene, and used this as a comparative measure of expression. Using gene models from the CADRE *A. nidulans* genome annotation, the frequency of read alignment to the exon bases of each gene was extracted from the mapped RNA-seq data. These frequencies were combined and divided by the total number of exon bases, producing the average coverage per exon base (Chapter 2.5.3). Genes with an average read coverage greater than 1 read per exon base were reported as expressed, facilitating the identification of poorly expressed genes while being sufficiently stringent to avoid false positives due to random background mapping. Using this software, 61.0 % of annotated genes were shown to be expressed in RNA library S1, and 57.4 % in S3.

Due to the extent of potentially novel or poorly annotated genes, it was hypothesized that significantly higher levels of antisense transcription would be identified when using RNA-seq data than when comparing existing gene positions. Software was developed to estimate occurrences of antisense transcription from mapped RNA-seq data. This software is described in Chapter 2.5.4, and worked by extracting gene models from the CADRE *A. nidulans* genome annotation, then interrogating corresponding base positions on the opposite strand for RNA-seq read alignments using the pileup data produced by Dr K. Ashelford (as previously discussed). With its most lenient settings, this software indicated 1703 occurrences of antisense transcription in library S1 and 1393 in S3. The number of these with antisense coverage greater than 10 % of the annotated gene length was significantly lower, 498 for S1 and 464 for S3, representing approximately 5 % of annotated genes. Inspection of the number of hits per chromosome revealed a high level of consistency in the proportion of genes with

antisense transcripts, with no significant bias (Fig. 3.4). The full software output for both libraries including gene lists with associated antisense coverage can be found in Appendix 2.4. It should be noted that this 10 % cutoff was used solely to introduce an arbitrary lower limit of antisense coverage as a form of control, against which the most lenient settings of the software could be compared. This cutoff is far from ideal as it will undoubtedly have introduced a bias for shorter genes, highlighting the need for more robust methods of antisense detection in future work (see Chapter 6.7).

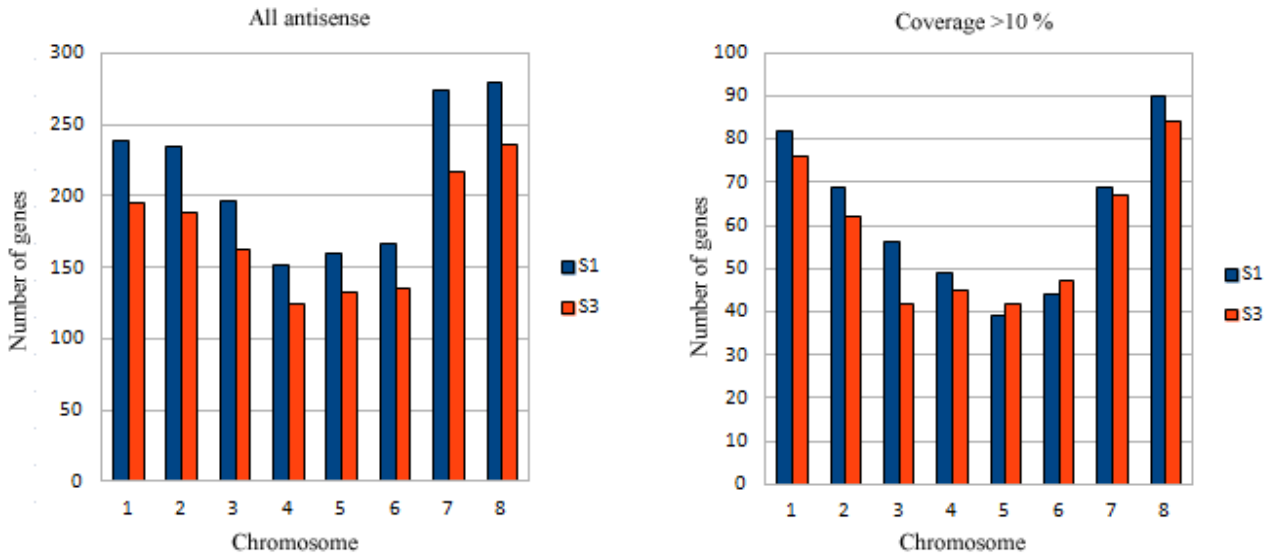


Figure 3.4. Histograms of estimated antisense transcript occurrence by chromosome. The number of genes with putative antisense transcripts identified using novel software and RNA-seq data is shown for libraries S1 (blue) and S3 (red). Datasets containing all genes exhibiting antisense transcription (left histogram) and only those with antisense transcription covering >10 % of the total gene length (right histogram) show similar patterns, and in each case the number of genes per chromosome was shown to be proportional to chromosome size.

3.5. Addressing fragmentation bias

The pronounced fragmentation bias evident in these libraries raised a major concern with the quality of the data produced by these libraries. A random, unbiased fragmentation would be expected to produce a far more even distribution of reads, where as RNA-seq data from S1 and S3 showed regions of both very high and very poor mapping within individual genes. The highly uneven distribution of reads across each gene made the data difficult to interpret when investigating gene structure or comparing expression levels between samples. One of the aims of this project was to present RNA-seq data to the *Aspergillus* community as a resource for further work. Therefore, data produced needed to be of the highest possible quality, prompting an investigation into the cause of this uneven mapping.

The protocol for SOLiD library preparation fragments RNA uses Ribonuclease III (RNase III), a double-stranded RNA specific endoribonuclease (Blaszczyk *et al*., 2001). This enzymatic fragmentation has the advantage of producing RNA fragments with the 5′ phosphate and 3′ hydroxyl groups required for SOLiD adaptor ligation. However, despite its ability to degrade a double-stranded substrate non-specifically, it has been shown to produce specific cuts in single-stranded RNA (Robertson *et al.,* 1968). This is due to the double-stranded specificity of the enzyme, which relies on the secondary structure of single-stranded substrates to provide dsRNA-like hair-pins at which it can act (Gan, 2005; Ji, 2008). The resulting structural bias of RNase III fragmentation translates through to the final library and the highly uneven mapping profile of the reads across each gene.

In an effort to bypass the issue of fragmentation bias, chemical shearing of RNA with divalent cations under elevated temperature was explored as an alternative fragmentation method. Metal ions have been shown to promote cleavage of RNA in aqueous solution.

Lanthanide ions, such as Eu^{3+},Tb^{3+}, and Yb^{3+}, are particularly efficient in catalyzing nucleotide cleavage, with rate acceleration of 3-4 orders of magnitude over the uncatalysed reaction (Breslow & Huang, 1991). Other metal ions such as Pb^{2+}, Zn^{2+}, and Mg^{2+} have also been shown to produce significant rate acceleration over the uncatalyzed reaction (Breslow & Huang, 1991). While not as efficient as Lanthanide ions, catalysis by these ions provides a more controlled fragmentation in a laboratory environment. Metal-catalysed heat fragmentation of RNA has been extensively used in sample preparation for microarray experiments (Browne, 2002) and is the preferred method of fragmentation in other high-throughput sequencing systems. Illumina utilise zinc acetate fragmentation in their RNA-seq protocols as they found "This is the most robust method for fragmentation of RNA. Fragmentation by this method has been shown to result in more uniform sequencing coverage compared with other methods" (Illumina, unpublished data).

Conditions for zinc acetate catalysed fragmentation of RNA have previously been optimised for the creation of 50-200 nt fragments in a number of systems (Vitale, 2001; Mehlmann *et al*., 2005) and this method is frequently used in sequencing applications (Gibbons *et al*., 2009; Vivancos *et al*., 2010). Metal ions are generally considered to cleave single-stranded RNA more effectively than double-stranded RNA (Hall *et al*., 1996; Husken *et al*., 1996; Kolasa *et al*., 1993; Zagorowska *et al*., 1998). However, the reaction temperature of 70°C denatures the secondary structure of the RNA, eliminating any potential for fragmentation bias.

A major drawback of this method is that it cleavage of RNA by divalent cations is usually preceded by an intramolecular transesterification of the internucleosidic 3′,5′-phosphodiester bond to a 2′,3′-cyclic phosphate and its subsequent hydrolysis to a mixture of 2′- and 3′-

phosphates (Zagórowska *et al.*, 1998). The resultant fragments therefore do not have the necessary 5′-phosphate and 3′-hydroxl groups for SOLiD adaptor ligation, and the ends must be repaired using T4 polynucleotide kinase (T4 PNK). This enzyme catalyzes the transfer of the γ-phosphate from ATP to the 5´-terminus of polynucleotides, and also functions as a 3′-phosphatase (Cameron & Uhlenbeck, 1977). This double function allows T4 PNK to produce the functional group required for SOLiD library preparation at each end of the RNA fragments.

Repaired RNA fragment libraries were used in conjunction with SREK to produce libraries for SOLiD sequencing. This proved unsuccessful, with RNA samples either degrading or being incompatible with some part of the library preparation protocol. This experiment was performed during a period in which the SOLiD system was still in its infancy and protocols changing rapidly. Library preparations were extremely temperamental during this period and the inability to produce a valid library using chemical fragmentation is most likely due to imperfections in the early SOLiD protocols and human error due to inexperience with the system.

Later attempts to use a chemical fragmentation in the preparation of RNA libraries for SOLiD sequencing were attempted without the use of SOLiD library preparation kits. An alternative protocol was developed using chemical fragmentation, oligonucleotides which matched the SOLiD primer sequences, and a random primer to perform reverse transcription. Synthesising cDNA prior to adaptor ligation allowed the use of DNA ligase instead of the standard SOLiD RNA ligation enzyme mix, and circumvents the requirement for a 5´-phosphate and 3´-hydroxl on the fragmented RNA. Removing the need for end repair aimed to decrease the number of steps in the protocol and the potential for sample degradation and loss. Libraries

created in this manner should also have provided more accurate and complete transcript mapping as the ligation compatibility of each fragment is not dependant on a T4 PNK reaction which will not have 100% efficiency.

RNA from one of the samples which had successfully been used in SOLiD sequencing was used to test the validity of this novel protocol. The RNA selected was from culture grown on minimal media with ammonia, however this was arbitrary as the purpose of this experiment was to provide proof of concept and the resulting libraries would not be sequenced. RNA samples underwent Poly(A) selection, DNA depletion with DNase I and were fragmented with zinc acetate. RNA was purified by ethanol precipitation and dissolved in 10 µl of nuclease free water.

First strand synthesis was performed by random priming with oligo "Random Priming P2 (barcoded)" (Chapter 2.1), and reverse transcription with KOD hot start polymerase (Novagen). RNase H (Fermentas) was used to selectively degrade the single-stranded RNA in RNA/cDNA duplexes created by first strand synthesis (Vorobjev & Zarytova, 2000). This reaction was first attempted using the manufacturer's protocol, however extremely low yields or complete loss of the fragment library was observed. Comparison of the first strand synthesis and RNase H reaction mixtures revealed that the pH of the Tris buffer was similar, and that the KCl, $MgCl_2$ and DTT required by the RNase H were present in the first strand synthesis reaction mix. While not at optimum concentrations, the presence of these components meant that RNase H reactions could be performed by adding the enzyme directly to the first strand synthesis reaction mix. The ethanol precipitation step prior to RNase H treatment was therefore removed to minimise the number of purification steps and potential for sample loss, while the heat deactivation step was also removed to prevent heat degradation of cDNA. Subsequent RNase H reactions were performed by adding the enzyme directly to the first strand synthesis reaction mix, achieving higher yields with no sample loss.

The presence of the active RNase H had no effect on subsequent reactions, and the enzyme was removed from the solution during downstream purification steps.

Following RNase H treatment, cDNA was isolated in solution with phenol:chloroform and purified by ethanol precipitation. To ensure the efficiency of hybridisation of the P1adaptor, two adaptor oligonucleotides were tested with different lengths of degenerate sequence overhangs. The complete, double stranded P1 adaptor was created by hybridising oligonucleotides P1 (comp) to either P1 random(6) or P1 random(10) (Chapter 2.1). Ligation of the adaptor to the cDNA was performed with DNA ligase, and proofreading Taq polymerase was used to synthesise the second strand. The resulting libraries were purified using a Minelute PCR purification kit (QIAGEN) and amplified by PCR with primers "P1 PCR" and "P2 PCR" (Chapter 2.1). The concentration of the cDNA was tested by spectrophotometer and the number of cycles chosen based on guidelines in the SOLiD library preparation protocol. No size selection was performed for these libraries as they were not to be sequenced and this step was the point at which the majority of problems arose in the creation of previous libraries.

The libraries were created with the N6 adaptor, N10 adaptor and control RNA (with N10 adaptor) were run on a 2% agarose gel next to a lane of Hyperladder V molecular weight marker (Bioline) before and after amplification. To produce comparable results, the same proportion of the sample was run on each gel. The resulting gels indicated successfully amplified libraries created using both adaptors, and a negative result from the –RNA control confirmed that the correct fragments were being amplified (Fig. 3.5).

The protocol was adapted to include the size selection step from the ABI SOLiD library preparation manual, and three new libraries and a negative control with no RNA were created with the completed protocol and the same starting RNA as the trial libraries (Fig. 3.6). As

both adaptors had produced successful libraries, the N6 adaptor was chosen for all future use as the shorter degenerate sequence had less chance of introducing mismatch errors into the finished library. Prior to PCR amplification, cDNA libraries were size selected alongside "Hyperladder V" molecular weight marker (Bioline) on 2% agarose gels (as indicated in the current version of SOLiD library creation kit protocols at the time). Gels were stained with ethidium bromide and visualised under UV. Visualisation of the gel indicated the presence of cDNA for each library, and a high proportion of fragments falling within the correct size range for SOLiD sequencing (Fig. 3.7). However, purification of the size selected library from the excised gel fragments proved extremely temperamental, with a high proportion of samples lost or degraded. This method of size selection had also been problematic in the creation of libraries with SREK. In an attempt to avoid these issues, different methods of size selection were investigated.

As an alternative to size selection on agarose gels, a FlashPAGE fractionator (Ambion) was used to isolate the required sizes of cDNA for SOLiD sequencing. This method had been recently released in updated SOLiD library preparation protocols for use with official SOLiD kits, and removed the need for a gel extraction step. Unlike previous size selection protocols, FlashPAGE fractionation was performed following the fragmentation step and before SOLiD amplified library construction in the method described in Fig. 3.6. The concentration of cDNA was tested by spectrophotometer prior to PCR amplification and the appropriate number of PCR cycles used. A second PCR amplification was performed with a higher number of cycles to ensure that any viable product had been sufficiently amplified.

Amplified libraries were visualised under UV on a 2% agarose gel stained with ethidium bromide (Fig. 3.8). However, libraries were deemed unsuitable for SOLiD sequencing.

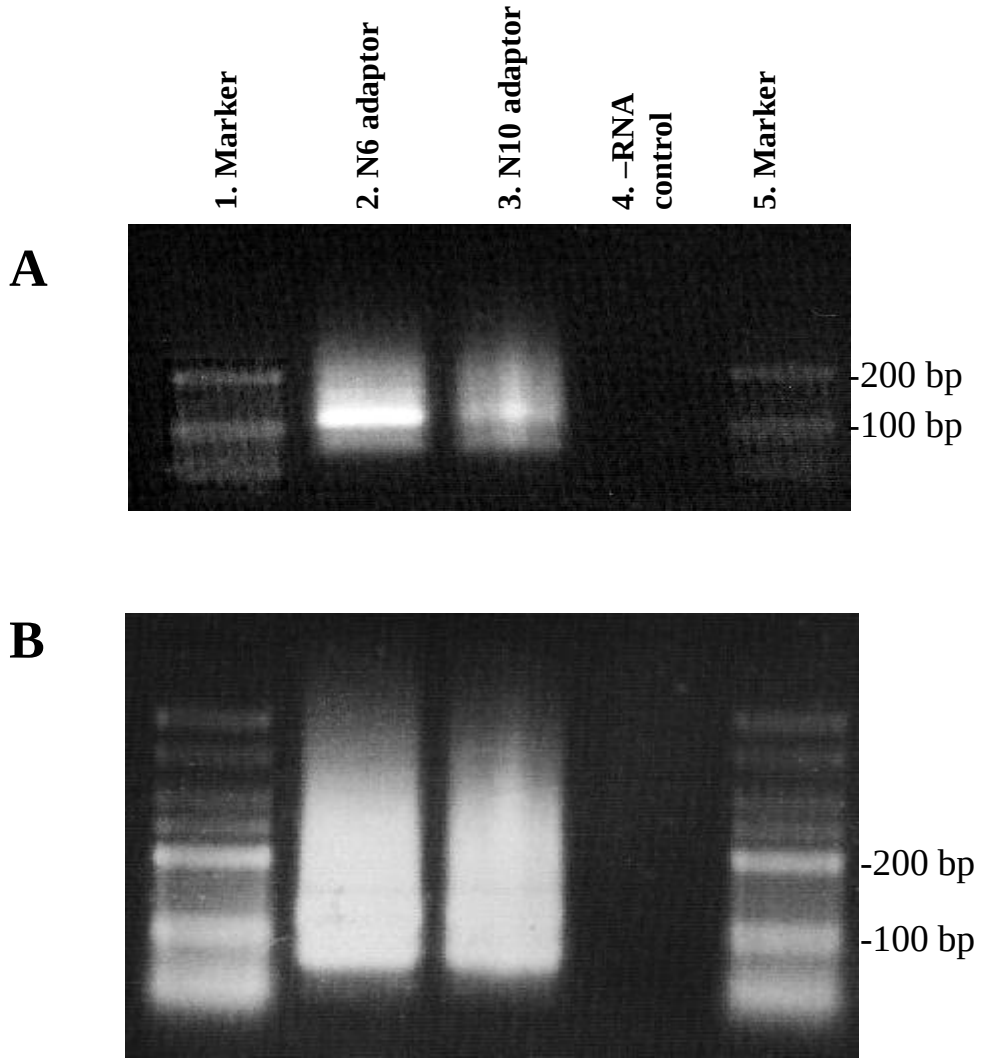


Figure 3.5. Trial libraries made with chemical fragmentation before (A) and after (B) PCR amplification. The same proportion of the library was run on each gel to produce comparable results. Streaks up the gel were due to these libraries not undergoing size selection. Both the N6 (lane 2) and N10 (lane 3) adaptors were shown to produce libraries which could be amplified by PCR, while the negative result produced by the –RNA control (lane 4) confirmed that it is the library being amplified rather than contaminants.

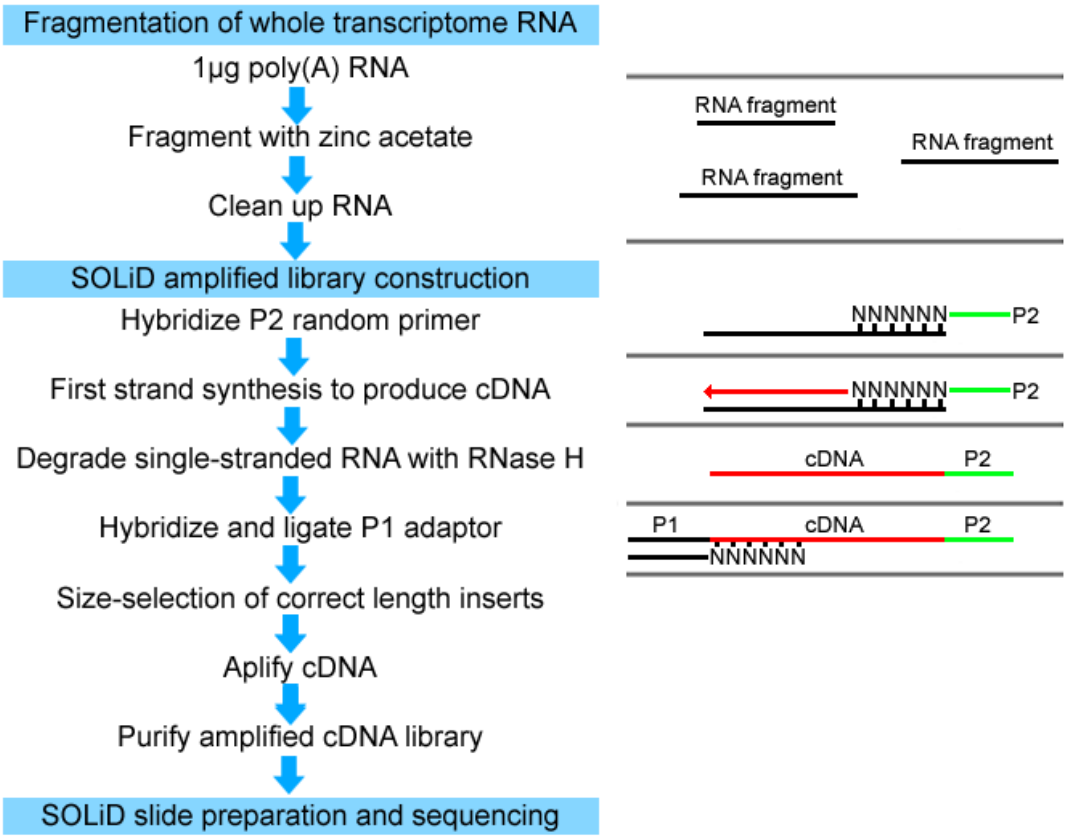


Figure 3.6. SOLiD library preparation protocol with zinc acetate fragmentation.

Hybridization of a P2 random primer followed by first strand synthesis generates cDNA bound to single-stranded RNA. producing single-stranded cDNA. The P1 adaptor is hybridized and ligated to the cDNA using DNA ligase and the incomplete strand removed by denaturing. Size selection of products 100-200 nt in length is performed by PAGE on 6% TBE-urea gels and PCR amplification used to increase the quantity of cDNA to that required for SOLiD sequencing.

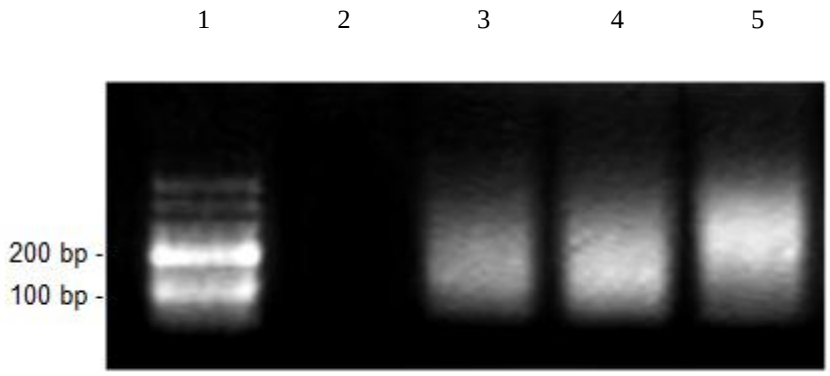


Figure 3.7. Size selection of amplified libraries on 2% agarose gel. Lane 1: Marker (Hyperladder V), Lane 2: Negative control, Lane 3-5: Amplified libraries. The streaks up the gel indicate a good fragmentation and show a large amount of product in the 100-200 nt size range required for SOLiD sequencing.

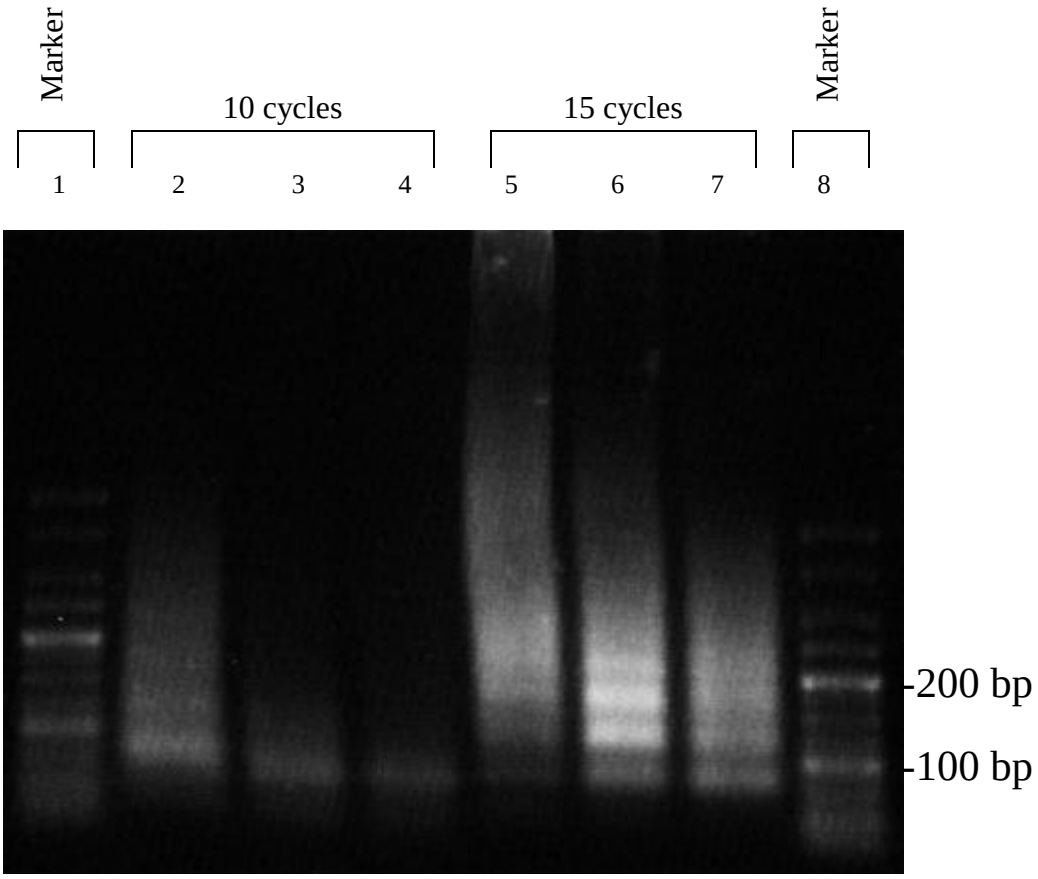


Figure 3.8. Amplified libraries created with chemical fragmentation and size selected by FlashPAGE. Libraries were amplified with either 10 or 15 cycles of PCR to ensure sufficient product. Bands of the appropriate size for SOLiD sequencing (150-200 bp) were present in some of the 15 cycle lanes (particularly lane 6), however the streaks up the gel suggest considerable contamination, most likely from primer complexes. None of the libraries were considered suitable for SOLiD sequencing.

Following advice from ABI representatives, development of this protocol was placed on hold as new SOLiD kits and protocols were soon to be released, and developments to the SOLiD sequencing platform would allow for deeper and more reliable sequencing. This promised greater coverage, giving a better representation of gene structure and minimising the issues caused by RNA fragmentation bias.

3.6. Discussion

The *Aspergillus nidulans* transcriptome has been sequenced on the SOLiD version 2 platform. Two whole transcriptome libraries prepared from G00 wild type cultures growth on nitrate with glucose and complete media were successfully mapped to the *A. nidulans* genome. The mapping percentage of these libraries did not appear to be optimal compared to the control, however there was no conclusive evidence that this was due to poor sample quality. It is more likely that the observed mapping figures were due to the relatively small size of the *A. nidulans* genome, resulting in fewer false positive hits due to repeat regions or random matches.

Gene expression analysis reported 60.1 % of genes expressed in *A. nidulans* cells grown on nitrate and glucose. Comparatively, 57.4 % of genes were expressed in cells grown on complete media. ~5% of annotated genes demonstrated a strong indication of antisense transcription. Natural antisense transcripts (NATs) have previously been shown to perform functional regulatory roles in RNA interference (RNAi), alternative splicing (Zhang *et al*., 2006), and influence transcription and protein activity (reviewed in Wilusz *et al*., 2009).

NATs are transcribed in intergenic regions and can vary in length from >200 to several thousand nt. They can be transcribed either *in cis*, meaning at the same genomic locus to the sense transcript and with perfect sequence complimentarity, or *in trans*, meaning at a different loci with imperfect sequence complimentarity. The majority of sense-antisense transcript pairs involve interactions between a coding transcript and a non-coding transcript, and are classified based on orientation of the overlapping region relative to the sense transcript. The four classes are; head-to-head (overlap at the 5' end), tail-to-tail (overlap at the 3' end), fully overlapping (the NAT covers the entire sense ORF) and embedded (the NAT is entirely covered by the sense ORF) (Lapidot & Pilpel, 2006; Su *et al*., 2010).

Interactions between sense-antisense pairs are able to regulate gene expression through a variety of mechanisms. The first of these is the formation of double stranded RNA (dsRNA), which can block active sites required for RNA-protein interactions (Beiter *et al*., 2009), preventing processes such as transport, splicing and polyadenylation. In the cytoplasm, the formation of dsRNA can also effect translation, rate of degradation, and in fungi has been shown to trigger the RNAi pathway (Faghihi & Wahlestedt, 2009). The second mechanism is chromatin remodelling, and NATs have been shown to recruit histone modifying enzymes such as histone methyltransferases to alter chromatin structure in fungi (Lapidot & Pilpel, 2006; Faghihi & Wahlestedt, 2009). This strategy requires the interaction of the NAT directly with the DNA, resulting in transcriptional regulation of the sense transcript (Munroe & Zhu, 2006). The extreme level of regulation achievable through this mechanism has been demonstrated in *S. cerevisiae* by the silencing of the highly expressed GAL1-10 cluster by NATs transcribed only once per hour (Gullerova & Proudfoot, 2010). Finally, NATs can cause transcriptional interference, where the simultanious transcription of two transcripts *in cis* interferes with the machinery of transcription, rather than the action of the NAT itself.

This interference can be due to either occlusion of the promoter or the two RNA polymerase II complexes blocking each other, preventing elongation of the transcripts (Werner *et al*., 2009; Lapidot & Pilpel, 2006). It has also been proposed that simultanious transcription of both strands places a topological constraint on the DNA and creates conditions which are unfavourable for transcription (Munroe & Zhu, 2006).

The discovery of NATs was met with some scepticism until recent years, and functional analysis has been limited considering the number that have been discovered through transcriptome analysis. The majority of experimental characterisation of NATS has currently been performed in *S. cerevisiae* (Donaldson & Saville, 2012), however global analysis has also been performed in *A. flavus*, leading to the discovery of 352 NATs, the majority of which were found to be expressed alongside a sense transcript, suggesting that these NATs may function post-transcriptionally (Smith *et al*., 2008). RNA-seq data indicated the rate of NAT occurrence in *A. nidulans* to be roughly consistent with those observed in *A. flavus* other systems (Yelin *et al*., 2003). This supports the hypothesis that natural antisense transcripts play an important role in the regulation of gene expression in *A. nidulans*.
690 intergenic regions with a large number of reads mapped to them were identified, representing either gross misannotation or putative novel genes. Further visual analysis also revealed many more examples of less extensive misannotations. This indicated that the *in silico* methods used to predict a large proportion of genes in this annotation were not comprehensive.

Attempts at modifying the SOLiD library preparation protocols to avoid fragmentation bias caused by RNase III were unsuccessful. Similar issues with size selection were reported by other parties preparing libraries for SOLiD sequencing at a meeting with representatives from

ABI in early 2009. More recent SOLiD protocols have moved away from agarose gels and now size select by PAGE on denaturing gels before the amplification step. All forms of gel extraction were also removed from the official protocols, with the gel fragments instead placed directly into the PCR reaction mix. This new method of size selection would have greatly benefited attempts to produce libraries using chemical fragmentation. Had successful SOLiD libraries been produced using the suggested protocol (Fig. 3.6), there was high probability that the degenerate sequence on the P1 adaptor would have introduced errors to the first 6 bases of the read due to imperfect matches between the degenerate sequence and cDNA during adaptor hybridisation and ligation. PAGE on denaturing gels would have removed the N6 adaptor sequence from the cDNA by denaturing and size selection. As the resulting single stranded cDNA would contain both adaptor sequences, PCR amplification could then have been performed without the need for second strand synthesis. This would have greatly simplified library creation and removed the potential for the introduction of errors in the first 6 bases of sequenced reads due to degenerate adaptor sequences. Despite the introduction of new library preparation kits and protocols, the use of RNase III and the associated fragmentation bias is a persisting issue. Work using the Illumina sequencing platform has reported similar coverage bias using chemical fragmentation, however work by Dohm *et al*. (2008) showed there to be no fragmentation bias using this method. Hansen *et al*. (2008) showed biases in Illumina transcriptome sequencing to be due to random hexamer priming. The fact that chemical fragmentation has been proven to be unbiased may means that alternative methods of RNA fragmentation may be worthwhile now that viable library creation and size selection protocols are more reliable. However, these findings mean that sequencing bias may still occur in SOLiD libraries created with an unbiased fragmentation method.

Constant and rapid developments in SOLiD sequencing technology caused this initial analysis to rapidly go out of date. The SOLiD version 2 sequencer was replaced by the version 4, allowing longer (50 bp) reads to be sequenced. Early sample preparation kits were replaced with the SOLiD Total RNA-seq kit (ABI), further improving the quality and reliability of library construction. Additionally, the release of new and/or improved software for mapping and handling whole-transcriptome data offered improved accuracy and more advanced downstream analysis. It was therefore beneficial to create new libraries using the most up to date SOLiD sequencing technology. Early work done with these initial samples was made largely redundant. However, it was used to guide future data analyses, as it had highlighted many points of interest which could be further and more accurately explored using these new libraries and software tools.

Chapter 4

4. 5′ specific RNA-seq

4.1. Identification of transcription start sites and functional promoter elements

Whole transcriptome RNA-seq has allowed for accurate transcript identification and gene structure prediction in relation to intron-exon junctions. However, the highly uneven nature of read mapping (discussed in Chapter 3.5) made the accurate identification of transcription start sites from these data alone extremely difficult and imprecise.

To supplement the whole transcriptome RNA-seq data already produced on the ABI SOLiD platform, a protocol for 5′ specific RNA-seq was developed for accurate, genome-wide identification of transcription start sites (TSS).

TSS identified by this analysis were used to define promoter regions for the identification of functional promoter elements. Previous research has highlighted the importance of conserved DNA motifs in the regulation of downstream pathways in a wide variety of organisms (Kellis *et al*, 2003; Conlon *et al*., 2001). Such regulatory sequences within promoters have been shown to control both the activation and repression of multiple pathways, functioning as regulatory elements in eukaryotes (Bregman *et al*., 2011; Trcek *et al*., 2011). These sequences are usually short, 6-15 bp in length, follow few rules and can tolerate some degree of sequence variation, making them much harder to detect than genes. Early examples were generally detected by experimental manipulation of individual promoter regions, however computational analysis has been successfully employed to identify regulatory elements associated with sets of related genes (Bailey & Elkan, 1994; Tavazoie et al., 1999; Stormo, 2000). Cross species sequence alignments have also been used to study promoters, and have been applied across the entire human and mouse genomes (Mouse Genome Sequencing Consortium, 2002), and in four species of *Saccharaomyces* (Kellis *et al*., 2003) to identify

regulatory motifs accross the entire genomes. Examples such as the binding site for the Gal4 transcription factor, which regulates genes involved in galactose metabolism, were highly conserved, with four experimentally validated sites showing perfect conservation between all four species of *Saccharaomyces* (Keegan *et al*, 1986; Kellis *et al.*, 2003). The Gal4 binding site motif was also shown to have a higher conservation rate in intergenic regions and upstream of flanking genes, while no such preferences were observed for control motifs (Kellis *et al*., 2003). A significant amount of work has also examined conserved functional regulatory elements in the promoters of numerous Aspergilli, including *A. nidulans (*David *et al.*, 2008). These include the binding sites for the CREA repressor responsible for carbon catabolite repression (Kulmburg *et al*., 2006), the PacC transcription factor required for regulation of the isopenicillin N synthase (*ipnA*) gene encoding a key penicillin biosynthetic enzyme (Espeso & Penalva, 1996), and the GATA transcription factor binding site (Ravagnani *et al*., 1997; Scazzocchio, 2000). Members of the GATA class of transcription factors have been shown to control a wide range of major physiological process in fungi, including development (Chae *et al*., 1995), circadian rhythm and blue light response (Ballario *et al*., 1996) and nitrogen metabolism (Fu & Marzluf, 1990; Kudla *et al*., 1990; Haas *et al*., 1995). Global analysis of conserved regulatory motifs is therefore vital to the understanding of transcriptional regulation, and mRNA expression data from microarrays has previously been used to discover regulatory motifs upstream intergenic regions in *S. cerevisiae* amongst other organisms (Roth *et al*., 1998; Tavazoie *et al*., 1999; McGuire & Church, 2000). RNA-seq data should therefore prove to be an excellent resource for global identification of conserved regulatory motifs in *A. nidulans*.

4.2. Development of 5′ specific RNA-seq

Early attempts to produce 5′ specific RNA-seq libraries on the SOLiD platform utilised a number of techniques previously employed during the exploration of alternative fragmentation methods for SOLiD library creation (chapter 3.5). RNA fragmentation with zinc acetate had previously been shown to produce fragments with 5′ and 3′ functional groups which were not compatible with SOLiD adaptor ligation protocols without further modification. Using this fragmentation method followed by removal of the 5′ cap structure with TAP created fragment libraries where only fragments representing the 5′ ends of transcripts were compatible for 5′ SOLiD adaptor ligation. However, due to the 3′ ends of these reads still possessing the incorrect functional group, only the 5′ adaptor would be added to these fragments. The 3′ adaptor therefore had to be added via an alternative method.

Creation of successfully PCR-amplified cDNA libraries had previously been achieved using randomly-primed first strand synthesis. This process utilised single stranded oligonucleotides comprised of a 5′ sequence complimentary to that of the SOLiD 3′ adaptor sequence, and a 3′ degenerate sequence (Fig. 4.1). This technique would perform first strand synthesis for all RNA fragments, regardless of functional groups. However, only those fragments representing 5′ ends would produce cDNA with the correct SOLiD adaptor sequences at each end. The degenerate sequence at the 3' end of the oligonucleotides was originally a 6N hexameric sequence as described by Fritz *et al.* (1991). This degenerate sequence was later amended to NNNNNC-3', as the addition of a cytosine residue to the 3' end has been shown to aid efficiency of hybridisation, as described by Hashimoto *et al.* (2004) in their protocol for 5'-end serial analysis of gene expression (5' SAGE).

This method was used to produce cDNA libraries which were then amplified using custom made primers as previously described. However, none of the libraries produced by this

method were suitable for SOLiD sequencing, with very little product recovered after cleanup from size selections.

5'- CTGCCCCGGGTTCCTCATTCTCTAAGCCCCTGCTGTACGGCCAAGGCGNNNNNC -3'

Figure 4.1. Design of Random priming P2 (barcoded) oligonucleotide for first strand synthesis. The sequence in black corresponds to that of the SOLiD 3' adaptor sequence, of which the blue sequence is the barcode. The degenerate sequence used for random priming is shown in red, with the addition of a cytosine residue at the 3' to aid efficiency of hybridisation. N represents any nucleotide (A,T, G or C).

4.3. Modifying the 5′ specific RNA-seq protocol to reduce sample loss

Attempts to produce alternative methods for SOLiD library preparation up to this point had all focused on reproducing a sequence of steps similar to those in the ABI SOLiD library preparation protocols. However, experience with SOLiD library preparation indicated that library quality decreased in accordance with the number of steps used. Kits and protocols for standard SOLiD library preparation had also evolved significantly by this point and were proving far more robust. The 5′ specific RNA-seq protocol was therefore revised to involve the fewest possible number of steps, and to utilise as many components of the SOLiD Small RNA Expression Kit as possible. These changes aimed to reduce sample loss while making library preparation as true to the standard protocol as possible in order to benefit from the increased reliability of these methods and reagents.

Evaluation of the 5′ specific RNA-seq protocol revealed that fragmentation of the RNA was not necessary. The random nature of the primer used for first strand synthesis naturally created a library of cDNA fragments with variable lengths from full length RNA substrates. The fragmentation step could therefore be removed and cDNA libraries created directly from full length mRNAs, greatly reducing the potential for sample loss and degradation caused by fragmentation and subsequent cleanup procedures.

To minimise sample loss, the different purification kits recommended by various official SOLiD protocols were tested on a non-size selected cDNA library created using the 5′ specific RNA-seq protocol as previously described, but without a fragmentation step. Equal quantities of the library were subjected to PCR cleanup using either the MinElute PCR purification kit (QIAGEN) or Purelink PCR micro kit (Invitrogen). The Purelink kit appeared to be superior for the removal of oligonucleotides from the sample, however, less product was observed in the 100-200 bp size range required for SOLiD sequencing (Fig. 4.2). Despite

the lower amount of product in the correct size range, the Purelink kit was selected for the purification of amplified SOLiD libraries due to the high level of purity required for libraries to pass quality controls for SOLiD sequencing.

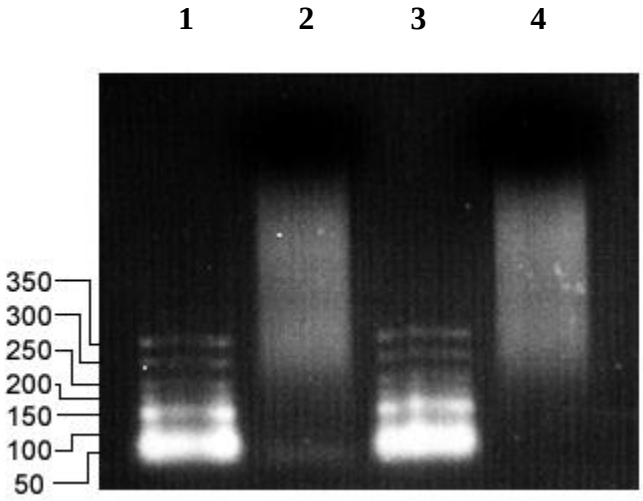


Figure 4.2. PCR cleanups with Minelute and Purelink kits. PCR cleanup products were run on agarose gels alongside a 50 bp size marker (lanes 1 and 3). The Minelute kit (lane 2) appeared to have slightly higher product concentration in the 100-200 bp size range required by SOLiD sequencing, but a clear band was visible at the bottom of the gel representing residual oligonucleotides in the sample. These did not appear in the products from the Purelink kit (lane 4), indicating a more robust cleanup protocol.

The number of cycles used for PCR amplification was also investigated. Previous unsuccessful attempts at library creation had led to the use of an increased number of PCR cycles in an attempt to produce a greater amount of product for SOLiD sequencing. Libraries which had not undergone size selection were again used for this experiment, as the size selection had been identified as the point at which most sample loss occurred. To investigate the effect of over-amplification, equal amounts of cDNA were amplified using either 20 or 25 cycles of PCR performed according to SOLiD SREK protocols. The maximum number of cycles recommended by the SREK was 18, so these values were chosen to demonstrate the result of both slight and severe over-amplification. PCR products were run on 2% agarose gels, stained with ethidium bromide and visualised under UV. 25 cycles appeared to produce significantly more product overall, as would be expected. However, a large proportion of this product had failed to run and was seen at the top of the gel. This resulted in less product appearing in the 100-200 bp size range required for SOLiD sequencing compared to the 20 cycle PCR products (Fig. 4.3). Over-amplification of SOLiD libraries was therefore counter-productive, and future attempts at 5′ specific RNA-seq used the standard number of cycles suggested in the SOLiD SREK protocols.

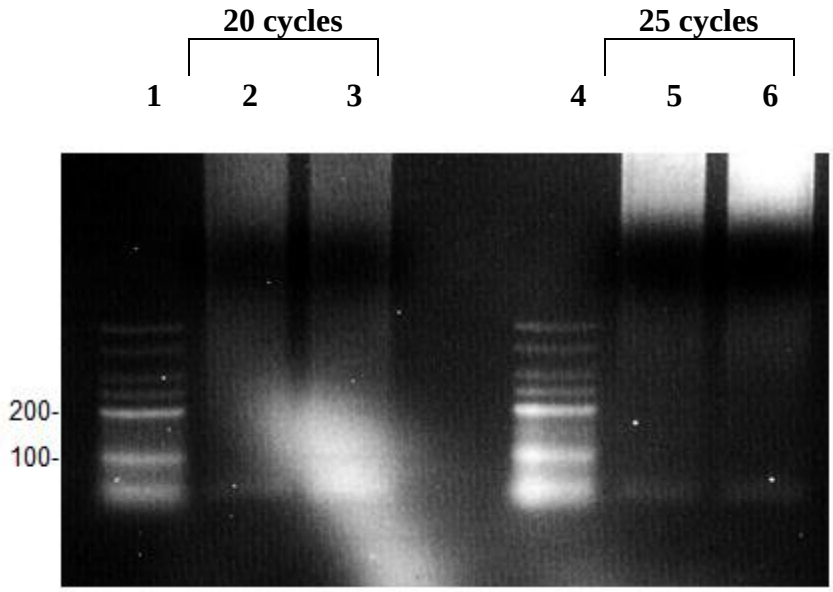


Figure 4.3. Effect of PCR over-amplification on SOLiD libraries. cDNA libraries were amplified by either 20 (lanes 2 and 3) or 25 cycles (lanes 5 and 6) of PCR, representing slight and gross over-amplification respectively. Hyperladder V molecular weight marker (lanes 1 and 4) was used to determine product sizes. A lower concentration of product in the 100-200 bp size range was observed with a greater number of PCR cycles, and much of the product appeared to be stuck at the top of the gel. Further library preparations were therefore performed with the minimum number of PCR cycles to warrant efficient recovery of fragments in the appropriate size range for SOLiD sequencing.

4.4. 5′ specific RNA-seq library

Advancements in the 5′ specific RNA-seq protocol proved successful in producing a library which passed quality control and was sequenced on the SOLiD version 3 Plus platform. To prepare this library, total RNA was extracted from G00 wild type *A. nidulans* grown on minimal media supplemented with nitrate and glucose. Total RNA was DNA depleted with DNase I, enriched for mRNA with poly(A) selection using oligo(dT), and the 5′ cap structure removed with TAP to facilitate 5′ adaptor ligation. No fragmentation step was performed prior to the hybridisation and ligation of adaptors from the SOLiD SREK, according to SREK protocols. First strand synthesis was then performed using a random priming P2 oligonucleotide (Fig. 4.1) as previously described. The downstream size selection required for fragmented libraries could be used to equal effect in size selection of libraries created in this fashion, and was performed according to standard SREK protocols using 6% TBE-urea polyacrylamide gels. The PCR amplification step was also performed using reagents from the SOLiD SREK and in accordance with standard SREK protocols and suggested number of PCR cycles. This successful protocol for 5′ specific RNA-seq library preparation is outlined in Fig. 4.4.

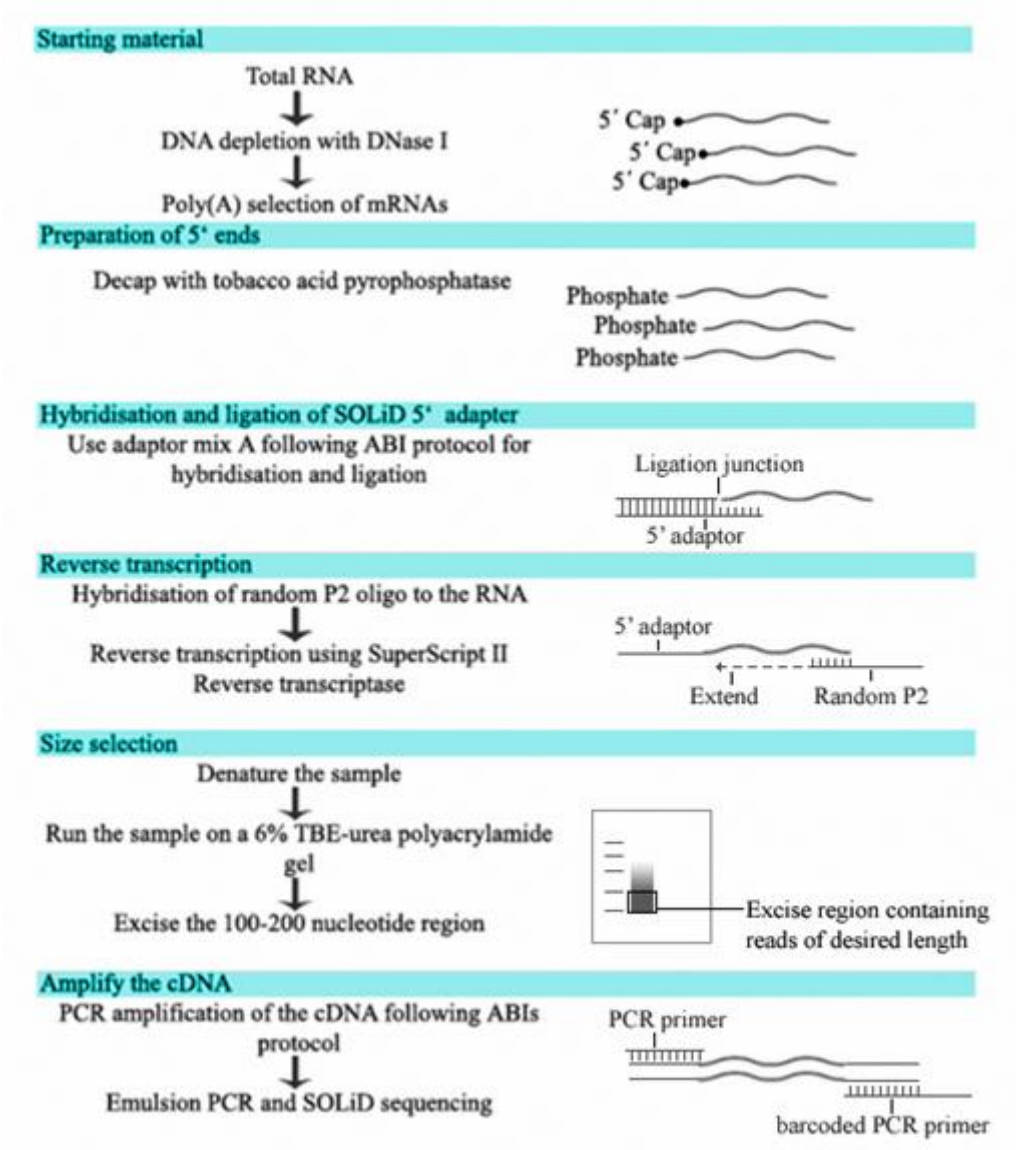


Figure 4.4. Protocol for 5′ specific RNA-seq on the SOLiD platform. Total RNA was extracted from *A. nidulans* culture and DNA depleted. Enrichment for mRNAs was performed by poly(A) selection using oligo(dT). 5′ caps were removed by TAP to allow subsequent ligation of the SOLiD 5′ adaptor. The P2 3′ adaptor was then added by reverse transcription using random priming, and size selection performed on denaturing polyacrylamide gels. Libraries were then amplified according to standard SREK protocols and sequenced on the ABI SOLiD platform.

Despite keeping as close as possible to standard SOLiD library preparation procedures and using the most robust kit for the cleanup of PCR products, quality control checks on an Agilent 2100 bioanalyser found the library to contain what appeared to be several artefacts producing large peaks at sizes 100, 150 and 253 bp (Fig. 4.5). These artefacts had the potential to interfere with the efficiency of the emulsion PCR step of SOLiD sequencing. However, so long as some usable data was produced it could be used as proof of concept, and further work could be done to produce new libraries with fewer artefacts and greater sequencing potential. Therefore, sequencing of the library went ahead, accepting this potential reduction in efficiency.

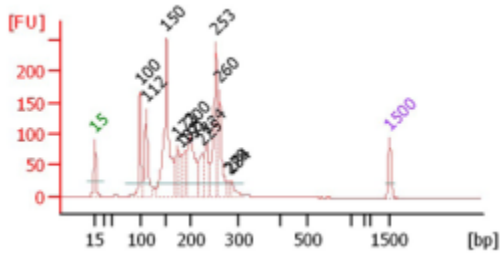


Figure 4.5. Agilent 2100 bioanalyzer plot of 5′ specific RNA-seq library. The library showed correct distribution for SOLiD sequencing, however artefacts were observed in the form of large peaks with sizes of 100, 150 and 253 bp. These had the potential to reduce the efficiency of the emulsion PCR step of SOLiD sequencing. Standard size markers were included at 15 and 1500 bp to ensure accuracy.

4.5. Mapping 5' specific RNA-seq reads to the *A. nidulans* genome

For 5' specific RNA-seq, the junction finding features of Tophat were not required as the main concern is the precise mapping of 5' ends. It was therefore decided to map reads to the *A. nidulans* genome with ABI's Bioscope software. Unlike Tophat, Bioscope was specifically designed to handle SOLiD colour space read data and uses progressive mapping to gradually trim the 3' ends of reads until optimal mapping is achieved. In the context of this experiment, Bioscope's method of read alignment should provide greater depth of coverage compared to the same data mapped with Tophat.

5' specific RNA-seq was performed as previously discussed, producing 64 million reads. Reads were mapped with Bioscope by members of the Liverpool Centre for Genomic Research (CGR), and14 million (22%) were aligned to the *A. nidulans* genome. This figure was similar to that achieved by whole transcriptome RNA-seq libraries, indicating that the artefacts observed during quality control checks did not produce a significant reduction in library quality. The data was therefore considered to be of suitable quality for further analysis.

At its most basic level, 5' RNA-seq data could be displayed in a genome browser and provided an excellent resource for visual interrogation of the *A. nidulans* transcriptome alongside reads from conventional whole transcriptome RNA-seq. Mapping of the 5' ends of transcripts allowed quick and accurate assessment of transcription start sites, and differentiation of individual genes in regions where genes are tightly packed and may have previously been annotated as a single gene model, or where a single gene has been incorrectly annotated into multiple gene models. These combined data have were used for the correction of several such misannotations (e.g. Fig. 4.6)

Figure 4.6. Correction of mis-annotation. Mapped RNA-seq coverage for the forward (green graph) and reverse (red graph) strands are displayed in Artemis against the forward (green track) strandof the *A.nidulans* genome assembly and the annotated exons in this region (dark grey). In this annotation (AN.4), two genes (AN10053.4 and AN10071.4) are predicted in this region. Whole transcriptome RNA-seq (lower graph) revealed the general structure of transcripts produced in this region. However, the uneven nature of the mapping makes accurate identification of the 5’ end difficult. 5’ specific sequencing (top track) gives a far clearer representation of the 5’ end. Re-annotation of this region using the combined whole transcriptome and 5’ RNA-seq data reveals that AN10053.4 and AN10071.4 lie within a single gene (bottom track) starting significantly upstream of previous estimates, including both previously annotated exons from both annotated genes (highlighted in blue), and two additional exons which had not previously been annotated (highlighted in pink). All RNA-seq data was produced from RNA extracted from *A. nidulans* culture grown on minimal media with nitrate (as previously described).

4.6. Conversion of mapped read data to readheads

While 5' specific RNA-seq provided an excellent resource even in its basic state, the format was not ideal for automated high-throughput interrogation of transcription start sites. While genome browsing software was available which could show individual reads aligned to the genome, the most useful format of data was the coverage plot such as those displayed in Artemis (e.g. Fig. 4.6). The coverage plot represents the number of reads aligned to each base position, and takes into account any alignment throughout the length of the read. However, the nature of 5' RNA-seq was such that only the aligned position of the first base in each read was significant.

In order to accurately assess putative transcription start sites, it was beneficial to alter the format of the mapped data from a series of 50 base pair reads to one which represents only the 5′ end of each read. This process identified 5′ ends from a pileup of aligned reads. By searching pileup data for each chromosome from 5′ to 3′, the number of reads starting at each position was calculated. Alignment frequencies of the following 49 positions were then reduced by the same number, so that only the 5′ end positions remained. Moving sequentially along each chromosome from 5′ to 3′ meant that each base position had any non-5′ end alignments removed before being processed, ensuring accurate representation of 5′ end positions and mapping frequencies in the resulting pileup. This process was designed in collaboration with Dr H. Wu, who developed novel software to perform this task. Appendix 2.3 contains software produced by Dr H. Wu for RH identification and all future elements of this analysis, as well as a description of software utilisation (provided by Dr Wu).

The chromosomal position of the first mapped nucleotide at the 5′ end of each read was identified as the readhead (RH). From this we were able to find the frequency of RH mapping

to each base position and convert read coverage data to a new format listing the number of reads and therefore the number of transcripts which start at each base Fig. 4.7.

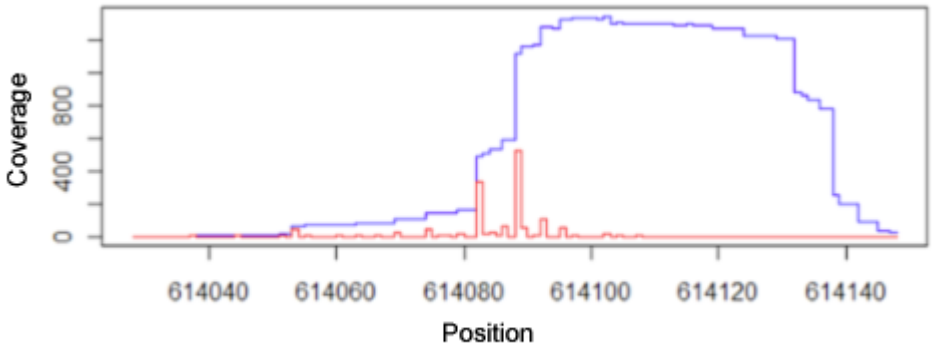


Figure 4.7. Conversion of mapped read alignments to RH frequencies. Standard RNA-seq read coverage data (blue) is shown against base position. The start position of each read was identified and mapped independently of other bases to convert this to RH coverage (red), which mapped only the base at the 5′ end of each read. Converting the data in this way greatly simplifies visualisation and further manipulation of TSS data.

Visualisation of both standard and RH mapping data in Artemis revealed a significant amount of background mapping. Many bases throughout the genome had been aligned with 1-2 reads, a number far too low to be considered a candidate for a transcription start site. While the majority of these bases appeared to map randomly, a large number were noted within annotated genes and shortly downstream of well mapped transcription start sites, possibly representing the 5′ ends of partially degraded mRNAs. This posed a problem when designing novel software for downstream analysis of the RH data, as such software might have identified low coverage positions as potential transcription start sites. To circumvent this issue, these positions were filtered out prior to any further analysis.

Filtering of any nucleotide position with low RH mapping frequency was performed using novel software to change the RH frequency to 0 at any base where the RH frequency was below a given threshold (Chapter 2.5.5). Various minimum RH values were tested and the resultant RH mapping files inspected visually in Artemis to determine an optimum cut off. A minimum RH value of 3 was found to eliminate the majority of background noise, while having a negligible impact on mapping around sites of high RH frequency.

Having removed background mapping, the remaining positions to which readheads were aligned represented the real 5′ ends of sequenced transcripts. This allowed regions of significant RH clustering to be more easily identified as potential transcription start sites.

4.7. Identification of TSS regions

Previous studies into the internal structure of core promoters have shown that transcription start sites are not absolute positions within the promoter. The FANTOM3 (functional annotation of mouse 3) project, applied Cap analysis of gene expression (CAGE) methods (Kodzius *et al*., 2006; Shiraki *et al*., 2003) to 20 tissues from mouse and human (Carninci *et al*., 2005; Carninci *et al*., 2006). In many cases, initiation of transcription was found to occur at multiple nucleotide positions within a core promoter region. This suggested that most core promoters do not have a single TSS, but rather a number of closely located initiation sites. These sites form distinct TSS regions and are conceptually different from alternative promoters, in which core promoters are separated by clear genomic space.

Visual inspection of the filtered RH data alongside standard whole transcriptome data and the current gene model annotation showed excellent correlation between regions of RH mapping and the annotated 5′ ends of genes. The majority of RHs mapped in clusters close to predicted start sites, often with a distribution of multiple large peaks surrounded by a range of positions of lower RH frequency. This supported the hypothesis that transcription start sites are not confined to a single base, instead presenting as regions in which transcription can be initiated. In an attempt to characterise these regions, a sample of 100 genes with high levels of RH mapping at the 5′ end were selected at random, and RH distribution investigated for the start site region of each gene.

Selecting the nucleotide with the highest RH mapping frequency as the predominant transcription start site for each region, the distribution of RH mapping both upstream and downstream of this position was investigated. In 98% of cases nucleotides with substantial levels of RH mapping were confined to within a 60 bp range in either direction from the predominant start site. In those cases where mapping was observed beyond this 60 bp limit,

the frequency of RH mapping was extremely low. While transcripts started at these positions are probably real, their apparent scarcity beyond this point meant that they were disregarded during further analysis, which focused on the TSS of more common transcripts. Therefore, a region of 121 bp centred on each predominant TSS was defined to investigate the spread of RH distribution on a global scale. These regions were identified on a global scale by Dr H.Wu.

4.8. Global analysis of RH frequency

To investigate the distribution in readhead mapping frequency to each TSS region on a global scale, the total RH frequency for each region was calculated. These RH frequencies represented the relative expression levels of transcripts originating from these regions and were analysed globally across chromosomes.

The RH frequencies of each position in a 121 bp TSS segment were combined to produce a total RH frequency for each region. These figures were used to generate box plots for all TSS regions identified on a chromosome. Performing this analysis for the first 4 chromosomes in *A. nidulans* revealed a high level of consistency between the statistics and near symmetrical normal distribution for each chromosome (Fig. 4.8). A large number of outlying positions of abnormally high RH mapping frequency were observed, many of which persisted even when the minimum and maximum thresholds were doubled from 1.5* to 3.0* the interquartile range (IQR). These outliers represent loci where transcription was found to initiate at an exceedingly frequent rate in statistical relation to other regions. While the majority of these outlying RH frequencies were most likely associated with genes with very high levels of expression, a low number of extremely high RH frequencies were observed, with some in excess of 100,000 RHs mapped. While these may simply have been associated with genes

that are very highly expressed, the extent of the frequency difference observed between these and the majority of other regions in the sample suggest that there may have been some experimental bias for reads mapped to these positions.

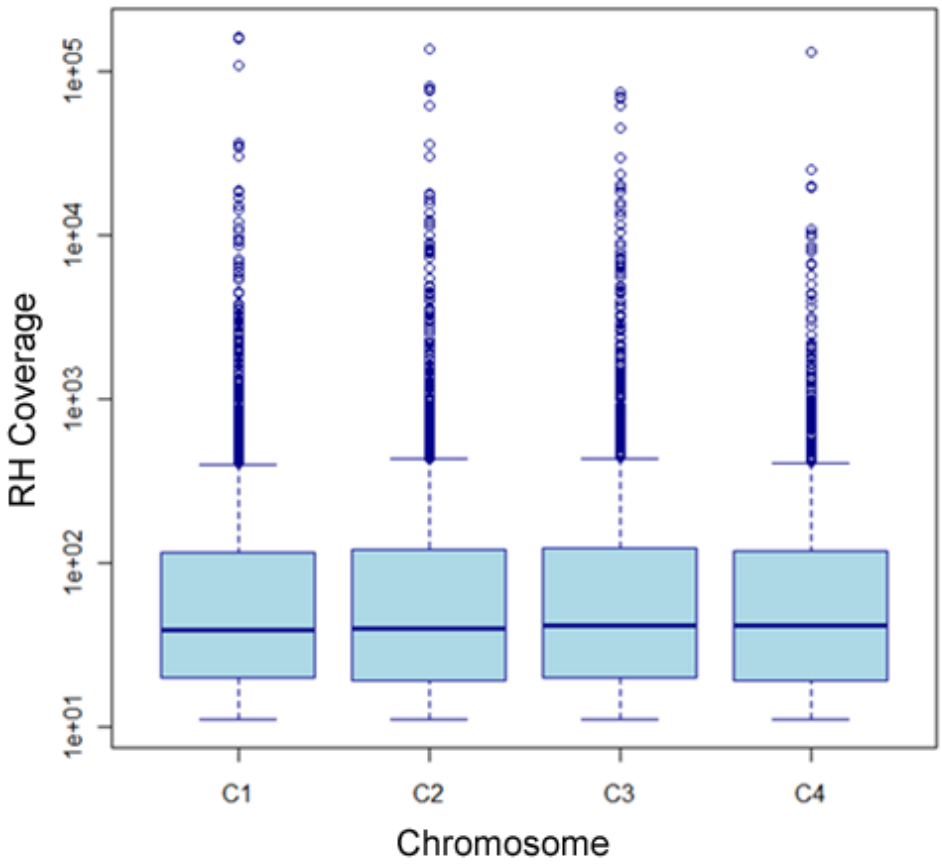


Figure 4.8. Box-plots of RH peak distribution for 4 chromosomes of *A. nidulans*. The five-number summaries for each chromosome are as follows: C1 (11, 20, 39, 115, 395), C2 (11, 19, 40, 121, 429), C3 (11, 20, 41, 124, 434), C4 (11, 19, 41, 117, 410). These numbers correspond to the five lines for each box-plot and represent the minimum, first quartile (Q1), median, third quartile (Q3), and maximum RH mapping coverage. A base position with coverage above the maximum is seen as an outlier indicated by dots. A high level of consistency between each chromosome and a bias for larger RH coverage is observed, with a large number of outliers appearing above the maximum level within the normal distribution, some peaks having >100,000 RH coverage.

4.9. Global analysis of TSS distribution within segments

Comparing total RH counts between regions gave a good indication of the variance in expression levels between start sites. However, the distribution of these readheads within the region itself was still unknown. Previous manual analysis of a sample of 100 well defined start sites indicated that readheads mostly fell within a 121 bp region centred on the highest peak. To confirm this observation and gain a deeper understanding of RH distribution within TSS regions, the RH frequency at each of the 121 bp positions within regions were summed up for all regions on a chromosome. To allow all regions from both strands to be compared simultaneously, all sequences were placed in 5′-3′ orientation. Regions which had previously been identified as outliers in total RH frequency (Chapter 4.8) were excluded as the extremely high RH frequencies in these regions were so large that they may have introduced bias to the RH distributions.

For ease of visualisation, RH positional distribution frequencies were combined into 13 bins to produce a graphical representation of global RH distribution across the regions in a chromosome. Chromosome 1 showed near normal distribution with a very strong central peak as expected due to this bin containing the primary TSS. Other abnormally large peaks away from this position would therefore represent secondary strong TSS, however RH frequencies trailed off towards the 0 and 121 bp positions. This distribution was consistent with previous findings regarding the spread and pattern of transcriptional initiation at start site regions. The analysis was then repeated for chromosomes 2, 3 and 4, all of which showed a near identical distribution pattern to chromosome 1 (Fig. 4.9).

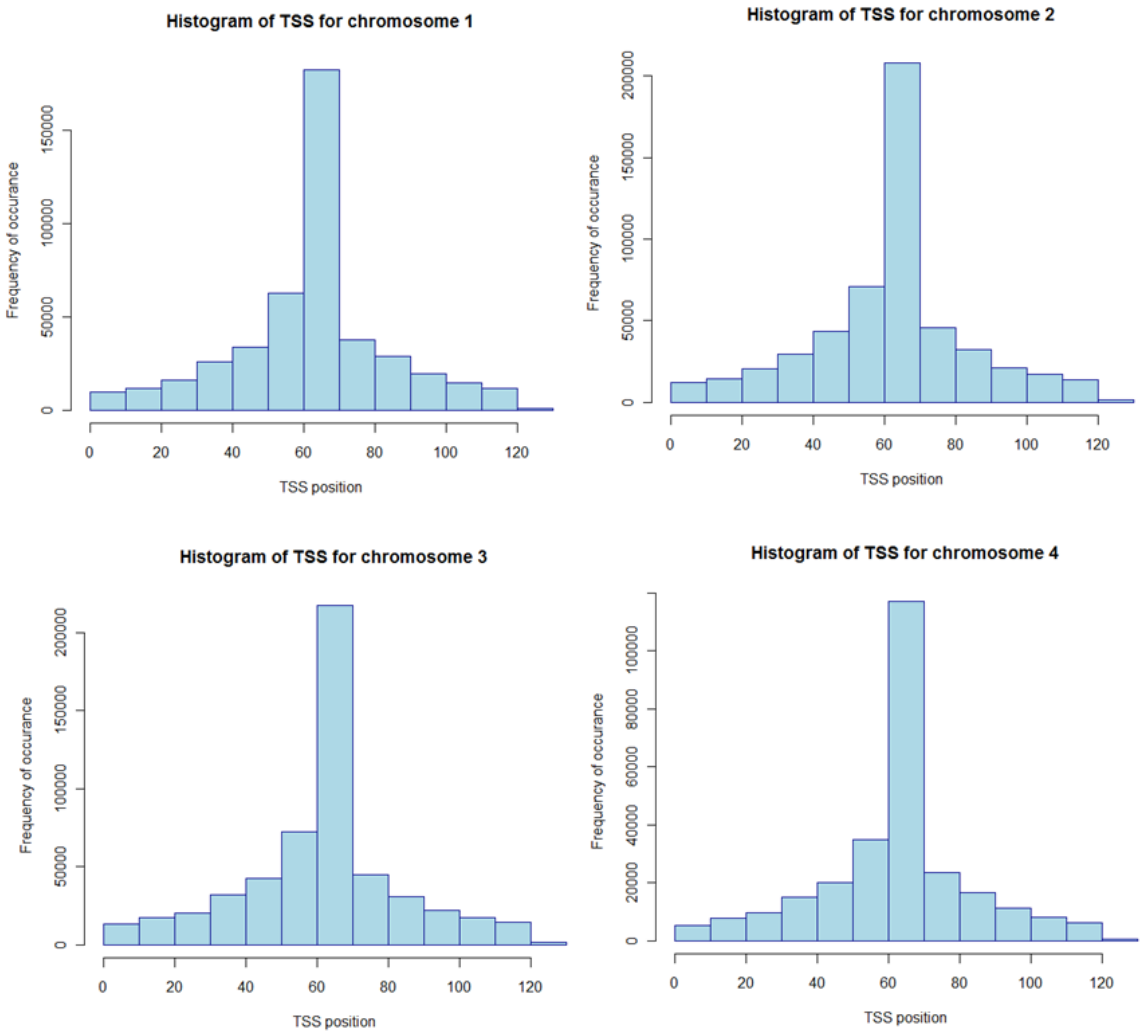


Figure 4.9. Readhead position distribution in TSS regions on chromosomes 1-4.

Readhead frequencies at each position of all identified TSS regions on each chromosome were combined. For ease of visualisation, positions were placed into 13 bins and their combined read frequencies plotted as histograms. The strong peak at the central position is an artefact caused by position 61 representing the strongest peak in each TSS segment. The presence of similar peaks at other positions would indicate RH enrichment due to alternative strong transcription start sites away from the primary TSS. These data indicate that the majority of TSS regions contain strong central start sites with high RH mapping. The level of RH mapping rapidly declined with increased distance from these central points, and this pattern was highly conserved between chromosomes.

4.10. Characterising TSS regions

While global analysis of RH distribution in TSS regions indicated a trend for normal distribution of transcriptional initiation around a primary start site, visual analysis of the data revealed that this was not always the case. While many regions presented as a near uniform RH distribution, others were observed with no single base possessing a significantly higher rate of RH mapping which would identify as a predominant start site, indicating that transcription started almost indiscriminately within these regions. Another observation was that many regions contained multiple bases with high RH mapping frequency, indicating the presence of multiple strong start sites within a single region (Fig. 4.10). This suggested variability in the specificity of each start site for a given nucleotide position. Based on this observation, a method was devised to statistically analyse the spread of RH mapping and produce a numerical representation of the variance in position of transcriptional initiation for each region.

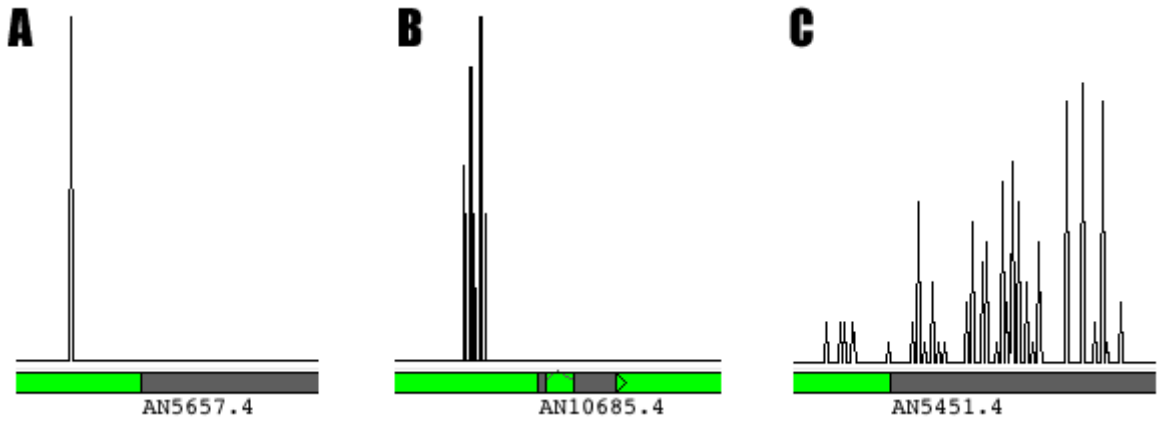


Figure 4.10. Examples of readhead mapping variance in transcription start sites. RH coverage data for three genes (black graphs) are displayed in Artemis against the *A. nidulans* genome annotation (green track), including genes present in each region (grey). Regions containing transcription start sites were shown to present with a number of different RH mapping profiles. These included a strong, single TSS location (A), multiple strong starts within a single TSS region (B), and widespread RH mapping indicating relatively indiscriminate transcription start sites (C).

Investigation into the variability of TSS specificity required the use of only well mapped regions. Regions of low expression may not present a true representation of the TSS and were therefore excluded. For this purpose, well mapped regions where the RH frequency of the highest single peak was ≥ 10 and the total RH frequency of the 121 bp region was ≥ 50 were identified for use in this analysis. Using these thresholds allowed the inclusion of both strong single peaks and regions of dispersed RH distribution, while aiming to further eliminate background noise and exclude regions which were poorly defined due to low mapping.

Using the above criteria, 15,475 of these 121 nt segments were identified as having significant RH mapping frequency. Each region was centred on the base with the highest RH mapping frequency, which therefore represented the preferred TSS within that region. By measuring the RH distribution around these primary TSS, it was possible to achieve a measure of variance in the location of transcriptional initiation in relation to that base position. This measure was presented in the form of a confidence interval. Regions with TSS found to be highly specific to a single base position were assigned a low confidence interval, while regions where transcription was shown to start across a wider range of positions were assigned lower confidence intervals.

Confidence values were assigned to each primary TSS based on the RH frequencies of surrounding bases. As previously discussed, RH frequencies in individual 121 bp regions rarely show normal distribution, more frequently presenting with a series of randomly dispersed and sized peaks. This prevented the use of a ready formula to calculate confidence intervals. Based on the normal assumption, positions with peak RH frequencies should represent the sample mean of each region. The data were therefore adapted by reversing the region around the peak position and adding the RH frequencies of the reversed region to those of the original to create a normal distribution while maintaining the variance of the unedited data. The relationship between variables could then be expressed by the equation

below, where z is the upper percentage point of the standard normal distribution and confidence interval length (CIL) is directly proportional to the standard deviation of RH distribution (σ) and inversely proportional to the sample size (n):

$$CIL = z\ 2\sigma / \sqrt{n}$$

This analysis was performed by Dr H. Wu, and resulted in two confidence values per TSS, representing the level of deviation in RH distribution upstream and downstream of the primary TSS. Combining these two values created a single statistic representing the frequency and distance of deviation in transcriptional start position from this primary TSS, and therefore a measure of heterogeneity for the position of transcriptional initiation. Lists of the primary (central) TSS location for each region and their associated CIL were produced. A subset of these data is shown in Table 4.1, full lists of TSS region positions and CI length scores can be found in Appendix 2.5.

Of the 15,475 regions analysed, 3,777 (24.4 %) were observed with CI lengths <2, indicating confidence values >95%. This represents very tight TSS clustering and a low level of positional deviation around a single, dominant start site. This indicates a high level of consistency in the position of transcriptional initiation within nearly a quarter of TSS regions in *A. nidulans*.

Table 4.1. TSS location confidence intervals and Length(bp) values. Data represents a subset of putative transcription start sites on the forward strand of chromosome I. Position is given as the central peak of each TSS region, representing the primary TSS. Confidence intervals represent the deviation possible in each direction from the position. These values are combined to produce the Length(bp) value, which provides a single figure as a statistical measure of TSS deviation within a given region.

Results for chromosome 1 of forward strand:

Position	Confidence Interval (0.95)	Length(bp)
41114	(-0.550 0.551)	1.101
175784	(-0.956 0.908)	1.864
228977	(-0.286 0.273)	0.559
229297	(-2.501 2.363)	4.864
229428	(-2.741 2.719)	5.460
229846	(-2.954 2.997)	5.950
245674	(-2.787 2.821)	5.608
247677	(-1.390 1.430)	2.820
286377	(-1.180 1.187)	2.367
322829	(-2.096 2.054)	4.150
322893	(-1.044 1.105)	2.149
324366	(-0.459 0.449)	0.908

4.11. Global analysis of TSS distribution

Global analysis of TSS distribution was performed by mapping the TSS regions across each chromosome. All TSS regions were mapped together, and separate tracks created to show start sites with low confidence intervals (< 2) which indicated strong positional specificity of transcriptional initiation, and the most highly expressed TSS determined by a total RH mapping frequency within the region of >1,000, indicating high levels of expression. No obvious clustering of transcription start sites and no indication of positional bias towards telomeric regions was observed, and statistical analysis by Dr Huihai Wu found no statistically significant distribution bias on any chromosome. An example of chromosome TSS distribution mapping can be seen in Fig.4.11, TSS maps for all chromosomes can be found in Appendix 2.6.

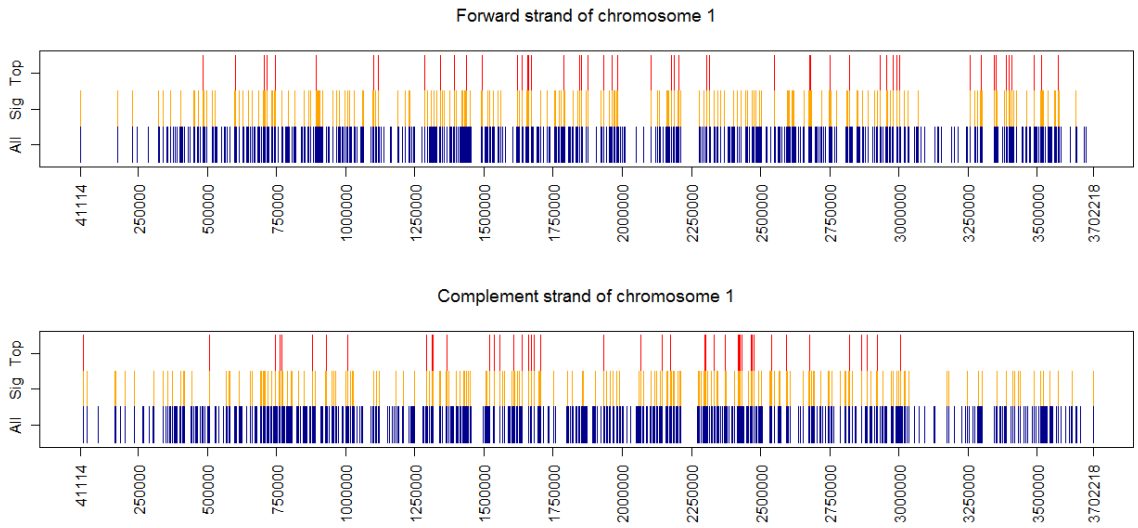


Figure 4.11. Chromosome map of identified TSS regions. Forward (Top) and the complementary (Bottom) strands of Chromosome 1 are shown with 3 tracks of data representing TSS positions. All identified TSS are indicated (blue), with additional tracks representing highly specific TSS with low confidence values (orange) and the most highly expressed TSS by total RH frequency (red). No statistically significant distribution bias was observed in any of the 8 chromosomes in the *A. nidulans* genome.

4.12. Motif finding and analysis

DNA sequence motifs are defined as short, recurring patterns in DNA presumed to have some biological function. Motifs located around the promoter are often present in sequence-specific binding sites for proteins such as transcription factors and therefore are often linked to gene regulation (D'haeseleer, 2006). Global identification of transcription start sites provided the opportunity to undertake promoter specific motif identification in *A. nidulans*.

To capture the promoter associated with each TSS, a 225 bp sequence ranging from 200 bp upstream to 24 bp downstream of the primary TSS position was extracted for each region. Based on previous classification of promoters in *A. nidulans* (Endo *et al*., 2008), this region was deemed to be of sufficient length to fully capture functional elements of promoter sequences. *De novo* motif finding was performed within these regions using the motif sequence finding softwares YMF (Sinha & Tompa, 2000) and MEME (Bailey & Elkan, 1994).

The effect of TSS distribution was investigated by creation of two TSS datasets for motif finding analysis. The first dataset (DS1) contained 3,777 regions with confidence values <2 representing very sharp TSS distribution as previously discussed. The second dataset (DS2) contained 3,959 regions with greater TSS distribution, identified by confidence values between 2 and 4.

The results of motif prediction by YMF analysis of DS1 were categorised into subsets with the software MATLIGN (Kankainen & Löytynoja, 2007), which converted the large list of motifs in the raw YMF output into small groups of similar, non-redundant motifs (Fig. 4.12).

Motifs found grouped by YMF z-scores (degree of over-representation) highest-lowest:

Group 1: CACGTG, CACGTGA, CACGTGAY
Group 2: AAAAWAA, AWAAWAA, ARARAAA, AAAAAAAA, ARARAAAA, AARARAAA
Group 3: TTTTTTW, TTWTTTW, TTTWTTW, TTTYTTY, TTTTYTTY, TTTYTTTY
Group 4: ACCRCC, CACCRCS, CWCCRCC, CACCRCC, ACCWCCA, CYCCGC
Group 5: GGCGGTS, GGYGGTS, RGGCGGTS, GYGGGGW, GCGGRGW
Group 6: GCCTSAGG, GCCTSAGGC
Group 7: CTCTCYY, CTCTTYT, ATAWATA, MATCAAC, MTCCAAC

Figure 4.12. Motifs identified from DS1. Motifs found by YMF were categorised into 7 groups and listed in order of descending over-representation.

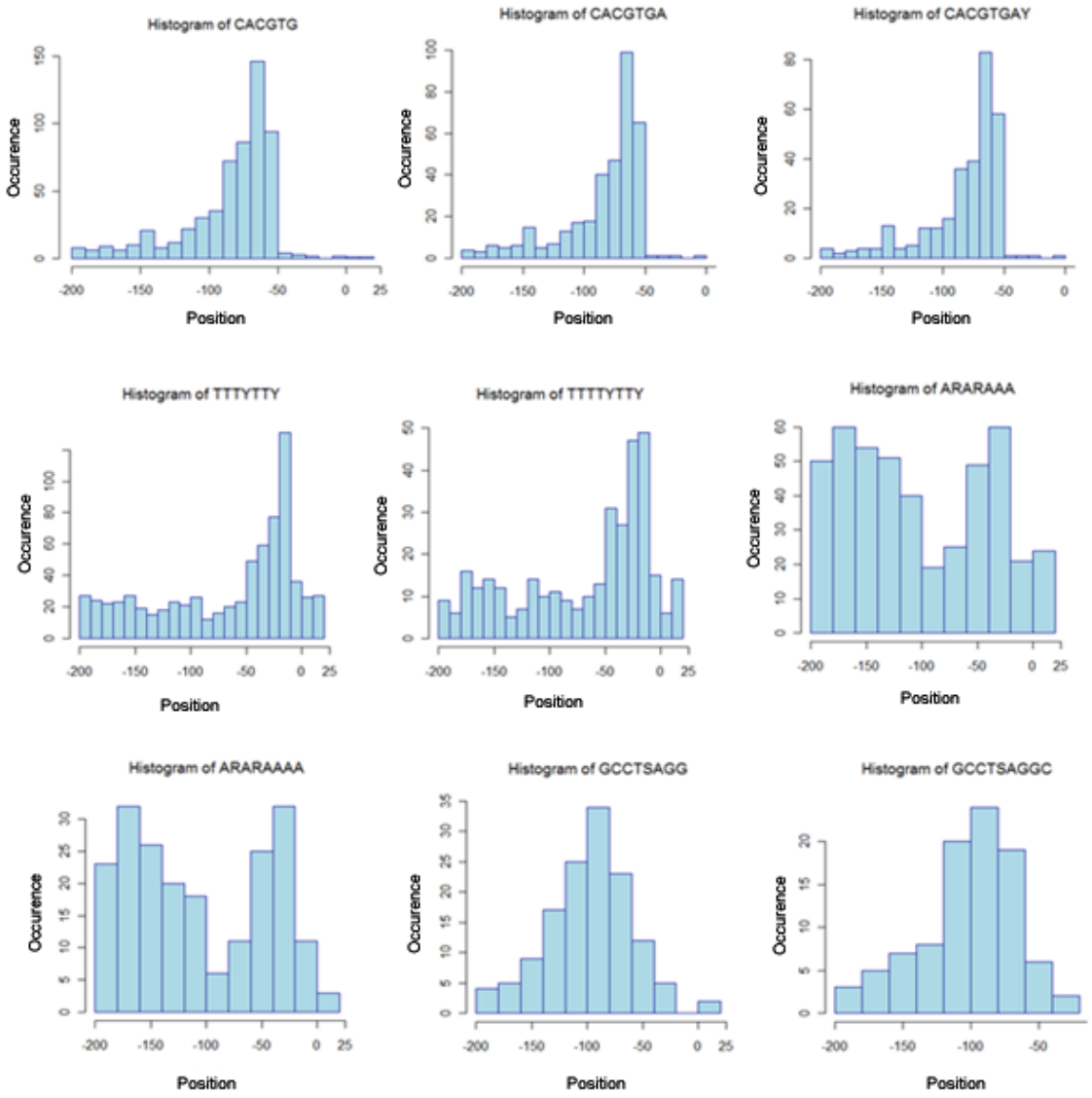
These motifs were largely consistent with those found by MEME, with all motifs identified belonging to the classes determined by MATLIGN and the strongest motif being the same in each platform (CACGTGA). Motifs predicted by both systems were grouped into 5 distinct subsets by Dr H. Wu. The most likely candidates for functional motifs were selected for further analysis and placed into classes based on the relative strength of the motifs. Motif strength was determined by a combination of Z-value (YMF), E-score (MEME), rate of occurrence and consistency of the motif based on any degenerate bases in the sequence (Fig.4.13).

The software used for motif discovery tests only for the degree of motif over-representation against a background sequence. From the list of motif groups identified (Fig. 4.13), it was clear that motifs in groups 2 (TTTYTTY, TTTTYTTY) and 3 (ARARAAA, ARARAAAA) consist of multiple base repeats, throwing doubt on their validity as real functional elements, despite their high level of enrichment against background sequences. While enrichment is a known characteristic of functional motifs (Frith *et al*., 2004), there is also evidence that strong, functional motifs are enriched within a narrow sequence region of the promoter (Vardhanabhuti *et al*., 2007). Motif position distribution was therefore analysed, firstly by producing histograms representing motif positions within promoters. This provided a visual representation of motif distribution in which all the motifs identified appear to have a positional bias within the promoter (Fig. 4.14), and these distribution varied dramatically between motifs. For example, motifs in class 1 (CACGTG, CACGTGA and CACGTGAY) showed very little distribution in the immediate vicinity of the TSS, but were strongly enriched around position -70. This very specific distance from the TSS supports the hypothesis that these motifs have a function within the promoter, requiring a specific distance from the TSS to perform their given role. As expected, Class 3 shows a much more evenly spread distribution pattern, and as such are less likely to possess a functional role. It is worth

noting however that class 2 motifs, while similar to class 3, express a high level of distributional enrichment.

Group 1: CACGTG, CACGTGA, CACGTGAY
Group 2: TTTYTTY, TTTTYTTY
Group 3: ARARAAA, ARARAAAA
Group 4: GCCTSAGG, GCCTSAGGC
Group 5: GGCGGTS, RGGCGGTS
Group 6: ACCRCC, CACCRCC

Figure 4.13. Classes of motif arranged from strongest (1) to weakest (6). These represent the most likely functional motifs from the combined outputs of MEME and YMF.



Continued overleaf.

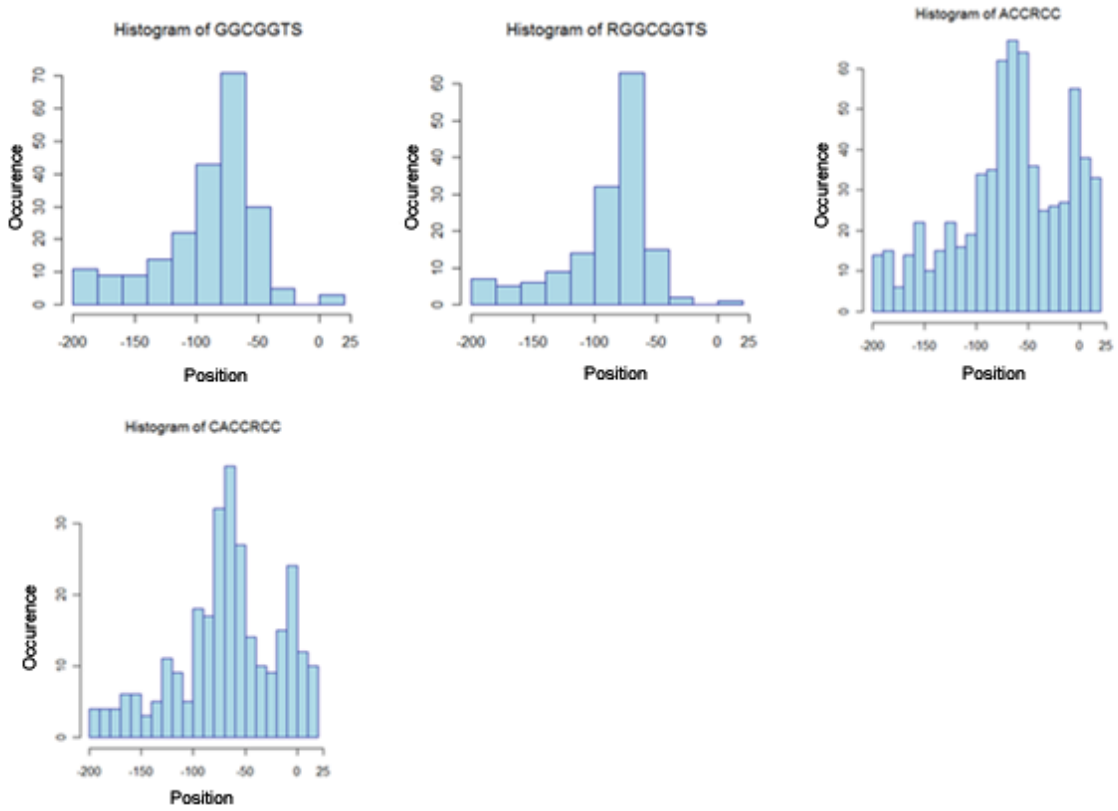


Figure 4.14. Histograms of enrichment relative to TSS location for 13 identified motifs. The occurence of motifs starting at each base position relative to the TSS (position 0) was plotted. Base posisions for motifs with higher occurence were divided into 22 bins. This was impractical for motifs with lower occurence as the bins were too small to small to easily interpret the resultant graphs. To facilitate ease of interpretation, the number of bins was halved to 11 for these motifs. A high level of enrichment around certain positions are indicated for a number of motifs, while others show more even distributions and are therefore less likely to perform functional roles.

To provide a statistical measure of distributional bias, Chi-square GOF uniformity tests were performed by Dr H.Wu, to determine goodness of fit for observed distribution against uniform distribution within the defined promoter region. Chi-square GOF uniformity tests determine the goodness of fit between observed and normal distributions. The smaller the p-value assigned, the more statistically significant the deviation from standard distribution. P-values lower than 0.05 are generally regarded as statistically significant. All motif positional distributions were found to have extremely low p-values of 0.0005, indicating significant variance from normal distribution, but providing no means of differentiating distribution between motifs. Dr Wu therefore coupled these data with kernal density estimation (Fig. 4.15) to re-order motif classes based on the level of distributional enrichment around the TSS (Fig. 4.16).

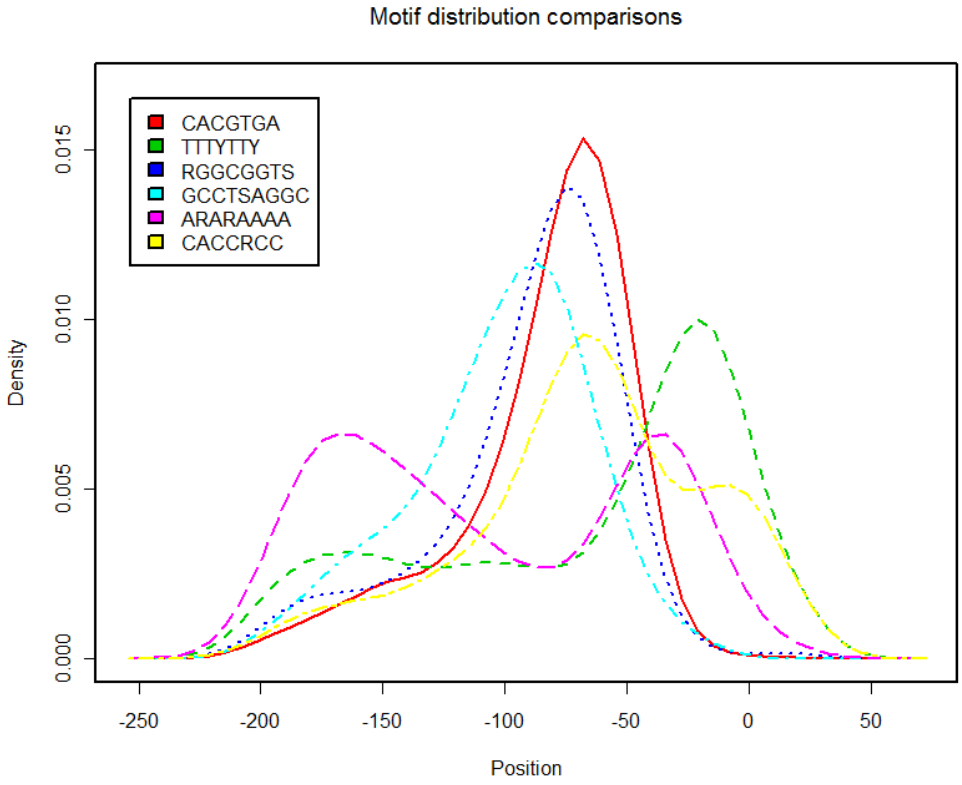


Figure 4.15. Distribution of motif classes by kernel density estimates. The relative density of a range of motifs at each nucleotide position around the TSS shows a clear and statistically significant bias, consistent with them having functional roles within the promoter.

Class 1: CACGTG, CACGTGA, CACGTGAY

Class 5: GGCGGTS, RGGCGGTS

Class 4: GCCTSAGG, GCCTSAGGC

Class 2: TTTYTTY, TTTTYTTY

Class 3: ARARAAA, ARARAAAA

Figure 4.16. Motif classes ordered by level of distributional enrichment. Combining the level of distributional enrichment with the scores from YMF and MEME, motif classes were reordered from strongest (top) to weakest (bottom).

Motif finding analysis was then repeated for the promoter regions of DS2. YMF produced a similar output. However, the motifs found were significantly weaker than those in DS1, with occurrence and Z-scores demonstrating a ~50% decrease in motif over-representation. Due to the weak motifs in this dataset, the motif finder MEME did not work well, finding very few significant motifs. The top 7 bp motif found in each dataset produced a perfect match (CACGTGA), and a number of other patterns emerged between the two datasets. Filtering motifs based on software scores and frequency of occurrence gave strong indication of functional distribution in additional motif groups: CTCTCYY, CTCTTYT, ATAWATA, MATCAAC and MTCCAAC.

Analysis of motif strength was again performed by Dr H. Wu. Combining the results from both DS1 and DS2 gives a total of 38 strong motifs in 7 groups (Fig. 4.17). Distribution analysis within promoters revealed 16 motifs most likely to possess a functional role (Fig. 4.18).

Group 1: CACGTG, CACGTGA, CACGTGAY

Group 2: AAAAWAA, AWAAWAA, ARARAAA, AAAAAAAA, ARARAAAA, AARARAAA

Group 3: TTTTTTW, TTWTTTW, TTTWTTW, TTTYTTY, TTTTYTTY, TTTYTTTY

Group 4: ACCRCC, CACCRCS, CWCCRCC, CACCRCC, ACCWCCA, CYCCGC

Group 5: GGCGGTS, GGYGGTS, RGGCGGTS, GYGGGGW, GCGGRGW

Group 6: GCCTSAGG, GCCTSAGGC

Group 7: CTCTCYY, CTCTTYT, ATAWATA, MATCAAC, MTCCAAC

Figure 4.17. Combined strong motifs from both datasets. 38 strong motifs were found in 7 distinct groups, listed in order of YMF Z-scores for degree of over-representation from strongest to weakest.

Group 1: CACGTG, CACGTGA, CACGTGAY

Group 2: TTTYTTY, TTTTYTTY, TTTYTTTY

Group 3: ACCRCC, CWCCRCC, CACCRCS, CACCRCC

Group 4: CTCTCYY, CTCTTYT, ATAWATA, MTCCAAC

Group 5: GGCGGTS, RGGCGGTS

Figure 4.18. 16 strongest candidates for functional motifs. Motifs are ordered based a combination of their scores from previous analysis combined with distribution analysis within the promoter.

Comparing motif finding results between DS1 and DS2 demonstrated a significant drop in motif over-representation as the specificity of the primary start site decreased. To further investigate this relationship, a new dataset was created (DStop) containing the top 10% of promoters in terms of RH frequency at the primary TSS. Using cumulative hypergeometric distribution, Dr Wu was able to calculate the probability of motif occurrences compared to the observed number, expressed as p-value of occurrence rate, with lower p-values indicating greater statistical significance. 34 motifs were tested against the 3 datasets (Table 4.2). From the average p-values, it was clear that the statistical significance of motif occurrences in DS1 was far greater than that in DS2, indicating a strong correlation between motif occurrence and specificity of the TSS. However, the average p-value of DStop was not significantly different to and in fact slightly higher than that of DS1. Therefore, as all promoters in DStop are also found in DS1, there appeared to be no correlation between motif occurrence and expression level.

Table 4.2. Significance of motif occurrences in different datasets. From the average p-values it is clear that the significance of motif occurrences in DS1 is far greater than that in DS2, indicating a strong correlation between motif occurrence and specificity of the TSS.

Datasets:	DS1	DS2	DStop
Motifs	P-value	P-value	P-value
CACGTG	1.7179E-18	1.0000E+00	2.1974E-04
CACGTGA	4.2916E-19	1.0000E+00	1.2218E-04
CACGTGAY	7.2512E-18	1.0000E+00	1.0279E-04
GCCTSAGG	2.2770E-03	9.9850E-01	1.2744E-02
GCCTSAGGC	7.8066E-04	9.9956E-01	2.8984E-03
GGCGGTS	4.4354E-03	9.9672E-01	3.4696E-01
GGYGGTS	3.3443E-01	6.9358E-01	5.6422E-01
RGGCGGTS	1.0953E-04	9.9993E-01	2.8046E-01
GYGGGGW	8.0896E-01	2.1834E-01	7.0614E-01
GCGGRGW	1.7388E-01	8.4715E-01	6.5796E-01
TTTTTTW	1.8928E-03	9.9861E-01	6.3331E-02
TTWTTTW	8.3815E-02	9.2690E-01	1.4004E-01
TTTWTTW	1.0008E-02	9.9186E-01	3.4062E-02
TTTYTTY	1.4002E-04	9.9989E-01	7.1137E-10
TTTTYTTY	2.0470E-05	9.9999E-01	5.9866E-08
TTTYTTTY	9.3205E-05	9.9993E-01	2.7824E-07
AAAAWAA	1.5936E-06	1.0000E+00	1.3973E-02
AWAAWAA	5.1683E-08	1.0000E+00	3.6517E-02
ARARAAA	6.5249E-08	1.0000E+00	1.2422E-01
AAAAAAAA	5.1909E-06	1.0000E+00	2.4465E-02
ARARAAAA	1.8580E-06	1.0000E+00	6.5643E-02
AARARAAA	7.3579E-06	1.0000E+00	2.0463E-01
ACCRCC	6.2715E-01	3.9540E-01	2.0602E-01
CACCRCS	3.4131E-01	6.8555E-01	6.4725E-01
CWCCRCC	4.3185E-01	5.9342E-01	8.8766E-01
CACCRCC	1.8547E-01	8.3655E-01	7.3450E-01
ACCWCCA	9.4434E-01	6.7786E-02	1.2929E-01
CYCCGC	2.3491E-01	7.8176E-01	9.1166E-01
CTCTCYY	6.8881E-01	3.3370E-01	2.7989E-08
CTCTTYT	6.1285E-01	4.1695E-01	1.1063E-06
ATAWATA	4.4312E-08	1.0000E+00	3.8243E-01
MATCAAC	3.3485E-01	6.9931E-01	8.2087E-01
MTCCAAC	7.4321E-01	2.8912E-01	7.5746E-01
Ave p-value	0.1989577	0.8112268	0.2653287

4.13. Association of TSS regions and motifs with annotated gene models

Regions of transcriptional initiation were associated with genes in the current *A. nidulans* annotation. Associations were based on either a direct overlap of a segment with an existing gene model, or where a segment fell within 500 bp upstream of a CDS. Distance from TSS to CDS has been studied *in Arabidopsis thaliana* and identified as 60-120 bp in that system (Yamamoto *et al*., 2011). Allowing matches between TSS and CDS up to a distance of 500 bp was therefore a fairly lenient threshold which was more than sufficient to ensure the identification of the maximum possible number of valid matches in A. *nidulans*.

Novel software (Chapter 2.5.6) was created to associate the 15,475 identified TSS regions with current gene annotations based on these criteria and perform motif finding within promoter regions surrounding these TSS. Each putative TSS was associated with a previously annotated gene by virtue of its position either within or under 500 bp upstream of the gene model.

The software then extracted promoter sequences surrounding gene-matched TSS regions and performed sequential alignment of identified motifs against these promoters (Fig. 4.19). The promoter region size of 225 bp (200 bp upstream and 24 bp downstream) was consistent with previous motif finding analysis. The 16 previously identified motifs were then aligned to these sequences, producing a list of genes associated with each motif. The number of gene matches for each motif can be seen in Table 4.3. Full lists of TSS:Gene associations can be found in Appendix 2.7, and lists of gene matches for each motif with the promoter region in FASTA format can be found in Appendix 2.8.

A second piece of software was created to count the number of TSS which had been associated with existing gene models, and determine the number of unique genes with at least one putative TSS (Appendix 2.1.5). 13,035 (84.2 %) of TSS were associated a total of 6,014

known genes, with 2,325 of these genes linked to multiple distinct TSS regions. This software also produced a list of gene IDs with multiple TSS associations which can be found in Appendix 2.7.1.

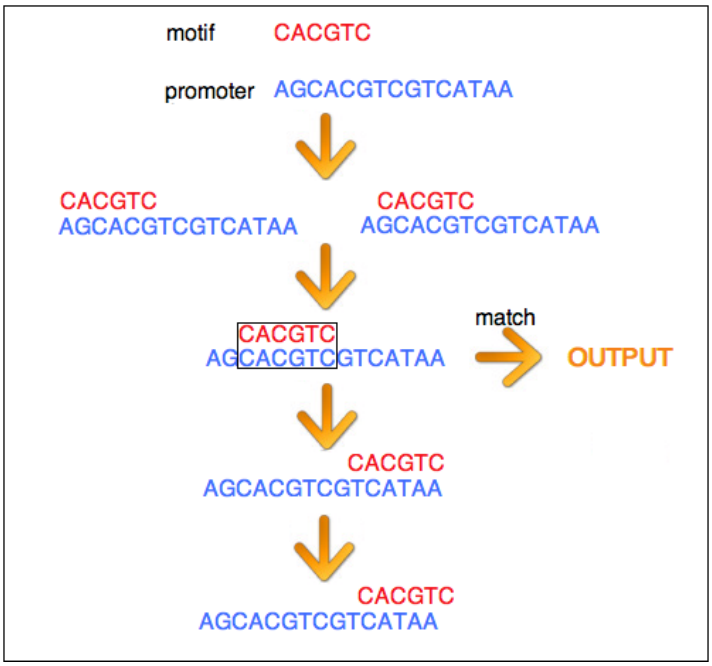


Figure 4.19. Sequential alignment of motifs to promoter sequences. Motifs were first aligned at the 5' end of the promoter, then moved sequentially towards the 5' end at single nucleotide intervals. This process continued until the motifs could no longer fully align. Upon successful alignment the gene associated with the current promoter sequence was recorded in an output file.

Table 4.3. Motif:gene matches. The table shows the number of genes with promoter regions aligned to each of the 16 strong motifs identified by motif finding analysis. The percentage of hits for each chromosome is displayed as well as the total number of genes associated with each motif.

Motif	CACGTG	CACGTGA	CACGTGAY	TTTYTTY	TTTTYTTY	TTTYTTTY	ACCRCC	CWCCRCC
Chr 1	76 (14.81%)	46 (15.03%)	40 (15.44%)	144 (15.14%)	87 (18.13%)	88 (18.64%)	88 (14.33%)	71 (14.26%)
Chr 2	69 (13.45%)	48 (15.69%)	40 (15.44%)	133 (13.99%)	61 (12.71%)	63 (13.35%)	105 (17.10%)	86 (17.27%)
Chr 3	68 (13.26%)	32 (10.46%)	27 (10.42%)	166 (17.46%)	90 (18.75%)	87 (18.43%)	83 (13.52%)	70 (14.06%)
Chr 4	37 (7.21%)	22 (7.19%)	18 (6.95%)	57 (5.99%)	26 (5.42%)	22 (4.66%)	41 (6.68%)	37 (7.43%)
Chr 5	58 (11.31%)	35 (11.44%)	31 (11.97%)	57 (5.99%)	26 (5.42%)	24 (5.08%)	52 (8.47%)	38 (7.63%)
Chr 6	60 (11.70%)	33 (10.78%)	27 (10.42%)	91 (9.57%)	41 (8.54%)	36 (7.63%)	66 (10.75%)	58 (11.65%)
Chr 7	65 (12.67%)	38 (12.42%)	33 (12.74%)	122 (12.83%)	57 (11.88%)	59 (12.50%)	67 (10.91%)	58 (11.65%)
Chr 8	80 (15.59%)	52 (16.99%)	43 (16.60%)	181 (19.03%)	92 (19.17%)	93 (19.70%)	112 (18.24%)	80 (16.06%)
Total	513	306	259	951	480	472	614	498

Motif	CACCRCS	CACCRCC	CTCTCYY	CTCTTYT	ATAWATA	MTCCAAC	GGCGGTS	RGGCGGTS
Chr 1	48 (13.26%)	40 (15.09%)	73 (12.33%)	28 (8.86%)	29 (11.24%)	15 (8.11%)	33 (16.10%)	23 (15.54%)
Chr 2	60 (16.57%)	47 (17.74%)	91 (15.37%)	56 (17.72%)	43 (16.67%)	33 (17.84%)	22 (10.73%)	14 (9.46%)
Chr 3	47 (12.98%)	38 (14.34%)	66 (11.15%)	40 (12.66%)	29 (11.24%)	25 (13.51%)	39 (19.02%)	32 (21.62%)
Chr 4	25 (6.91%)	17 (6.42%)	53 (8.95%)	25 (7.91%)	16 (6.20%)	12 (6.49%)	14 (6.83%)	12 (8.11%)
Chr 5	28 (7.73%)	16 (6.04%)	53 (8.95%)	31 (9.81%)	31 (12.02%)	15 (8.11%)	8 (3.90%)	5 (3.38%)
Chr 6	44 (12.15%)	34 (12.83%)	49 (8.28%)	27 (8.54%)	25 (9.69%)	22 (11.89%)	24 (11.71%)	14 (9.46%)
Chr 7	40 (11.05%)	27 (10.19%)	91 (15.37%)	52 (16.46%)	49 (18.99%)	24 (12.97%)	29 (14.15%)	20 (13.51%)
Chr 8	70 (19.34%)	46 (17.36%)	116 (19.59%)	57 (18.04%)	36 (13.95%)	39 (21.08%)	36 (17.56%)	28 (18.92%)
Total	362	265	592	316	258	185	205	148

4.14. Consensus sequences around gene associated start sites

5′ heterogeneity has been shown to occur in eukaryotic organisms (Gowda *et al*., 2006). Later studies noted that this phenomenon also occurs in fungi (Gowda *et al*., 2007; Spanu & Doyle, 2009) and is most probably due to inaccurate transcription by RNA polymerase in the first stages of the process. TSS regions with particularly strong single peaks and tight distributions are therefore likely to have an additional feature within the promoter which guides the machinery of transcription to initiate at a highly specific locus.

To investigate this hypothesis, novel software was developed to interrogate a 21 nucleotide sequence surrounding putative TSS locations, representing 10 bp in each direction upstream

and downstream of the central peak. This software took a list of TSS positions and aligned each TSS to the *A. nidulans* reference in FASTA format. The 21 nt sequence surrounding each TSS was obtained from the reference, and the nucleic acid at each nt position was recorded using a set of 4 arrays (representing A, C, G and T). These arrays were 21 elements in length, with each element representing a single nt position, and used to maintain counts of the nucleic acids found at each position (Chapter 2.5.7).

TSS with strong positional specificity were more likely to be surrounded by a consensus sequence to guide the machinery of transcription to this location. Therefore, only tightly distributed TSS with CIL < 2 were used. The resultant counts of nucleic acid occurrence at each position were used in conjunction with WebLogo 3.0 (Crooks *et al*., 2004) to generate a graphical representation of patterns within a multiple sequence alignment, also known as a sequence logo (Schneider & Stephens, 1990). Logos consists of stacks of letters, one stack for each nucleotide position in the sequence, with the overall height of the symbols within the stack reflecting the relative frequency of occurrence at that position (Crooks, *et al.* 2004) Fig. 4.20.

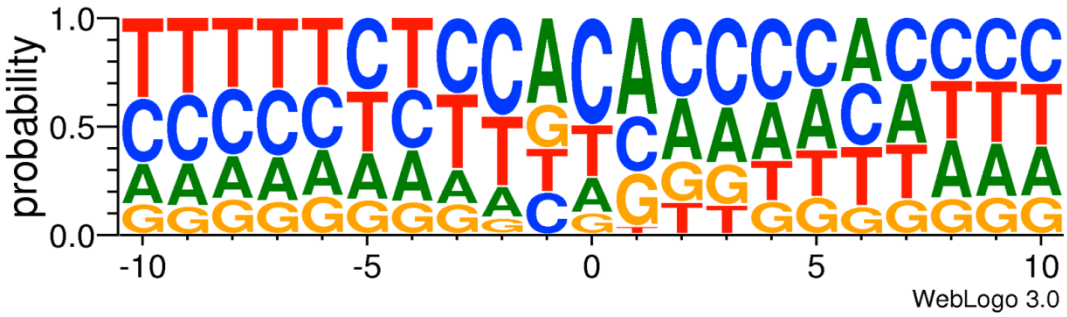


Figure 4.20. Sequence logo for the region surrounding transcription start sites in *A. nidulans*. The data used to generate this sequence logo consists of the -10 to +10 nucleotide regions around each TSS (shown here at 0) while the relative height of the letters at each position represents the relative probability of the respective base appearing at that location.

From the sequence logo produced, there did not appear to be any particular motif which could cause transcription to initiate at a given locus. These TSS must therefore rely on an upstream element of the promoter guiding the machinery of transcription to produce their tight distributions.

4.15. Further development of 5′ specific RNA-seq protocols

4.15.1. Investigation of mRNA decay

During visual analysis of 5′ specific RNA-seq readhead data, a significant number of readheads were observed downstream of some transcription start site regions. These generally presented as a series of peaks with RH frequencies which trailed off with increasing distance from the TSS. The most likely cause of this phenomenon was the presence of mRNAs with truncated 5′ ends in the original sample. Natural degradation of mRNA 5′ ends in *A. nidulans* cells occurs primarily by 5′ to 3′ exonucleolytic decay, and therefore produces 5′ ends compatible with SOLiD adaptor ligation. Decapping and subsequent 5′ degradation is triggered by shortening of the mRNA 3′ poly(A) tail to around 15 residues (Morozov *et al.*, 2010), meaning that mRNAs with degraded 5′ ends often still possess poly(A) tails (Morozov *et al.*, 2010; Morozov & Caddick, 2012), and can therefore be enriched by poly(A) selection using oligo(dT). Natively degraded 5′ ends are therefore likely to be represented in the final SOLiD library.

Large amounts of what appear to be mRNAs with naturally decapped and decayed 5' ends had not been anticipated in the design of a 5'-specific sequencing protocol. The data produced as a result of this could potentially be extremely interesting for the analysis of mRNA degradation in living cells. However, the inclusion of these degradation products was not

ideal for the identification of true transcription start sites. The extra readhead peaks could potentially interfere with the statistical analysis of the TSS location and can cause uncertainty as to the exact location of the start site. The protocol for 5'-specific RNA-seq was therefore modified to make these degradation products ineligible for adaptor ligation, by removal of the 5′ phosphate using alkaline phosphatase (ALP). Performing this step prior to decapping meant that the 5′ ends of full length transcripts were protected from the action of the ALP, and therefore only these 5′ ends would be represented in the final library.

Alkaline phosphatases are enzymes present in many organisms ranging from *E. coli* to humans (McComb *et al*. 1979). The three-dimensional crystallographic structure of *E. coli* alkaline phosphatase has been determined and the reaction mechanism for this group of enzymes has been inferred (Coleman, 1992). With few exceptions alkaline phosphatases are homodimeric, nonspecific phosphomonoesterases, with each catalytic site comprising of three metal ions (2 ZN^{2+} and one Mg^{2+}). These three metal ions are required for enzymatic activity (Millan, 2006) and the conservation of these ions as well as the residues which comprise the catalytic zinc binding site and the substrate binding residues, suggests conservation of the reaction mechanism throughout evolution (Kim & Lipscomb, 1990). There are significant structural differences between different alkaline phosphatases, with several loop regions being inserted and/or deleted between *E. coli* and mammalian enzymes. This huge variability in structure is believed to cause the huge differences in catalytic activity known to be a feature of alkaline phosphatases, with mammalian enzymes displaying kcat values 10-100 orders of magnitude above that of those found in *E. coli* (Murphy *et al.,* 1995).

rAPid Alkaline Phosphatase from Roche is supplied as a recombinant enzyme isolated from bovine intestine and expressed in the yeast *P. pastoris*. This enzyme catalyzes the dephosphorylation of 5' phosphates from DNA and RNA, as well as nucleotides and proteins.

It was chosen as an alternative to shrimp or calf alkaline phosphatases as it is irreversibly inactivated by heat treatment for 2 minutes +75°C and is active in restriction enzyme buffers, allowing restriction enzyme digestion, dephosphorylation, enzyme inactivation, ligation, and 5'-end labelling to be performed without purification steps (Roche). This makes it ideal for use in this 5' sequencing protocol as it can be used to remove the the 5'-phosphoryl termini required by ligases, preventing ligation of the uncapped 5' ends to the SOLiD primers, while requiring fewer purification steps which could lead to sample loss or contamination.

Early attempts at using this enzyme and the standard dephosphorylation procedure suggested by the manufacturer resulted in complete degradation of the RNA samples being treated, despite claims that the product is nuclease-free. We found that this could be prevented by the addition of an RNase inhibitor and that an incubation of 30 minutes at 37°C using the manufacturer's dephosphorylation procedure was sufficient to completely dephosphorylate the uncapped RNA present in the sample.

To test the effectiveness of this ALP treatment, a new 5'-specific RNA-seq library was created (+TAP+ALP). The library preparation protocol was as previously described, but RNA was treated with rAPid Alkaline Phosphatase prior to decapping. This transcription start sites identified by this new library should therefore match start sites identified by libraries prepared with the original protocol, but with vastly reduced contamination due to degradation products.

As a control, a library was prepared from RNA which had been neither treated with ALP or decapped with TAP (-TAP-ALP), meaning that only natively decapped mRNAs should be sequenced. 5' ends identified by this library would not necessarily represent the location of transcriptional initiation, instead representing the products of decapping and 5' degradation. These RH locations should therefore still match to locations identified in the original 5'-

specific RNA-seq library, but not to those identified by the library treated with ALP. If this method proved successful in isolating natively decapped transcripts, libraries created in this manner could be used for future investigation of mRNA decay in Aspergilli and other organisms.

Finally, a new library was prepared using the original 5'-specific RNA-seq protocol (+TAP-ALP), with no ALP treatment prior to decapping. This enabled all three libraries to be prepared using the same total RNA as starting material, and to be sequenced together on a single SOLiD version 3 Plus slide, making them directly comparable.

Alignment of SOLiD reads from these three new libraries to the *A. nidulans* genome was performed using the ABI Bioscope software suite as previously described. As an additional quality control step, only reads where the entire 5' end of the read could be aligned to the *A. nidulans* genome were converted to readheads. Novel software (Chapter 2.5.8) was developed to identify reads where the 5' ends had been truncated by Bioscope to produce a valid alignment. These were removed from the dataset as truncation of the 5' end meant they would not necessarily identify the true 5' ends of transcripts. To reduced background mapping and false-positive TSS identification, this software also removed reads which had been successfully mapped to multiple loci.

Aligned read data were converted to readheads using a different method to that described in Chapter 4.6. Instead of using pileup data, novel software was developed to generate pileups of RH mapping directly from the mapped read data (Chapter 2.5.9). This method was more robust and efficient than previous RH conversion software, and allowed RH pileups to be generated directly from full length and uniquely mapping read data.

Additional software was developed to align two sets of readhead data and report the number of matching RH positions (Chapter 2.5.10). To further reduce background noise and false positive transcription start site identification, positions with RH mapping frequencies <3 were removed from each dataset prior to alignment. This ensured that only positions with significant mapping and therefore representing real transcript 5′ ends would be compared.

As predicted, transcription start sites identified by the alkaline phosphatase treated +TAP+ALP library showed excellent correlation with those identified by the library prepared using the original non-ALP treated protocol (+TAP-ALP). 86% of RH positions identified by +TAP+ALP aligned to those identified by +TAP-ALP, however this represented only 15.3% of RH positions in the latter dataset. Additionally, only 6.7% of +TAP+ALP positions could be aligned to those identified by the –TAP-ALP library, showing near complete elimination of natively decapped and degraded mRNAs from the ALP treated library. These statistics confirmed the presence of a large number of transcripts with partial 5′ degradation in libraries prepared without the use of ALP, and demonstrated ALP treatment to be extremely effective at removing these transcripts from the finished library while maintaining the integrity of TSS identification (Fig. 4.21). ALP treatment therefore represents a vast improvement over previous protocols for the robust analysis of full length mRNA 5′ ends and therefore the true positions of transcriptional initiation.

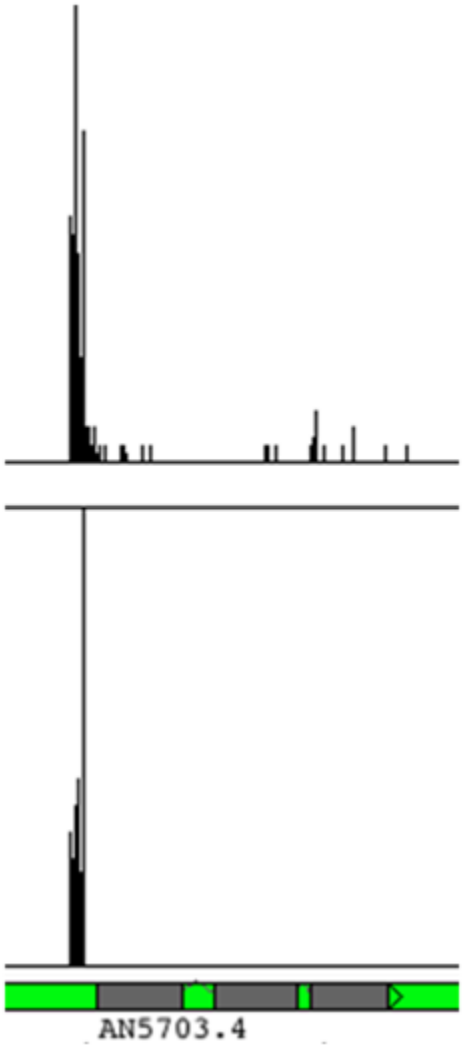


Figure 4.21. Comparison of 5′ specific RNA-seq readhead data from libraries produced with and without Alkaline phosphatase (ALP) treatment. RH coverage data is shown (black graphs) against the *A. nidulans* genome annotation (green track). Genes are shown on the annotation track in grey. The non-treated library (top graph) includes a trail of readheads immediately downstream of the TSS, and a number of other positions of readhead mapping further downstream. The exclusion of these readhead positions in the library treated with ALP (bottom graph) indicates that they represent the products of 5′ degradation, and ALP treatment is shown to produce a far more accurate representation of TSS positions.

To explore the validity of –TAP-ALP treated libraries for the investigation of 5′ degradation, -TAP-ALP RH positions were aligned to +TAP-ALP. A high level of consistency had been shown in the investigation of capped, full-length mRNAs, however only 31.4% of –TAP-ALP RH positions could be aligned to +TAP-ALP, showing relatively poor consistency in the identification of degradation products.

As degradation products were expected to represent a relatively low proportion of total-RNA, the filtering of RH positions with mapping frequency <3 was removed and alignment of –TAP-ALP to +TAP-ALP repeated for the unfiltered readhead data. An almost ten-fold increase in the number of alignments was observed, however the total number of RH positions had increased by a factor of 19, resulting in a lower percentage match. This was due to the high frequency of low level and random background mapping, and highlighted the importance of filtering out positions with low mapping frequency for accurate TSS identification. The data were therefore extremely difficult to interpret, as less than 1/3 of positions were identified in both datasets.

4.15.2. Increased library quality using a shorter Random priming P2 oligonucleotide

Although successful 5′ specific RNA-seq libraries had been created and sequenced on the SOLiD platform, a number of issues arose during library creation and data analysis. Despite using the most robust method of PCR cleanup, the library was found to contain what appeared to be several artefacts producing large peaks at sizes 100, 150 and 253 bp. The high concentration of these artefacts in relation to the rest of the sample meant that they could potentially interfere with the efficiency of the emulsion PCR. While this did not appear to

reduce the quality of the library to below that of whole transcriptome RNA-seq libraries, it may still have impacted on efficiency and required further investigation.

It was interesting to note that the sizes of the artefacts observed were all approximate multiples of 50. This was a similar size to the Random Priming P2 (barcoded) oligonucleotide primer used during first strand synthesis, which may therefore be responsible for the artefacts in the library. To investigate this issue, a shorter version of the Random Priming P2 oligonucleotide was developed which would be easier to remove by purification and size selection procedures. This shorter "Random priming P2 (non-barcoded)" oligonucleotide also presented the opportunity to more easily multiplex 5′ specific RNA-seq libraries. By designing the primer to include only the 3′ internal adaptor sequence in addition to the random priming degenerate sequence (Fig. 4.22), barcodes could be added during amplification using the standard SOLiD barcoding kit reagents and protocols.

Libraries produced using this shorter oligo did not contain any of the previously observed artefacts when checked for quality on an Agilent 2100 bioanalyser (Fig. 4.23). A small, unexpected peak was observed at 112 bp, which could be easily removed by a secondary round of size selection. However, this was deemed unnecessary as the concentration of the artefact was so low that it would have no noticeable effect on the efficiency of emulsion PCR. This shorter P2 oligo therefore represented a marked increase in library quality compared to the longer, pre-barcoded P2, as well as providing multiplexing compatibility for future 5′ specific RNA-seq libraries.

Random priming P2 (barcoded):

CTGCCCCGGGTTCCTCATTCTCTAAGCCCCTGCTGTACGGCCAAGGCGNNNNNC

Random priming P2 (non-barcoded):

CTGCTGTACGGCCAAGGCGNNNNNC

Figure 4.22. Comparison of Random priming P2 oligonucleotides. The two primers are aligned and presented in 5'-3' orientation. The shorter version of the random priming P2 contains only the internal adaptor before the SOLiD barcode (blue) and the degenerate sequence (red). The internal adaptor provides a binding site compatible with any of the barcoded primers in the PCR primer mix from ABI, allowing any barcode to be added to a sample once the cDNA has been generated.

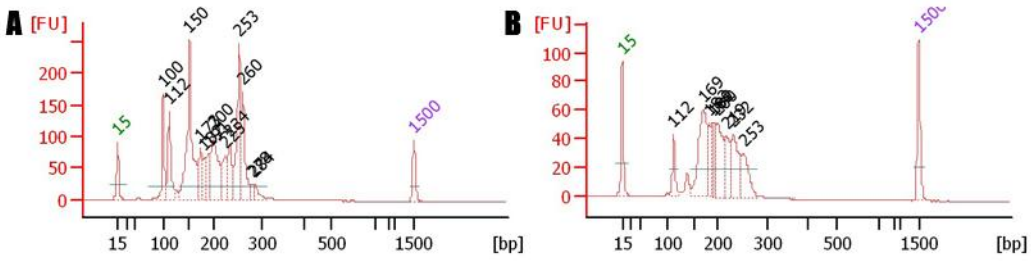


Figure 4.23. Agilent bioanalyser plots for samples prepared with the barcoded and non barcoded Random Priming P2 oligonucleotide. Library preparation with a barcoded Random Priming P2 oligonucleotide (A) resulted in unwanted peaks at 100, 150 and 252 bp, indicating contamination of the SOLiD library by PCR artefacts. The shorter Random Priming P2 oligonucleotide (B) produced higher quality libraries in which the sizes of these peaks were greatly reduced. Peaks at 15 (green) and 1500 (purple) are standard size markers.

To fully assay the improvement in library quality as a result of this new, shorter P2, the creation of a new SOLiD library was required, which could be compared to previous libraries created with the longer P2 oligo. Comparisons between protocols were performed using the newer libraries rather than the first 5'-specific RNA-seq data, as newer libraries were sequenced on the updated SOLiD version 3 Plus system, where as the previous 5' RNA-seq library had been sequenced on the older SOLiD version 2. Comparing libraries sequenced on the same version of the SOLiD system therefore provided a more robust assay of library quality.

A biological repeat of the previously discussed TAP+ALP library was created using the new, shorter P2 oligo. To ensure the two libraries were comparable, the same TAP and ALP treated RNA sample was used in the creation of this new library, and the only change in the protocol was using the shorter P2 to prime first strand synthesis. The library was multiplexed and sequenced on the same proportion of a SOLiD version 3 Plus slide as previous libraries, again to ensure comparable data. 4,053,500 reads were mapped to the genome with Bioscope, of which 3,143,927 (77.6 %) mapped to unique positions. Therefore, the new library demonstrated a marked increase in quality over the library prepared with the longer P2 oligo, which produced just over half this number of mappable reads, of which only 66.9 % were mapped to unique positions. This resulted in a nearly 3 fold increase in high quality, mappable reads, which should translate to greater coverage and therefore a more robust identification of transcription start sites, especially those with lower levels of expression.

To test this hypothesis, the RH positions identified by the two libraries were filtered of positions with RH frequencies <3, then aligned and compared (as previously described). 87.5 % of transcription start sites identified by the full length oligo TAP+ALK library were also identified by the new library, indicating a high level of consistency between runs. Increasing the minimum RH frequency threshold for filtering to 20 increased the percentage

of TSS matches to 97 %, showing that nearly all highly expressed TSS positions identified by TAP+ALK library were conserved in the new, shorter P2 library. The high level of consistency between the two libraries confirmed robust TSS finding protocol. However the greater coverage of the shorter P2 library resulted in more than double the number of putative transcription start sites (66,807 compared to 30,312) being identified with RH frequency <3. As such this represented a marked improvement in library quality and a more complete TSS analysis resulting from the use of this shorter oligo.

Protocols for 5'-specific RNA-seq were therefore updated to use the shorter P2 oligonucleotide, as it had been shown to produce both higher quality fragment libraries with fewer PCR artefacts, resulting in a greater throughput of data from SOLiD sequencing and a more robust analysis of transcription start sites.

4.16. Discussion

A protocol for 5′ specific RNA-seq on the ABI SOLiD platform has been developed and used to map the transcription start sites of mRNA transcripts in *A. nidulans*. Reads from 5′ specific RNA-seq library mapped to the *A. nidulans* genome at a similar rate to previous whole transcriptome libraries, indicating similar library quality. 14 million transcript 5′ ends were mapped and used to define 15,475 TSS regions of significant mapping.

A high level of TSS diversity was observed between transcripts, with TSS regions generally extending up to 60 bp either side of the primary TSS location. This produced TSS regions up to 121 bp in length, with global analysis RH mapping within TSS showing a consistent pattern for TSS diversity both within and between chromosomes. While the vast majority of TSS regions appear to fit within these limits, some outliers with greater diversity did exist.

This is reasonably consistent with TSS region sizes identified in mice, which were shown to have an average length of 134.2 bp (Kawaji *et al*., 2006).

TSS regions have been categorised based on a statistical analysis of RH distribution. Of the 15,475 regions analysed, 3,777 (24.4 %) had well defined start sites with transcription predominantly confined to a single base pair, with very little diversity (CIL <2). Work in mice by Kawaji *et al*., (2006) also classified TSS regions based on distribution. Four classes were identified, presenting as either: a single dominant peak a general broad distribution, a broad distribution with a dominant peak, and a bimodal or multimodal distribution. Only the first class represents a narrowly defined TSS location, whereas the remaining classes are categories of broad regions with multiple TSSs. Kawaji *et al*., (2006) found that 23% of TSS regions in mice fell within the first category of narrowly defined TSSs. This is almost identical to the 24.4 % of TSS regions observed in *A. nidulans* with CI lengths <2.

Narrowly defined transcription start sites were investigated for potential consensus sequence which could guide the machinery of transcription to these very specific locations. No strong bias in sequence was observed in this dataset, indicating that there is likely to be a more distant element within the promoter which influences TSS specificity. RNA polymerase II dependant promoters of vertebrates have previously been studied and divided into two groups, categorised as either TATA or CpG types. The former has sharp and peaky TSS clusters with the peak TSS at a strict distance from the TATA box, while the latter generally has contrastingly broad TSS clusters (Suzuki *et al*., 2001; Carninci *et al*., 2006). Further work focusing on non-vertebrate promoters in *Drosophila melanogaster* found both sharp and broad TSS clusters, but while *Drosophila* possessed TATA type promoters, it did not possess CpG type (Hoskins *et al.,* 2011). These findings were consistent with those in *Arabidopsis thaliana*, where the TATA promoter is again associated with sharp TSS clusters (Yamamoto

et al., 2011). It is therefore likely that TSS regions identified as having low CI length in *A. nidulans* are associated with TATA type promoters.

Motif finding within promoter sequences surrounding TSS regions with both narrow and broad TSS distributions has been performed. Motif finding results from multiple pieces of software were combined with distribution analysis within the promoter to identify 16 candidates most likely to represent strong, functional motifs. The degree of overrepresentation of these motifs was far greater in the dataset representing narrowly defined TSS regions, further indicating functionality and supporting the hypothesis that upstream elements play a vital role in determining the position of transcriptional initiation.

Since this analysis, the 5′ specific RNA-seq protocol had been include ALP treatment and use a shorter, non-barcoded Random Priming P2 oligonucleotide. These changes were shown to produce higher quality SOLiD libraries and a more robust analysis of transcription start site locations. However, further analysis of these data using the methods previously described in this Chapter was not performed due to time constraints. It would therefore be beneficial to repeat TSS identification and promoter motif analysis using the more accurate data from a library prepared in this way.

These developments also resulted in a proposed method for the investigation of mRNA decapping and 5′ degradation. mRNA degradation is a vital process for the control of mRNA expression through rapid reduction of mRNA concentration, and has been studied in a number of organisms (Belasco & Brewerman, 1993; Ross, 1995; Abler & Green, 1996; Caponigro & Parker, 1996). Degradation of transcripts can occur through three principal mechanisms; 5'-3' exonuclease or 3'-5' exonuclease mediated degradation, or endonuclease cleavage followed by exonuclease mediated degradation (Caponigro & Parker, 1996). Overall stability of the transcript is determined by features such as primary sequence and tertiary

structure, as well as specific structures such as the 5' cap and poly(A) tail. These features cause some transcripts to be fundamentally unstable, while others are very stable, often linked with the activity of the transcript and if its action is required to be short or long lasting (Herrick *et al*., 1990). There are also many documented examples of transcripts with regulated stability (Surosky & Esposity, 1992; Klausner *et al*., 1993; Cereghino & Schaffler, 1996). AreA is an example of this, and its degradation has been well characterised in *A. nidulans* (Morozov *et al*., 2000). Degradation is generally preceded by poly(A) shortening (Caponigro & Parker, 1996), which triggers decapping when the poly(A) tail is shortened to around 15 residues and facilitates rapid degradation of the transcript by 5'-3' endonucleolytic decay. In *A. nidulans*, decapping is often preceeded by modification of the shortened poly(A) tail with the addition of CUCU, which has been proposed to act as a precurser to decapping (Morozov *et al*., 2012). The exact mechanism by which this modification influences degradation is still unknown, as is the precise order of events in which degradation occurs. Investigation of mRNA degradation products in mRNA-seq data may therefore yield valuable insight into this process, by providing a snapshot of all degradation products in a given sample of Total RNA.

Initial investigation of the resultant data indicated significant inconsistency between degradation products identified by two different RNA-seq libraries. This may be due to the progressive nature of exonuclease degradation, the products of which are randomly sampled by RNA extraction at a single time point. However, positions that were identified in both libraries may represent specific endonuclease cuts, or points at which exonuclease activity is paused. While these data were therefore extremely interesting, further and more detailed analysis was required to fully understand the process of 5′ degradation in *A. nidulans*.

Work performed by Cowley (2012), conducted under my supervision, utilised motifs identified in this thesis to perform global analysis of motif distribution. We designed software

to perform motif alignment and association with DNA features in the *A. nidulans* genome. The results of this analysis showed statistical significance of motif distribution throughout DNA features, further supporting the hypothesis that they are functional elements within the promoter sequence and throughout the genome. In addition to the enrichment previously observed in promoter regions (Chapter 4.12), motif enrichment was also observed downstream of transcription start sites and within introns. Investigating these sequence motifs in more depth at exon-intron junctions may link these to alternative splice sites or indicate roles in the regulation of splicing events (Holse & Ohler, 2008). Cowley (2012) also performed analysis of motifs associated with annotated gene models, using gene ontology (GO) to search for enrichment of these genes within specific functions, pathways or components of the genome. Genes associated with two motifs; AARARAAA and TTTYTTY, showed significant enrichment in DNA response and DNA/RNA metabolic processing respectively, indicating distinct biological functions.

Chapter 5

5. New whole transcriptome libraries

5.1. Introduction

Since the sequencing and analysis of the SOLID RNA-seq libraries described in Chapter 3, significant advances were made in both library preparation and SOLiD sequencing. In addition to increased bead density, new slide chemistry for greater bead retention during sequencing and improved algorithms for detecting and distinguishing bead signals during ligation cycles enabled higher throughput. Optimisation of the ligation chemistry also led to significant improvements in accuracy and performance compared to previous SOLiD versions. These enhancements enabled the SOLiD 4 system to generate up to 100 GB of mappable sequence or 1.4 billion reads per run, with an increased read length of 50 bp and accuracy of >99.94% (Ichikawa *et al*., 2010).

To take advantage of these developments in sequencing technology, new whole transcriptome libraries were created using the latest reagents and protocols. The SOLiD Total RNA-seq kit had replaced both the Whole transcriptome Analysis and Small RNA Expression kits, providing streamlined and more robust protocols for library creation. Using this kit in combination with the SOLiD RNA barcoding kit, samples could be multiplexed and run together on the same slide, lowering the sequencing cost per sample. We were therefore able to produce more libraries and significantly more data compared to previous sequencing runs.

A wide range of new read mapping software had also recently been released or adapted to support colour space reads produced by SOLiD sequencing. This included Tophat, a splice-junction mapper for RNA-seq reads, which gained colour space support in release 1.1.0 on 10/03/2010. Tophat aligns RNA-Seq reads to genomes using the ultra high-throughput short read aligner Bowtie (Langmead *et al*., 2009), and then analyzes the mapping results to identify splice junctions between exons. The ability to split reads and map them across splice

junctions allowed Tophat to accurately define the intron-exon structure of genes, locate previously unreported junctions and identify potential occurrences of differential transcript splicing (Trapnell *et al*., 2009). This not only provided more accurate identification of gene structure, but gave greater scope for the investigation of transcript splicing, making Tophat the preferred method of read mapping for this analysis.

5.2. Library preparation and sequencing

Using RNA extracted from G00 wild-type *A. nidulans* cultures grown in a variety of conditions, we aimed to sequence transcripts from a wide range of expression profiles, therefore capturing as much of the transcriptome as possible. Two of the conditions were repeats of the previous S1 and S3 libraries, made using RNA from culture grown on either minimal media with nitrate as sole nitrogen source, or on complete media. Another library was created from culture grown on minimal media with ammonium replacing nitrate as an alternative nitrogen source. These three conditions therefore represented A. nidulans grown on abundant nutrients, a sole primary nitrogen source, and a sole secondary nitrogen source. The uptake and utilisation of different nutrients requires unique cellular mechanisms which are tightly regulated by a number of genes associated with each metabolic pathway (Caddick *et al*, 1994) and will therefore produce different expression profiles. Two additional libraries were created from cultures grown initially on minimal media with nitrate, then transferred to minimal media with no nitrogen supplement for either 4 or 72 hours (see Chapter 2.2.4 for detailed growth conditions of all cultures). The two lengths of nitrogen starvation aimed to produce different gene expression profiles from alternative points in the cell's stress response, which has been shown to activate in low nitrogen conditions (Etxebeste *et al*., 2010).

RNA extracted from these cultures underwent poly-A selection, 5′ decapping, and whole transcriptome library preparation with the SOLiD Total RNA-seq and SOLiD RNA barcoding kits (ABI), following the manufacturer's protocol. Completed libraries were submitted to the Liverpool Centre for Genomic Research (CGR) for SOLiD sequencing. Following quality control checks performed by the CGR, libraries were pooled to create a single sample containing all libraries in equal concentration. The multiplexed libraries were then sequenced on the SOLiD 4 sequencing platform. Reads associated with each barcode are separated by software on the SOLiD machine, giving individual outputs for each sample. The number of reads produced by each library is displayed in Table 5.1.

Table 5.1. Whole Transcriptome libraries sequenced on the SOLiD 4 system. Growth condition and number of reads produced by SOLiD sequencing is given for each library

Library condition	Number of reads
Minimal medium + nitrate	60,003,026
Complete medium	60,257,046
Minimal medium + ammonium	46,929,150
4 hour nitrogen starvation	71,009,476
72 hour nitrogen starvation	65,995,033

5.3. SOLiD read mapping

The *A. nidulans* genome assembly (version CADRE 2.5) was downloaded via the ensemblgenomes ftp server (ftp://ftp.ensemblgenomes.org/pub/) and used to assemble a Bowtie compatible FASTA index (Chapter 2.4.2).

To ensure the relative quality of each fragment library was consistent between samples, a preliminary test of mapping quality was performed with Bowtie alone. Reads were trimmed at the 3′ end to a final length of 35 bp as the error rate in SOLiD reads increases rapidly with each additional base position beyond 35 bp from the 5′ end (Applied Biosystems, 2008). Stringent mapping of the trimmed reads against the *A. nidulans* genome was performed in Bowtie, allowing 0 mismatches per read in order to best assess sequencing quality for each sample. Non-uniquely mapping reads were assigned randomly to a single position to prevent total read mapping statistics from being artificially inflated by reads mapping to multiple positions. Using these settings, all libraries mapped to the *A. nidulans* genome at a low rate, but demonstrated excellent balance of reads mapping to each strand. Percentage read mapping between samples was also reasonably consistent, with only the 72 hour nitrogen starvation library showing significantly increased mapping compared to other samples (Table 5.2).

Table 5.2. Read quality assessment mapping. Read distribution between forward and reverse strands was even for each sample and mapping figures showed a high level of consistency between samples, with only the 72 hour nitrogen starvation sample showing a significantly higher level of mapping.

Library condition	Number of reads	Mapped to forward strand	Mapped to reverse strand
Minimal medium + nitrate	60,003,026	2781874 (4.6%)	2810440 (4.7%)
Complete medium	60,257,046	3523898 (5.8%)	3501683 (5.8%)
Minimal medium + ammonium	46,929,150	2122804 (4.5%)	2142050 (4.6%)
4 hour nitrogen starvation	71,009,476	4098809 (5.8%)	4095745 (5.8%)
72 hour nitrogen starvation	65,995,033	6990113 (10.6%)	7145774 (10.8%)

Full mapping of reads from newly sequenced samples was performed with Tophat in conjunction with Bowtie using basic settings (methods 2.4.3). No reference genes were supplied to Tophat for this mapping. Had gene models been provided, Tophat would have extracted the transcript sequences and assembled an artificial transcriptome, then use Bowtie to map reads preferentially to this construct, before mapping to the rest of the genome. This would have led to a strong mapping bias at the loci of current gene models, which our previous data had shown to be incomplete and of poor quality. By mapping without a gene model annotation we aimed to remove this bias and produce a more accurate representation of the *A. nidulans* transcriptome.

Resulting BAM files were indexed and sorted using SAMtools to facilitate visualisation of mapped reads in the Broad Institute's Integrative Genomics Viewer (IGV) software (Robinson *et al*., 2011; Thorvaldsdottir *et al*., 2012). IGV is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets, capable of displaying both reads which map across splice junctions, and separate tracks to show junctions predicted by Tophat (Robinson, *et al.*, 2011; Thorvaldsdottir *et al*., 2012).

Visual analysis of the mapped data in IGV displayed a large number of Tophat-defined splice junctions which correspond to those in the annotation, but included many more which spanned extremely long genetic regions, often traversing several genes (Fig. 5.1). One of the ways Tophat identifies putative splice junctions is by splitting reads which partially align to the genome into two segments, mapping the partial alignment and the rest of the read independently. If the second segment is mapped downstream of the first, Tophat identifies the gap between them as a splice junction (Trapnell *et al*., 2009). The high frequency and obvious error of these extremely long junctions would have led to difficulties and inaccuracy in data analysis. This led to a reassessment of mapping criteria to reduce the occurrence of

this phenomenon, and subsequent remapping of each sample to obtain more accurate junction predictions.

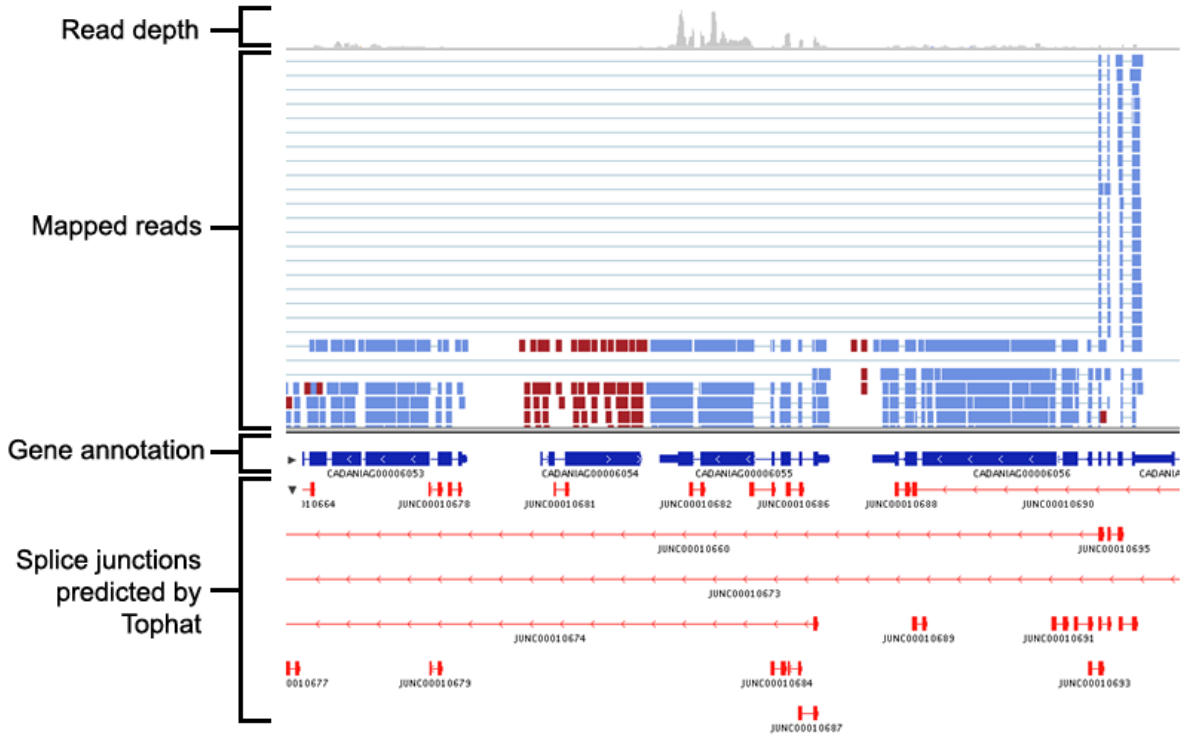


Figure 5.1. Tophat based mapping of extremely long splice junctions displayed in IGV. While the majority of splice junctions predicted by Tophat fit the annotation, a large number of junctions (shown as horizontal red lines in bottom track) were predicted to span several genes (genes are shown in blue in the gene annotation track). This was due to segments from split reads mapping at distant loci and Tophat defining the gap in between as a splice junction. Instances of this read splitting are shown in the mapped reads track, blocks of red (forward strand) and blue (reverse strand) indicate reads aligned to the genome, while the horizontal lines indicate gaps between split reads.

5.4. Assessment of Tophat junction mapping

To address the issue of excessively long junction finding, it was necessary to limit the maximum intron length when using Tophat to perform read alignment. The default maximum intron length in Tophat is 500,000, and when searching for junctions *ab initio*, TopHat will ignore donor/acceptor pairs farther than this many bases apart. This is many times larger than any known junction found in the *A. nidulans* genome and provided a reason for Tophat reporting junctions which spanned unrealistic distances and across multiple genes.

As a test of this hypothesis, Tophat mapping was performed with a maximum intron length of 21,000, the size of the largest gene in the current annotation. Being many times smaller than the default 500,000 bp length, this greatly reduced the number of extremely long introns found by Tophat. However, a number of extremely long introns persisted.

To determine an appropriate limit for intron length in *A. nidulans,* currently annotated intron lengths were used as a starting point. Software was written to extract a list of all intron lengths from the current *A. nidulans* gene annotation. While the software output contained the length of every intron in the *A. nidulans* genome, further processing was required to extract meaningful data. To identify the range of intron sizes, the software was updated to sort the values in the intron length array by size. This produced an output file of intron lengths listed from shortest to longest, assisting the rapid identification of the range of intron sizes in *A. nidulans* and the maximum intron size found (Chapter 2.5.11)

To gain an overview of all intron lengths, graphical representations of the resulting dataset were produced in R (R Development Core Team, 2008) (Fig. 5.2). The box and whisker plot produced indicated an extremely low interquartile range with a large number of outliers. To provide an alternative view of the data, a histogram was also created in R (Fig. 5.3), providing a clearer indication of the number of introns of each length.

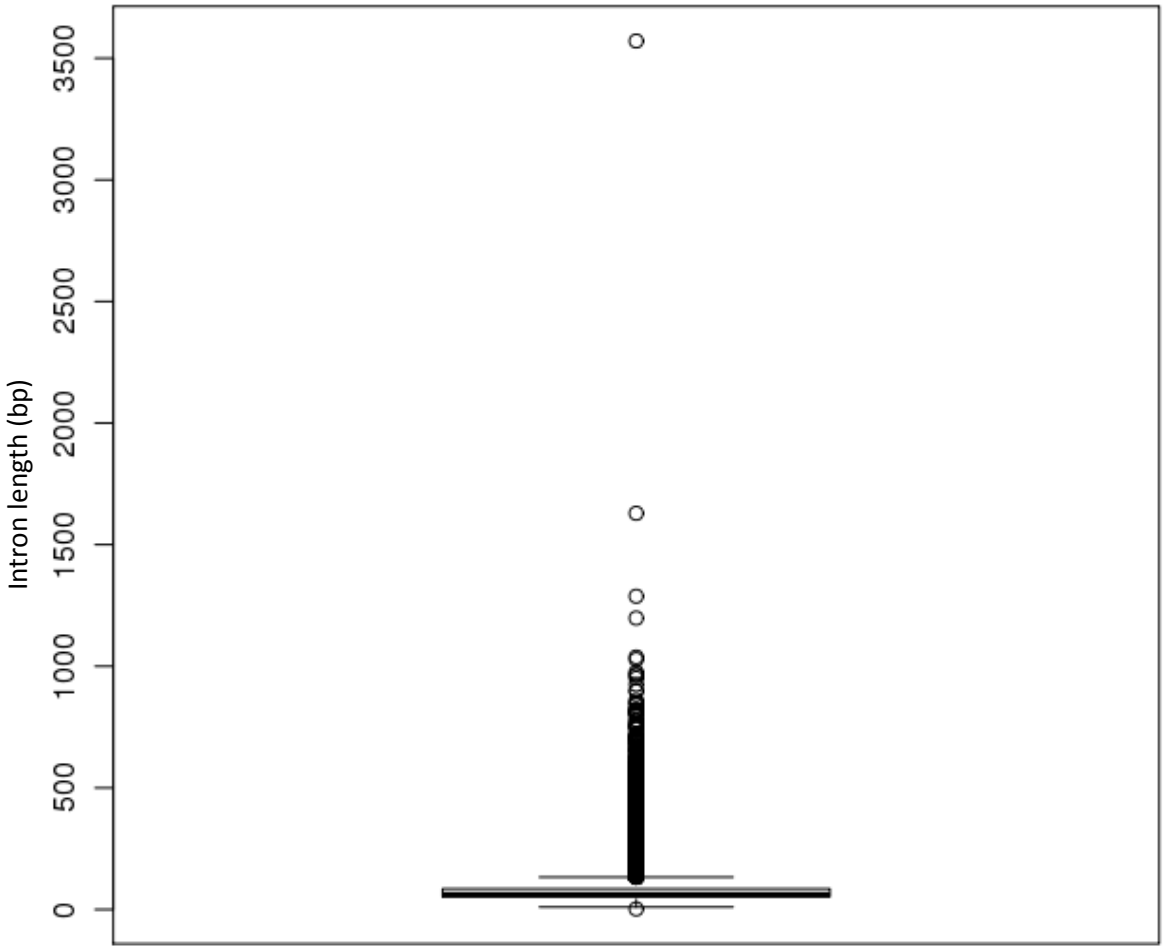


Figure 5.2. Box and whisker plot of intron lengths in *A. nidulans*. The length of all introns is plotted, with the five number summary indicated by the box and whiskers. The box indicates the lower quartile, median and upper quartile, while the whiskers indicate the lowest and highest values determined by statistical methods, indicating that the majority of introns fall within this length range. The circles represent lengths determined to be statistical outliers, meaning that the majority of intron lengths fall within the boxed region around 40-100 bp, with the median value being 82. However, there is a significant number of outliers observed up to approximately 1200 bp in length, beyond which only very few introns are observed.

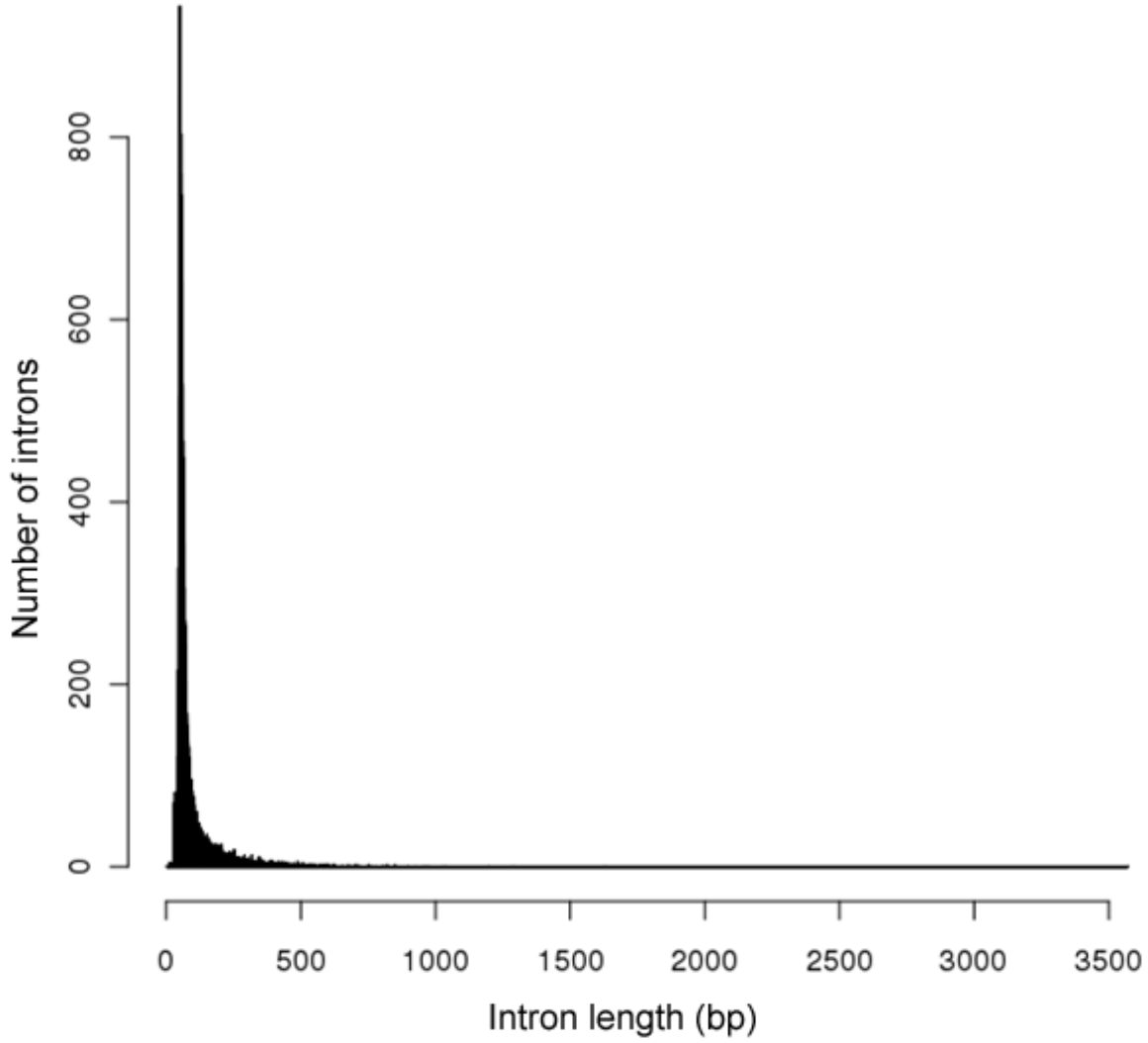


Figure 5.3. Histogram of intron lengths in *A. nidulans.* The number of introns of each length is shown, indicating the vast majority fall within the 40-100 bp length range as was suggested by Fig. 5.2.

Annotated intron length in *A. nidulans* was shown to range between 2 and 3571 bp. However, from Fig. 5.2 and Fig, 5.3 it was apparent that nearly all annotated introns are < 1200 bp, with the dataset containing a small number of larger outliers. The largest intron was more than double the length of any other at 3571 bp, raising questions about the validity of this annotation. To identify the gene containing this intron, the software was updated to record the ID of genes as they were processed and record the ID of the gene containing the largest intron. The finished version of this software is described further in Chapter 2.5.11.

The modified software was used to identify the largest intron as being annotated in gene AN4390.4 (CADANIAG00006067). While our RNA-seq data showed no evidence of an intron at this locus (Fig. 5.4), it is still possible that introns of this length would exist in *A. nidulans*. Lowering the maximum intron length beyond this point would potentially have prevented real introns from being found by Tophat. A maximum intron size therefore had to be over 3571. An optimal limit of 5,000 was selected as it appeared to give excellent mapping results on visual inspection in IGV, while being appropriately large so as not to exclude any real junctions of greater than average length. Maximum intron lengths below this value were also tested, however this caused no appreciable improvement in mapping quality and only a minor decrease in junctions found (32 fewer were found with maximum intron length of 4,000).

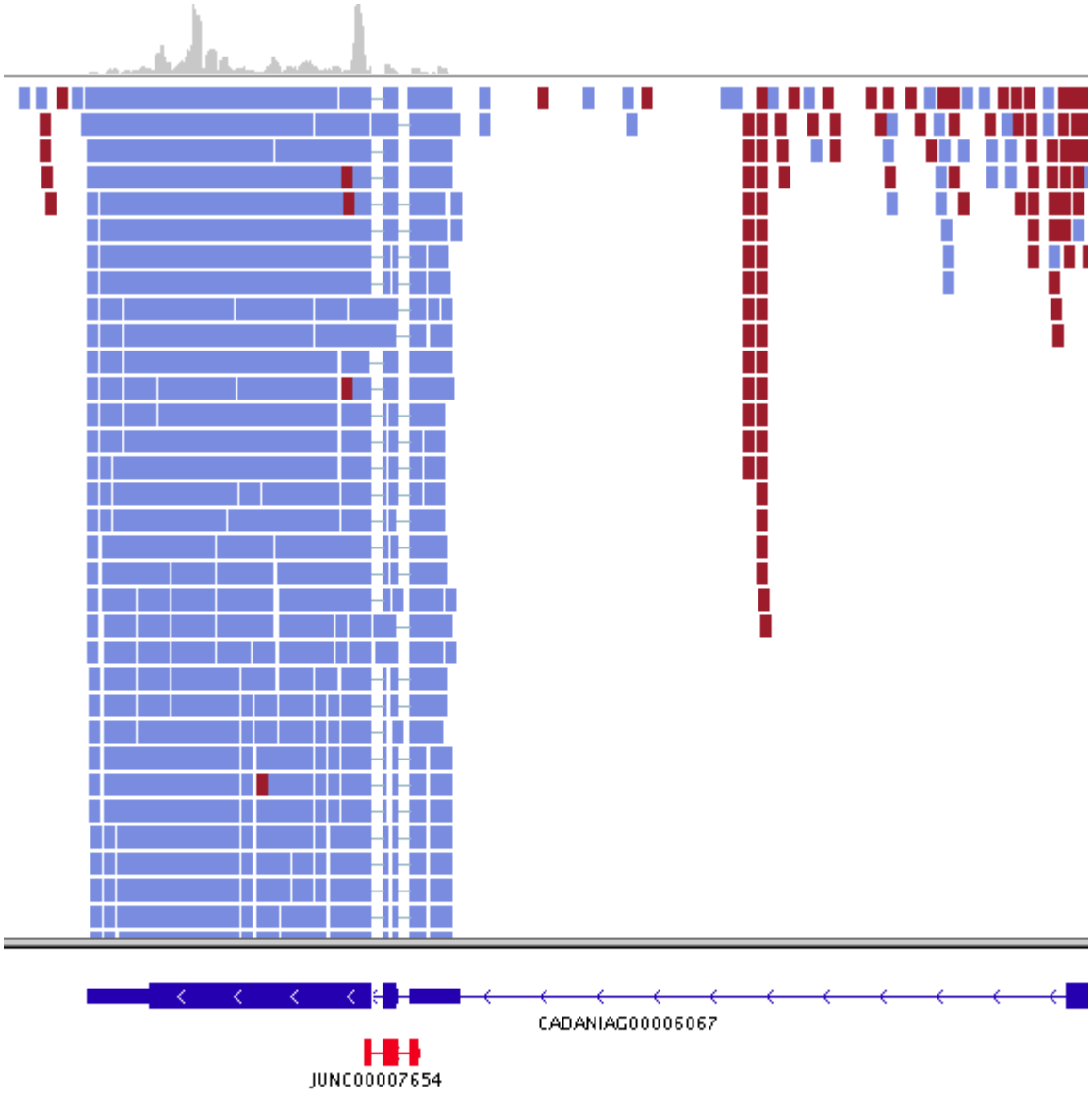


Figure 5.4. RNA-seq and junction data for gene AN4390.4 (CADANIAG00006067).

Tophat aligned total RNA-seq reads displayed in IGV is shown, including read coverage (top track), individual for the forward (blue) and reverse (red) strands (middle track). The bottom track shows the annotated gene in this region (blue) and the splice junctions predicted by Tophat (red). This confirmed the presence of 2 small introns, however there was no evidence for the existence of the large 3571 bp intron at this locus.

More recent annotations of the *A. nidulans* genome which were released subsequent to this analysis no longer contain an intron at this locus. Studies into the structure of AN4390 have shown the gene to start at the exon displayed before the first confirmed junction in Fig. 5.4 (de Groot *et al*., 2009).

5.5. Modified Tophat mapping

In order to produce a more accurate representation of the *A. nidulans* transcriptome, mapping of all five libraries sequenced on the SOLiD 4 system was repeated using Tophat in conjunction with Bowtie, with maximum intron length set to 5,000 bases.

A mapping run was also performed using all five libraries as input. This combined run aimed to produce the maximum possible depth of coverage and widest expression profile. Table 5.3 contains the alignment and junction finding results for each library.

Table 5.3. Mapping and splice junction identification by Tophat with a maximum intron length of 5000 bp. Total read mapping is displayed in column 2, and the number of unique mapping reads in column 3, also expressed as a percentage of total reads sequenced in each library. The number of unique splice junctions identified by Tophat is displayed in column 3.

Sample	Reads mapped (million)	Unique mapping reads (million)	Splice junctions
Minimal medium + nitrate	13.8 (22 %)	12.0 (20 %)	10232
Complete medium	12.0 (20 %)	11.5 (19.2 %)	11549
Minimal medium + ammonium	8.0 (17 %)	7.4 (15.9 %)	9995
4 hour nitrogen starvation	14.0 (20 %)	13.2 (18.7 %)	12247
72 hour nitrogen starvation	20.4 (31 %)	19.8 (30.0 %)	15352
Combined	76.7 (25.2 %)	73.3 (24.2 %)	20988

No gene model annotation was provided to Tophat and Bowtie for any of the mapping runs shown in Table 5.3. As a form of control, a single mapping run was conducted using the minimal media + nitrate library with a supplied gene model annotation in GTF format. Tophat uses a gene model annotation to guide read mapping preferentially to these regions, and supplying this annotation resulted in a significant reduction in both total and unique read mapping, and also in the number of splice junctions identified. It was theorised that this observation was caused by preferential mapping to annotated gene models, reducing the number of reads mapped to other locations. A third mapping run was performed with Tophat, to further investigate this issue, using the GTF annotation in conjunction with the "—no-novel-juncs" option, which prevented Tophat from reporting any junctions which were not supported by the gene models in the GTF annotation. The results of these runs are shown in Table 5.4.

Table 5.4. Mapping and splice junction identification by Tophat with alternative settings. Three Tophat mapping runs were performed for the same SOLiD read library, using different settings for each run. For the first run listed, no GTF gene annotation was supplied. The second run was guided by the GTF but still allowed the discovery of novel junctions. The final run did not report any junctions which did not match to those in the GTF annotation. Total read mapping, uniquely mapped reads and the number of reported splice junctions is given for each run.

Tophat settings	Reads mapped (million)	Uniquely mapping reads (million)	Splice junctions
No GTF	13825877	12044976	10232
GTF	10168951	9544085	7827
GTF and –no-novel-juncs	10160824	9546179	10216

While both mapping runs which were guided by the GTF annotation produced similar read mapping figures, the number of splice junctions reported increased when no novel junctions were permitted. To investigate the implications of these data, software was written to compare the junction positions reported by each mapping run. By comparing each set of junctions from the run which allowed novel junctions, to those which included junctions found in the GTF, the number of novel junctions reported could be determined in each case.

Initially this software compared the positions of junction features from two junctions.bed files produced by Tophat, however this produced extremely high numbers of mismatches between datasets. The software was then rewritten to use feature length data contained in another part of the junctions.bed files to adapt the feature positions and obtain the precise junction start and stop sites. The finished software is described further in Chapter 2.5.12. While this significantly increased the accuracy of the analysis, the number of mismatches between runs was still high. Despite the low number of junctions reported by the GTF only run, 2282 of these were shown to not match those of the annotation, while this number increased to 5765 in the run where no GTF was supplied.

These results imply that the use of a GTF annotation hinders the depth of read mapping to previously unannotated regions of the genome, and therefore the discovery of novel genes and splice junctions. All future work into novel feature discovery therefore utilised Tophat mapping where no gene model annotations had been supplied.

5.6. Genome utilisation indicated by Tophat mapping of RNA-seq reads

Investigating the proportion of the *A. nidulans* genome to which reads could be aligned gave an indication of genome utilisation, and the breadth of transcriptional activity. Studies in

other organisms have previously indicated extremely broad transcription across their respective genomes. At least 93% of the human genome is thought to be transcribed (The ENCODE project consortium, 2007; Cheng, *et al*., 2005; Bertone *et al*., 2004), and similar levels of transcription have been observed in the mouse genome (The FANTOM Consortium, 2005).

The proportion of the *A. nidulans* genome being transcribed was calculated based on the number of bases to which reads had been aligned. To access this statistic, the mpileup feature of SAMtools was used to convert read alignment data from Tophat-produced BAM files, into a pileup format summarising the base calls of aligned reads to the reference FASTA. The resulting pileup file contained a newline-separated list of only those reference bases covered by aligned reads, excluding bases where no alignment occurred (Chapter 2.4.5). The number of these bases was determined using a linecount of the pileup file, and used to calculate the percentage of the genome being transcribed.

The analysis was repeated for a combined dataset of RNA-seq reads from all five libraries. By combining the reads from different conditions, we aimed to produce a dataset in which the maximum number of genes were being expressed, and therefore a more accurate assessment of the proportion of the genome which could be transcribed in *A. nidulans.*

With RNA-seq reads from all 5 libraries combined into a single dataset, Tophat and bowtie were able to align reads to 24096731 bp of reference genome, indicating that 80.1% of the genome is transcribed as mRNA. This seemed extremely high, as previous analysis of the *A. nidulans* genome predicted 51.5% total protein coding DNA (Galagan *et al*., 2005). Reflection on the cause of this discrepancy revealed an error in the analysis, as the pileup format displays alignments from both strands on a single line. This analysis therefore showed

the proportion of the genome transcribed on either strand, rather than being nucleotide specific, thus artificially inflating the amount of the genome which appeared to be transcribed.

To circumvent this issue, the fifth column of each pileup file was isolated. This column contains information on match, mismatch, indel, strand, mapping quality, starts and stops of reads aligned to a given position. By searching these data for characters indicating the strand to which reads had been aligned at each position, a count of strand specific alignments was performed.This produced far more accurate estimates of genome utilisation (Table 5.5).

Table 5.5. Nucleotide specific alignments to the A. nidulans genome and proportion of the genome transcribed. Nucleotides to which RNA-seq reads had been aligned on each strand were counted. These figures were combined to give the total proportion of the genome transcribed in each condition.

Sample	Bases expressed on forward strand	Bases expressed on reverse strand	Proportion of genome transcribed
Minimal media + nitrate	9237980	9316688	18554668 (30.8%)
Complete media	8882344	8932957	17785301 (29.6%)
Minimal media + ammonium	8218262	8285336	16503598 (27.4%)
4 hour nitrogen starvation	9254205	9314739	18568944 (30.8%)
72 hour nitrogen starvation	11679565	11783429	23462994 (39.0%)
Combined	14218016	14356861	28574877 (47.5%)

The proportion of the genome found to be transcribed by this analysis was far more in line with previous reports. Libraries were created with poly-A selected mRNA, meaning rRNA was not represented. rDNA in *A. nidulans* has previously been defined as a tandemly repeated 7.8 kb unit (Borsuk *et al*., 1982) with a copy number of ~45 (Ganley & Kobayashi, 2007), adding another 351 kb of transcribed DNA to this total. Small RNAs were also excluded due to the 50 bp read length of whole transcriptome RNA-seq libraries, as sequences shorter than 50 bp result in a high number of mismatches during alignment and the read being rejected as low quality. Sequencing studies of sRNAs in the filamentous fungus *Neurospora crassa* have been conducted using high-throughput techniques on the Illumina/Solexa genome analyser and 2,180,272 sRNA (17-30 nt) sequences which matched the *Neurospora* genome assembly (accession number GSE21175) were identified (Lee *et al*., 2010). The genome of *Neurospora crassa* has been reported as completely sequenced in 2003 (Galagan *et al*., 2003). Current estimates suggested that the 41 megabase genome encodes 9,733 genes. While this genome was found to be approximately one third larger than that of *A. nidulans*, the amount of sRNA transcription in *Neurospora* indicates that a significant increase in genome utilisation would be observed should sRNAs be included in this analysis.

Significant discrepancies were observed between the number of bases aligned to RNA-seq reads on both strands (table 5.6), compared to strand specific alignments based on individual nucleotides (table 5.5). The higher numbers observed in the latter analysis indicate a large number of bases with reads aligned to both forward and reverse strands. For the combined dataset, 4,478,146 bases were observed where transcription appears to be occurring on both strands. This supports previous estimates of antisense transcript numbers in *A. nidulans* from RNA-seq data (Chapter 3.4).

5.7. Discussion

The mRNA transcriptome has been sequenced for *A. nidulans* cultures grown in 5 different conditions on the SOLiD version 4 platform. Alignment of RNA-seq reads to the reference genome with Tophat and Bowtie has been performed, with a maximum intron length of 5,000 to prevent the identification of incorrect splice junctions, and using no reference gene model to achieve the highest possible level of mapping and novel feature discovery.

Compiling individual libraries into a single dataset shows a significant increase in the proportion of the genome being expressed, indicating variance in gene expression profiles between samples grown in different conditions.

A high level of transcription has been observed, with approximately 47.5% of the genome aligned to RNA-seq reads, indicating broad transcriptional utilisation of the *A. nidulans* genome, in keeping with findings from previous sequencing studies. These data also indicate a significant level of antisense transcription, supporting previous estimates and warranting further investigation. However, this analysis did not take into account regions such as introns where no mapping should occur, and the fact that these libraries will not include sRNAs. The total coverage may also be affected by the uneven mapping profile observed due to the fragmentation method employed during SOLiD library preparation (as discussed in Chapter 3.5), however the increased read length and depth of coverage in these newer libraries appears to drastically reduce the number of bases in expressed exonic regions where the frequency of read mapping is reduced to 0 due to fragmentation bias. It is therefore expected that any effect this bias has on overall genome coverage will have been minimal.

Additionally, Tophat mapping with alternative GTF annotation settings indicated that read mapping to junctions was guided to a greater degree when novel junctions were not permitted. However, the depth of mapping was greatly reduced when a GTF annotation was supplied. A

potential workaround for this could be to first perform Tophat mapping with no reference annotation, thereby achieving the greatest possible depth of coverage and identifying splice junctions based purely on RNA-seq data. The resulting junction data could then be used to build an annotation that could be then supplied to Tophat to guide a second mapping run, potentially increasing the accuracy of mapping across novel splice junctions.

Chapter 6

6. Identification of putative novel and antisense genes in *Aspergillus nidulans*

6.1. Previous observations

Whole transcriptome RNA-seq data from early SOLiD libraries S1 and S3 had been aligned to the *A. nidulans* reference genome using Corona lite. Visual analysis of these data in the Artemis genome browser had revealed a high level of read mapping to intergenic regions, indicating the existence of a large number of putative novel genes at these loci (Chapter 3.4). Many of these genes appeared to be located in antisense configuration with existing gene models, and an estimated 5-18 % of annotated genes were associated with some form of antisense transcription. To further investigate these observations, estimation of the number of novel transcripts was performed using early RNA-seq data produced on the SOLiD version 2 platform (Chapter 1). Subsequent developments in RNA-seq technology and data analysis tools led to more robust identification and analysis of novel transcripts, and facilitated further analysis of putative antisense relationships between transcripts.

As these putative novel genes were not detected by previous gene prediction analysis, it is likely that the majority will be non-coding. Non-coding RNAs fall into two broad categories: regulatory and house-keeping. Regulatory ncRNAs are often expressed during development or in response to environmental conditions such as nutrient defficiency, and are often restricted to subcellular compartments based on their function (Prasanth & Spector, 2007; Clark & Mattick, 2011; Tisseur *et al*., 2010). House-keeping ncRNAs on the other hand are constitutively expressed and include small nuclear, small nucleolar, ribosomal, transfer and spliceosomal RNAs (Wright & Bruford, 2011). This means that while ncRNAs do not encode proteins, they still contain information and perform functions vital to cell viability, and have many features similar to protein coding genes, including the ability to be differentially spliced

or polyadenylated, and either nuclear or cytoplasmic (Ponting *et al*., 2009; Amaral *et al*., 2011).

Of particular interest are NATs (as previously discussed) and ncRNAs involed in the RNA interference (RNAi) pathway, which in eukaryotes is triggered by the formation of double stranded RNA, causing short interfering RNAs (siRNA) to mediate post-transcriptional gene silencing (PTGS), maintain genome integrity and protect against viruses (Taft *et al*., 2010). A range of these small RNAs have been reported in fungi, most notably in *Neurosporta crassa* (Lee *et al*., 2010), and through bioinformatic analysis of a wide range of fungal genomes for the conservation of key proteins involved in RNAi (Nunes *et al*., 2011). siRNAs have also been shown to trigger PTGS through RNAi-pathways *N. crassa* and *C. neoformans*, viral suppression in *A. nidulans*, *M. oryzae* and *C. parasitica*, and heterochromatin formation in *S. pombe* (Schumann *et al*., 2010; Li *et al*., 2010; Dang *et al*., 2011). However, the non-coding nature of these transcripts makes global analysis extremely difficult *in silico*. RNA-seq is therefore a potent tool for the global discovery and analysis of ncRNAs both in *A. nidulans* and other systems.

6.2 Initial estimation of novel genes using SOLiD V2 RNA-seq libraries

Based on the observations discussed in Chapter 6.1, software was developed to detect potentially novel genes from mapped RNA-seq data. This software utilised coverage data in pileup format, produced as previously discussed (Chapter 3.4).

Identification of putative novel genes using these pileup data was performed in three distinct steps, with separate pieces of software designed for each. The first piece of software (Chapter 2.5.13) used the current gene annotations in .gbk format to augment the pileup data with an additional column indicating which bases were covered by an existing gene model. As

previously discussed in Chapter 3.4, visual analysis of the aligned RNA-seq data had indicated a significant number of poorly annotated genes, many of which were shown to extend beyond their annotated start and stop sites. To prevent these extended regions from being identified as putative novel genes, all gene models were extended by 500 bp at each end. This process of "masking" regions beyond the annotated gene model was intended to remove any regions of significant transcript mapping which were close enough to an existing gene as to potentially be associated with that gene, thereby minimising the identification of false positive novel genes. The 500 bp distance used was determined by testing a number of different distances and inspecting the resultant data in Artemis to see if all reads that appeared to associate with existing genes had been removed. The software used these extended gene models and the original RH data pileup to create an augmented pileup with two additional columns. These columns indicated chromosome position and coverage (or lack thereof) of each position by an extended gene model. Lines of these data which indicated RH mapping in regions not covered by existing gene models were written to a new pileup file.

A second piece of software (Appendix 2.1.13) then removed any lines from this new pileup file where RH coverage was below a user defined threshold. This threshold was required to remove positions where low level, random alignments had occurred, which could have affected novel gene model assembly.

The base positions in the resulting output corresponded to those where significant RH coverage was found in intergenic regions, and at a suitable distance from existing gene models to identify them as putative novel transcripts. A final piece of software (Appendix 2.1.14) was developed to search these data for runs of sequential positions which could represent a putative novel transcript. The minimum length of transcript which the software would report was set to 200. This figure was chosen as it was long enough to ensure that one or two randomly mapped reads would not be combined to form a putative novel transcript,

while being as short as possible to minimise the exclusion of real, short transcripts. A maximum distance between two bases for them to be considered part of the same transcript was also used, which accommodated the presence of introns and regions of low mapping due to fragmentation bias as previously discussed. As bases with low alignment frequency had already been removed, this distance was set to 500 bp, allowing a large gap between bases of significant mapping frequency to identify as many putative novel transcripts as possible, while not being so large that a significant number of these transcripts would be joined together.

Initial estimations of the number of putative novel genes indicated by intergenic alignment of SOLiD RNA-seq reads were performed using the S1 whole transcriptome library (minimal media + NO_3). As with previous investigations involving read frequencies, the minimum read coverage for a position to be reported was set to three to reduce the effect of random background alignments. 679 potential novel genes were identified; the experiment was then repeated with the minimum read coverage set to five. Increasing the minimum coverage to this point ensured that only regions with a high frequency of read alignment were reported as containing putative novel genes. Using these settings, the number of predicted genes fell to 418, however the ratio of genes reported between each chromosome remained fairly consistent (Table 6.1). The full outputs from the software used in this analysis can be found in Appendix 2.9.

Table 6.1. Putative novel genes reported in the S1 SOLiD library using novel software. Software was run with a minimum read frequency per base of 3, then repeated with a minimum coverage of 5 to ensure only intergenic regions with a high level of read mapping were reported as producing putative novel transcripts.

Chromosome	Novel genes (minimum coverage = 3)	Novel genes (minimum coverage = 5)
1	77	47
2	82	55
3	107	75
4	66	38
5	55	26
6	61	41
7	115	73
8	116	63
Total	**679**	**418**

6.3. Transcript assembly with Cufflinks

The switch to Bowtie and Tophat as the preferred method of RNA-seq read alignment produced aligned read lists in BAM format. This was the required format for the "Cufflinks" transcript assembly software, which was used to predict transcript structure from aligned whole transcriptome RNA-seq reads from libraries sequenced on the SOLiD version 4 platform.

Cufflinks constructs a parsimonious set of transcripts that aim to explain the observed alignment of reads produced by RNA-Seq experiments. To do so, Cufflinks reduces the comparative assembly problem to a problem in maximum matching in bipartite graphs. The software then implements a constructive proof of Dilworth's Theorem, which characterizes the width of a partially ordered dataset (Dilworth, 1950). It does this by constructing a covering relation on the read alignments, then finding a minimum path cover on the directed acyclic graph to explain this relation. Cufflinks tries to find the correct parsimonious set of transcripts by performing a **minimum cost** maximum matching. The cost of associating splicing events is based on the "percent-spliced-in" score developed by Wang *et al*. (2008) (Trapnell *et al*., 2010)**.**

Initial, trial assemblies were performed using the SOLiD whole transcriptome RNA-seq library prepared from RNA extracted from *A. nidulans* culture grown on MM + NO_3. Cufflinks predicted 13742 transcripts expressed in this growth condition. However, visualisation of the assembled transcripts in the IGV indicated a large number of transcripts which did not fit to the aligned read data. Closer investigation showed that Cufflinks was using read alignment data from both strands of DNA to assemble a single, strand-specific transcript (e.g. Fig.6.1). These errors were unexpected as the format of the RNA-seq library had been explicitly defined within the Cufflinks input.

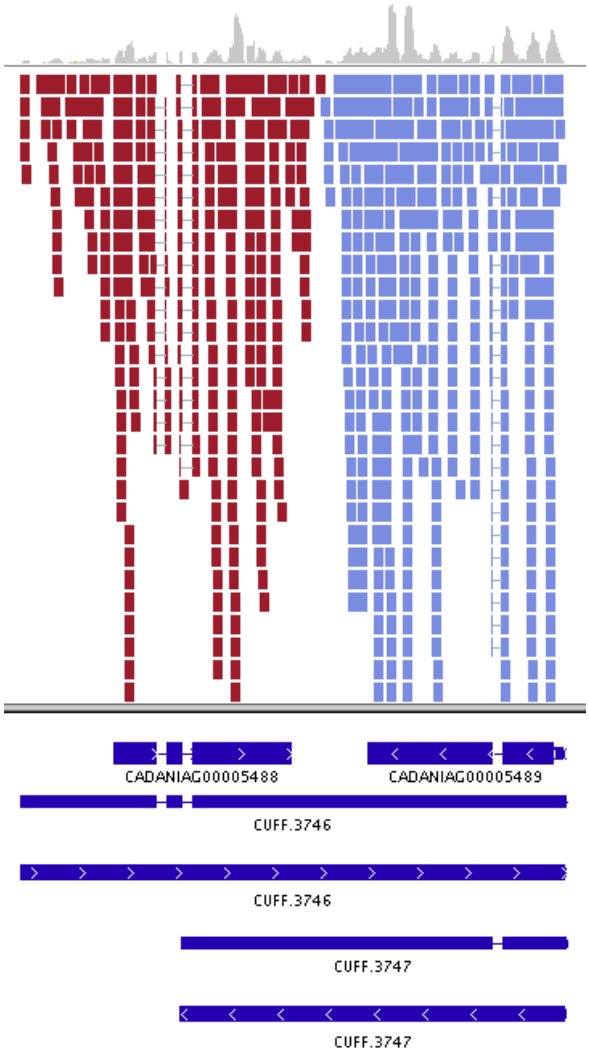


Figure 6.1. Cufflinks assembly of transcripts using aligned reads from both strands. Strand specific read alignments (red for forward strand, blue for reverse strand) is shown in IGV (top track), along with the existing gene model annotations for the genes in this region (CADANIAG00005488 and CADANIAG00005489). Transcripts assembled by Cufflinks are shown (thinner blue tracks) below this annotation. CUFF.3746 indicates a transcript on the forward strand which runs through the entire length of both gene annotations. This transcript therefore appears to have been assembled from the reads on both strands. The 3' end of CUFF.4747 on the reverse strand also appears to extend significantly beyond the point at which read alignments suggest it should end, again indicating involvement of reads from the opposite strand during transcript assembly with Cufflinks.

To overcome this issue, Cufflinks transcript assembly was repeated for strand specific alignment data. This was achieved by splitting the BAM into two separate files, each containing all the reads aligned to a single strand. Cufflinks was then used to process reads aligned to each strand separately, and was therefore unable to use reads aligned to the opposite strand when assembling transcripts. This method was shown to provide significant improvements in the accuracy of transcript structural prediction and goodness of fit to aligned RNA-seq reads (Fig. 6.2).

Utilising this modified Cufflinks assembly protocol, 16065 putative transcripts were assembled from RNA-seq data in *A. nidulans*.

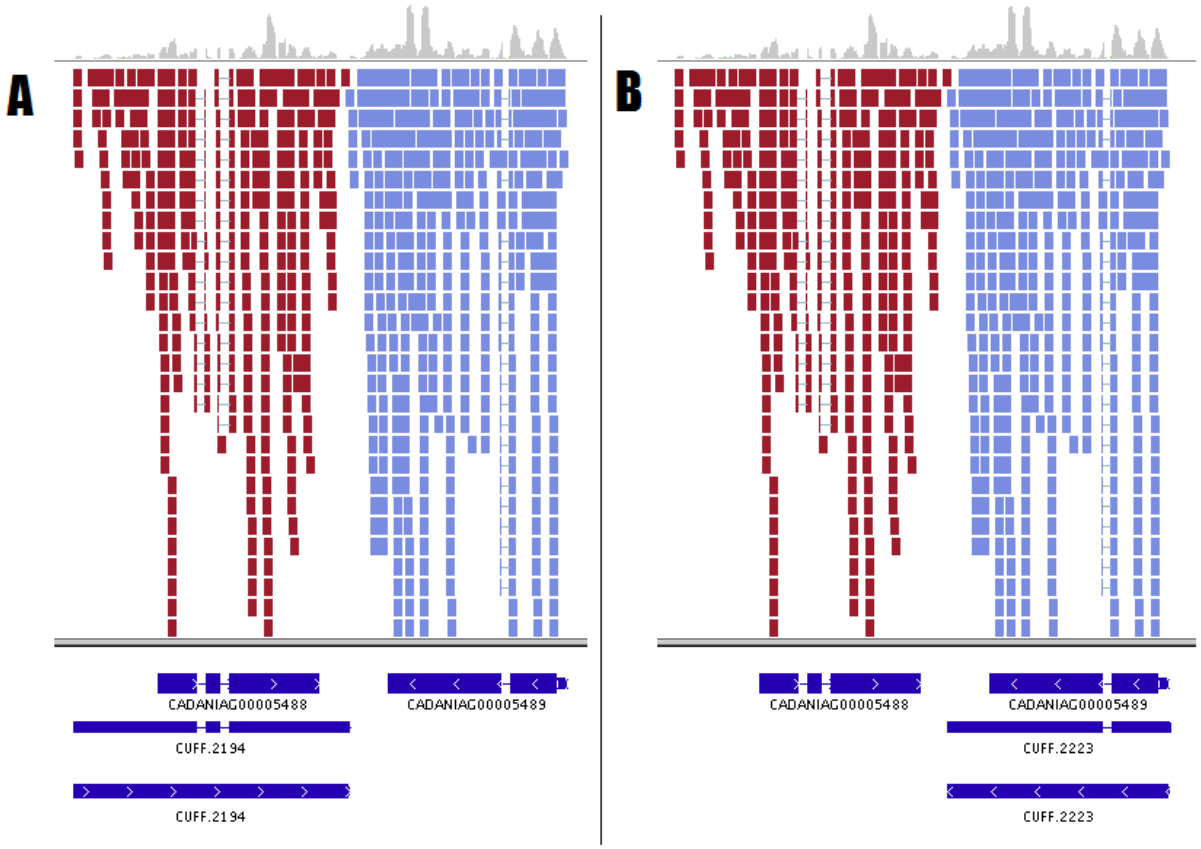


Figure 6.2. Strand specific transcript assembly with Cufflinks. Strand specific read alignments (red for forward strand, blue for reverse strand) is shown in IGV (top track), along with the existing gene model annotations for the genes in this region (CADANIAG00005488 and CADANIAG00005489). Transcripts assembled using Cufflinks and strand specific read alignments for both the forward (A) and reverse (B) strands are shown below existing gene models. Transcripts assembled in this way demonstrate excellent correlation with strand specific RNA-seq read alignments, with no involvement of reads aligned to the opposite strand.

6.4. Novel gene finding using Cufflinks assembled transcripts from SOLiD V4 RNA-seq libraries

Analysis of SOLiD version 2 whole transcriptome RNA-seq read alignment to intergenic regions in the *A. nidulans* genome had provided an estimation of the number of novel genes indicated by these data. Since this analysis, the release of the SOLiD version 4 had facilitated the creation and sequencing of newer whole transcriptome libraries with greater accuracy and higher coverage compared to those produced on the SOLiD version 2. New software had also been released for RNA-seq data analysis, allowing for the discovery of splice junctions (Tophat) and the prediction of transcript structure from aligned reads (Cufflinks). Investigation into the position of Cufflinks assembled transcripts provided a vast improvement in the robustness of novel gene prediction in *A. nidulans* compared to previous methods.

Given the large number of putative novel genes indicated by a single library of SOLiD version 2 RNA-seq data, the size of the dataset for this analysis was limited to the reads from a single SOLiD version 4 library. The library produced from RNA extracted from *A. nidulans* culture grown on minimal media supplemented with NO_3 was chosen as this was the standard growth condition, matching that of the previous analysis.

For this library, 16065 potential transcripts had previously been defined using Cufflinks. Comparison of these *de novo* transcripts with the CADRE2.5 *A. nidulans* gene model annotation in .gtf format was performed using the Cuffcompare feature of the Cufflinks software package (Trapnell *et al*., 2010; Roberts *et al*., 2011a; Roberts *et al*., 2011b).

Cuffcompare produced a tab delimited file in ".tmap" format, which listed the most closely matching reference transcript for each Cufflinks transcript. The class code column of this file contained information regarding the type of match for each Cufflinks transcript and the

reference gene with which it has been associated. Where a transcript could not be associated with a reference gene model, it was placed into the category of "unknown intergenic transcripts", indicated by the letter "u" in the class code column of the cuffcompare .tmap output file. Isolating the number of transcripts which fell into this category indicated that 1112 previously unannotated transcripts were expressed in RNA libraries extracted from *A. nidulans* grown in this condition.

6.5. Functional analysis of putative novel transcripts

To determine if any of these putative novel transcripts had the potential to represent previously unannotated protein coding genes, it was first necessary to analyse their sequences for the presence of open reading frames (ORFs). An open reading frame is a sequence of nucleotides which begins with a translation start codon and can be divided into a set of consecutive, non-overlapping triplets, each of which codes for an amino acid, and terminating in a stop codon. These features are required for the translation of the amino acid sequence into a viable protein. Long ORFs are therefore often used as an initial indicator of potential protein coding sequences, however they alone do not constitute sufficient proof that a sequence is translated (Deonier *et al*., 2005).

To obtain the sequence of each novel transcript, a list of intergenic transcript IDs was obtained from the cuffcompare.tmap file as previously described. Novel software was created to match these transcript IDs with those in the original Cufflinks output "transcripts.gtf", containing transcript positions (Chapter 2.5.14). The chromosome positions for each intergenic transcript were output to a new file, and a new GTF format file created containing only intergenic transcripts for visualisation in a genome browser. A second piece of software was developed to use the resulting transcript positions to extract the predicted sequence for

each transcript from the FASTA format CADRE2.5 *A. nidulans* genome annotation, and place them into a new FASTA format file to be used as input for ORF finding software (Chapter 2.5.15).

Open reading frame identification within intergenic transcript sequences was performed using Getorf from EMBOS (Williams, 2000). Minimum nucleotide length was set to 60 and finding of complimentary ORFs was disabled as transcripts on each strand had been aligned, assembled and were processed separately. Software was designed to process the resultant ORF list and obtain relevant statistics. This software counted the total number of unique ORFs, and interrogated the associated cufflinks transcript IDs to determine the number of transcripts in which these had been found (Chapter 2.5.16). 2593 open reading frames were identified across 868 unique transcripts.

To determine if any of these ORFs had the potential to produce functional proteins, the ORF sequences were aligned to known sequences coding functional protein domains. To provide a more robust analysis, two separate protein domain databases were used.

The first of these databases was Pfam 26.0, a large collection of protein families represented by multiple sequence alignments (Punta *et al*., 2012). Pfam protein families fall into two categories; Pfam-A and Pfam-B. Only Pfam-A entries were used, as this category contains only high quality, manually curated families. Pfam-B contains supplementary proteins from the Automatic Domain Decomposition Algorithm (ADDA) database **(**Heger & Holm, 2003), which are of significantly lower quality and as such are useful for detecting conserved regions, but not ideal for functional analysis of protein sequences. Initial attempts at using Pfam resulted in errors due to the presence of ORF sequences which Pfam determined were not biologically viable. These sequences were removed from the ORF input file to allow Pfam to run successfully. Additionally, Pfam would not accept lists of ORFs in the format

produced by ORF, due to the presence of duplicate hits and additional newline characters within the FASTA sequence. Novel software was developed to process the Getorf output and remove these features, allowing the resulting files to be used in Pfam (Chapter 2.5.17).

The second database used was the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG is a database resource for understanding high-level functions and utilities of biological systems, such as the cell, organism and ecosystem. In particular it provides a manually curated gene database against which sequences can be aligned (Kanehisa at al., 2004). KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by BLAST comparisons against the manually curated KEGG GENES database (Moriya *et al*., 2007).

290 protein domain hits were reported by pfam (Appendix 2.11.1), but only 51 by KAAS (Appendix 2.12.1). Comparison of the ORFs with confirmed protein domain matches in each system was performed using novel software (Chapter 2.5.18), to produce a list of ORFs which aligned to functional protein domains in both databases. This identified 37 transcribed ORFs which were predicted to produce functional proteins in both databases. These ORFs were located in 32 distinct intergenic transcripts, identifying these as strong candidates for putative novel protein coding genes, the strongest candidates being those transcripts which aligned to multiple protein domains in both systems. Full software outputs and novel transcript lists can be found in Appendix 2.13.1.

6.6. Identification of novel transcripts with non-AUG start codons

Using its default settings, Getorf identifies the start of an ORF by the presence of an AUG methionine start codon. However, it has long been established that translation can initiate from non-AUG start sites, with examples of this phenomenom reported in fungi as early as 1993 (Gutierrez *et al*., 1993). More recently, non-canonical start codons have been observed in *A. nidulans*, used by the GATA transcription factor gene areB (Conlon *et al*., 2001).

As current *A. nidulans* references were largely predicted *in silico*, it was likely that a number of protein coding transcripts with non-canonical translation start sites would have been missed. To identify putative novel protein coding genes from RNA-seq data, ORF finding with Getorf was repeated, using the "-[no]methionine" command to allow ORFs with non-methionine start codons.

841 ORFs were identified in 446 transcripts. These were aligned to the protein databases Pfam and KEGG as previously described. Pfam identified 719 hits against ORFs (Appendix 2.11.2), while KEGG identified only 65 (Appendix 2.12.2). Comparison of ORF hits between databases was performed with novel software as previously described. This identified 45 ORFs in 38 transcripts, 6 of which had not been identified by previous analysis using only ORFs with AUG start codons. The full software outputs and novel transcript lists can be found in Appendix 2.13.2.

6.7. Investigation of antisense transcripts using Cufflinks assembly

Natural antisense transcripts (NATs) are subsets of non-coding RNAs (ncRNA), defined as transcripts lacking evidence of a functional ORF (Ponting *et al*., 2009). The regulatory roles of NATs have recently been characterised in mammalian systems (Faghihi & Wahlestedt, 2009), and a large number have since been identified in range of fungal species including *A. flavus* (Smith et al., 2008). The discovery of natural antisense transcripts NATs in fungi greatly increased our understanding the regulation of gene expression through mechanisms such as transcriptional interference, chromatin remodelling and dsRNA formation. Transcriptome analysis has produced current estimates of NAT association with ORFs in fungi ranging from 0.8-3.6 % in systems with low coverage, to as high as 85.2 % in comprehensively analysed transcriptomes (Donaldson & Saville, 2012). Preliminary estimates of antisense occurrences using early RNA-seq data (Chapter 3.4; Chapter 5.6) placed this figure at 5-18 % in *A. nidulans.* More recent RNA-seq experiments on the SOLiD version 4 system (Chapter 5), coupled with vast improvements in transcript prediction through the use of Tophat and Cufflinks facilitated a more robust analysis of these putative antisense transcripts, which were identified using overlaps on opposite strands between transcripts assembled by Cufflinks.

Custom software created to facilitate both identification and characterisation of each antisense transcript (Chapter 2.5.19). Using this software, pairs of overlapping transcripts were identified by investigating their start and stop positions from strand specific Cufflinks assemblies. Pairs of overlapping transcripts were then placed into one of three categories, either with one transcript totally internal to the other (class 1), overlapping at the 3' ends (class 2) or overlapping at the 5' ends (class 3). A total of 2291 antisense occurrences were identified, meaning 28.5 % of transcripts demonstrated some form of antisense overlap. Of these occurrences, 697 were class 1, 1205 were class 2 and 287 were class 3. Class 1 showed

no significant enrichment for which strand contained the smaller transcript. However, there was significant bias for antisense transcripts overlapping at the 3' end, with more than half the observed antisense occurrences falling into this category. The full software output including lists of antisense pairs can be found in Appendix 2.14.

Amongst the complementary transcripts identified, some appeared to be associated with previously unannotated genes. To investigate the number of novel transcripts observed in antisense configuration, the transcripts were aligned to the CADRE2.5 *A. nidulans* gene model annotation using Cuffcompare as previously described. The class column of the resulting ".tmap" file indicated transcripts which did not match to any annotated gene model on the same strand, but which demonstrated an exonic overlap with a transcript on the opposite strand, therefore representing a novel transcript in antisense configuration with an existing gene model. These transcripts were indicated by the class code "x", and 55 examples were identified using this method.

In order to verify these data, four novel antisense transcripts were randomly chosen to be investigated experimentally. As the sequence data were produced utilising poly(A) selected RNA, RT-PCR was conducted using oligo(dT) in combination with gene specific 5' primers (2.1., Antisense A-D). Each primer was designed by visualising the aligned read data in the Artemis genome browser, and extracting the sequence data for regions of significant mapping frequency which overlapped an annotated gene model on the opposite strand, and that RNA-seq data indicated would produce a PCR product approximately 200 bp in length (Fig. 6.3). Sequences were used as input for the "Primer 3" DNA primer design software (Rozen & Skaletsky, 2000), with settings to design primers with an optimal length of 18 nt to achieve a Tm below 50 °C.

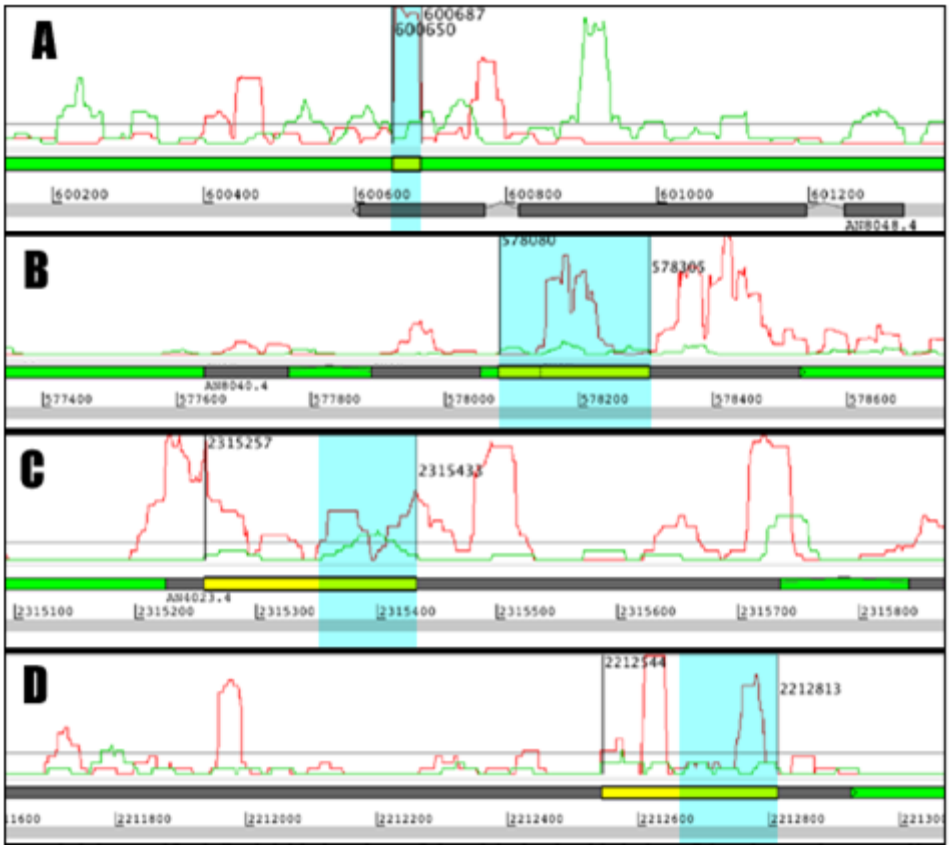


Figure 6.3. Novel antisense regions selected for experimental investigation. Whole transcriptome RNA-seq data for the forward (red) and reverse (green) strands is displayed against the forward (green track) and reverse (light grey track) strands of the *A. nidulans* genome annotation in the Artemis genome browser. Four putative novel antisense transcripts overlapping annotated *A. nidulans* gene models (shown in grey on the annotation tracks) are shown (A-D). Sequence data for the regions highlighted in blue was extracted to produce PCR primers to prove the existence of these transcripts experimentally. These regions were chosen as they appeared to have reads mapping to both strands, indicating antisense transcription.

Total RNA was extracted from wild type G00 *A. nidulans*. RNA underwent DNaseA treatment and cleanup by phenol:chloroform separation followed by ethanol precipitation to prevent DNA contamination of RT-PCR. First round RT-PCR was performed against oligo dT (25mer) using Superscript III reverse transcriptase (Invitrogen) according to the manufacturer's protocol. 30 rounds of PCR were performed with KOD hot start polymerase (Novagen) according to the manufacturer's protocols. PCR products were run on a 1.5 % agarose gel against a 50 bp ladder, stained with ethidium bromide and visualised under UV. PCR products were TA cloned (Chapter2.3.18) and sequenced. To ensure strand specificity, sequences were searched for the presence of a poly(A) tail at the end of the predicted fragment sequence. Based on these data, the presence of antisense transcripts was confirmed in all four cases, each involving a previously unannotated transcript in antisense configuration with a known gene.

One specific example of antisense RNA was investigated in more detail. The transcription factor MeaB is a gene has a regulatory associated with nitrogen metabolite repression (Polley, Caddick, 1996). From examination of RNA-seq data for the transcript sequence for *meaB,* it appeared that there is an antisense transcript which initiates within the first intron of the gene. Analysis of the intron revealed a GATA motif indicating a possible functional association with a second transcription factor AreA (Muro-Pastor *et al*., 1999). Recent work had shown *areA* to partially regulate the transcription of *meaB (*Wagner *et al*., 2010). To confirm the presence of the antisense transcript and test the possibility that it is under regulation by *areA* we utilised northern analysis using a single stranded probe from exon 1 (-ve strand). As can be seen from Fig. 6.4, the antisense transcript is differentially regulated in response to nitrogen regime and its expression is dependent on functional area.

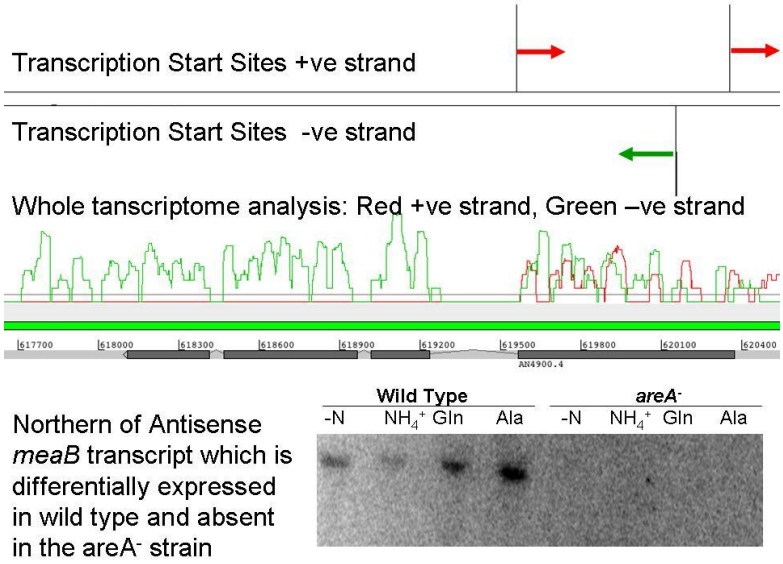


Figure 6.4. Confirmation of an antisense transcript in *meaB*. The position and direction of transcription start sites indicated by 5' specific RNA-seq are shown (top two tracks) alongside mapped whole transcriptome RNA-seq data for the forward (red graph) and reverse (green graph) strands in the region of meaB. These data are displayed here in Artemis alongside the A. nidulans genome annotation for this region (green track for forward strand, grey track for reverse). The presence of an antisense transcript in *meaB* was first indicated by the overlapping region of the strand specific coverage plots from these whole transcriptome RNA-seq data, and confirmed by the presence of an antisense TSS identified by 5' specific RNA-seq. From northern analysis of the antisense meaB transcript from wild type and areA- strains of *A. nidulans* (bottom right), this antisense transcript was identified and found to be modulated by nitrogen regime, and its expression to be AreA dependent.

For *meaB* the scarcity of the transcript meant that oligo dT was not optimal as a primer for reverse transcription. In this case we performed RT with the meaB sequence specific primer "meaB R" to minimise background interference during PCR. Two PCR reactions were performed with forward primers "meaB F1" and "meaB F2" in combination with reverse primer "meaB R" (Chapter 2.1). These primers were designed to produce fragments 400 bp and 800 bp in length. PCR products of the correct lengths were TA cloned and sequenced, confirming the presence of the antisense transcript. These results were validated by a control utilising no RT reaction, which did not give any product.

6.8. Discussion

Early attempts at novel transcript identification using SOLiD version 2 data were hindered by the lack of available software. Due to inexperience in software creation at this early time point, the software designed for this analysis was basic. Utilisation of a number of methods such as "masking" of regions beyond the ends of annotated gene will undoubtedly have led to a number of real novel transcripts being excluded from this analysis due to their proximity to existing gene models. Additionally, the use of a minimum transcript length was intended to avoid background mapping, but will have caused even further loss of real novel transcript identification, as many transcripts in *A. nidulans* are smaller than the 200 nt limit, with genes as small as 36 nt currently annotated by CADRE. Transcriptome analysis of *A. fumigatus* has also revealed a large population of non-coding RNAs, ranging from 21-22 nt long microRNAs (Bartel, 2004; Bartel & Chen, 2004) to lengths of around 500 nt (Jöchl *et al*., 2008). While this method of whole transcriptome RNA-seq will not have been able to detect the smallest of microRNAs, this suggests that a significant number of previously unidentified ncRNAs below 200 nt in length will have been excluded from this analysis in *A. nidulans*.

The development of Tophat and Cufflinks RNA-seq analysis software to work with SOLiD colour space data presented a drastic advance in the potential for novel transcript identification. Transcript assembly with Cufflinks utilises Tophat defined splice junctions to pull together reads aligned to exons from the same transcript, and was therefore far more robust than the novel gene finding software developed for this thesis. However, Cufflinks also presented its own challenges due to issues with strand specificity, assembling transcripts which included reads aligned to both strands of DNA. While Cufflinks has been shown to work well with unpaired RNA-Seq reads, it was designed with paired reads in mind (Trapnel, 2012), so SOLiD fragment libraries such as those used here were not ideal. Continued development of the software should result in higher accuracy transcript assembly using SOLiD reads in future releases. Until then, a valid workaround has been achieved by splitting the BAM file by strand, which vastly improved the accuracy of transcript assembly from strand specific fragment libraries.

Using this technique, 16065 putative transcripts have been identified from whole transcriptome RNA-seq data in *A. nidulans*. This presents a vast increase over the 10,827 transcripts currently predicated in the most recent CADRE *A. nidulans* genome annotation (version 3a). Of these transcripts, 1112 were predicted to originate from putative novel genes. ORF finding and functional domain analysis revealed that the vast majority of these transcripts are likely to be non-coding, explaining why they may not yet have been predicted *in silico*. An abundant class of intergenic ncRNAs which do not contain canonical ORFs has been described by Guttman *et al*., (2009, 2010), many of which are thought to produce functional RNAs with roles in transcriptional regulation and gene expression (Guttman *et al*., 2011).

32 strong candidates for protein coding genes were identified from ORFs in transcripts aligned to intergenic regions of the *A. nidulans* genome. The use of two protein domain databases to identify these genes will have inevitably led to a gross underestimation of these figures, and while those genes identified in both systems are the most likely to produce functional proteins, transcripts containing ORFs which aligned to only one protein database should not be discounted. Six additional strong candidates (matches in both systems) and a large number of additional weaker candidates (matches in 1 system) for novel protein coding genes were identified from ORFs with non-canonical start codons. Recent work by Ivanov *et al.,* (2010) has highlighted the potential importance of non-AUG start codons in the regulation of translational initiation, where only a small number had been identified previously.

2291 occurrences of transcripts in antisense configuration were also identified, with a significant bias observed for overlaps between the 3' ends of transcripts. This bias was consistent with previous findings in yeast (David *et al*., 2006) and *Aspergillus flavus* (Smith *et al*., 2008). While this represents the number of transcripts shown to overlap between the two strands of DNA, it does not necessarily represent the number of antisense transcripts with a functional relationship; as such relationships require experimental proof of existence. However, combining this with whole transcriptome and 5' specific RNA-seq data has been shown to provide an excellent resource for investigation into where such relationships may exist, and to guide experimental design into antisense regulation for specific genes (e.g. meaB).

Additionally, 55 strong candidates for putative novel antisense transcripts have been identified using Cuffcompare, 4 of which have been proven to exist experimentally. This number is likely to be a gross underestimation, as the numerous class codes and strict criteria

used by Cuffcompare to apply those codes, are likely to have excluded a large number of putative novel transcripts from being identified by the selection process used.

Chapter 7

7. Investigation of alternative splicing in *Aspergillus nidulans*

7.1. Alternative splicing

Alternative splicing of mRNA was first identified in 1980 when membrane-bound and secreted antibodies were shown to be encoded by the same gene (Alt *et al*., 1980; Early *at al*., 1980). This phenomenon has become increasingly studied as it allows alternative isoforms of proteins to be created from a single gene, potentially providing a huge increase in the diversity of proteins without the need for additional genetic information. While once thought of as a rare occurrence, the estimated number of genes which use alternative splicing of transcribed pre-mRNA to encode more than one protein (or protein isoform) has dramatically risen over time. Recent studies using high-throughput sequencing suggest that 95–100% of human pre-mRNAs with sequences corresponding to more than one exon can be differentially processed to yield multiple distinct mRNAs (Pan *et al*., 2008; Wang *et al*., 2008).

The majority of exon sequences are constitutively spliced, meaning that they are always included in the mature mRNA. Some exons however are differentially regulated and can be spliced in or out of the final mRNA. These are known as cassette exons, and may be regulated either at an individual level, or in a mutually exclusive mannar alongside additional cassette exons (Smith & Nadal-Ginard, 1989; Schmucker *et al*., 2000). Some exons have also shown to be variable in length due to alternative splice sites as a result of alternative promoter usage in the case of 5'-terminal exons, and alternative polyadenylation sites in the case of 3'-terminal exons. There are also cases where the removal of introns through splicing does not occur, leading to the intron sequence being retained in the final mRNA (Black, 2003).

Splice variants have been widely studied and four broad categories identified: retained introns (RIs), cassette exons (CEs), competing 5' splice sites, and competing 3' splice sites (Fig. 7.1). CEs have shown to be the predominant form of splice variation in multicellular eukaryotes (Sugnet *et al*., 2004; Ast, 2004; Thanaraj *et al*., 2004), while RIs are more prevelent in multicellular plants such as *A. thaliana* and *O. sativa* (Ner-Gaon *et al*., 2004; Campbell *et al*., 2006; Wang & Brendel, 2006), as well as in the fungi *Cryptococcus* and *A. nidulans*, and also in yeast (Kupfer *et al*., 2004; Loftus *et al*., 2005; Collins & Penny, 2006; Romfo *et al*., 2000).

Splicing of pre-mRNA is guided by short, conserved sequences known as splice sites, which are present at intron/exon junctions. The 5' splice site occurs at the exon/intron junction, and is a semi-conserved sequence containing a GU dinucleotide. Downstream of the 5' splice site at the intron/exon junction are three conserved elements which form the 3' splice site. These consist of the branch-point, a polypyrimidine tract, and finally a terminal AG at the 3' end of the intron (Burge *et al*., 1999). A macromollecular complex of five small nuclear ribonucleoproteins (snRNPs) is assembled onto the intron during splicing. These snRNPs are collectively known as the spliceosome, and are responsible for cleaving the pre-mRNA at the 5' splice site and lariat formation by the ligation of the free intron 5' end to the branch-point. The spliceosome then cleaves the pre-mRNA at the 3' splice site to release the intron (now in lariat configuration), and ligates the two exons to reform the spliced mRNA molecule.

Splice site consensus sequences alone are generally not sufficient to determine if a given sequence will be spliced, and regulation of splice site utilisation can occur in a number of ways. Firstly the choice of splice site is thought to be affected by the initial binding of spliceosome components to the pre-mRNA, and the formation of spliceosome complexes (Lallena *et al*., 2002). This can be influenced by a number of factors, including the strength of

splicing signals and the length of introns and exons. Components of the spliceosome have been shown to interact between splice sites at either end of an exon, stimulating the removal of flanking introns in a process called exon definition (Hoffman & Grabowski, 1992). However, suboptimal distances between splice sites have also been shown to interfere with splicing due to steric hinderance between machinery at flanking splice sites (Dominski & Kole, 1991). Many additional non-splice site regulatory sequences have also been shown to strongly affect splicing. Splicing enhancers are sequences which stimulate splisosome assembly, while splicing silencers or repressors act to block spliceosome assembly and prevent splicing events with which they are associated. Both splicing enhancers and repressors have intronic and exonic varieties, and act either through influencing the secondary structure of the pre-mRNA or as binding sites for proteins, both of which are able to influence splice site recognition (Libri *et al*., 1991; Jacquenet *et al*., 2001; Black, 2003).

Specific examples of alternative splicing have been described in Aspergilli (Trevisan, 2011; Maruyama, 2005), although such examples are rare. RNA-seq data has the potential to dramatically increase the number of alternative splicing events identified and greatly improve our understanding of alternative splicing in Aspergilli.

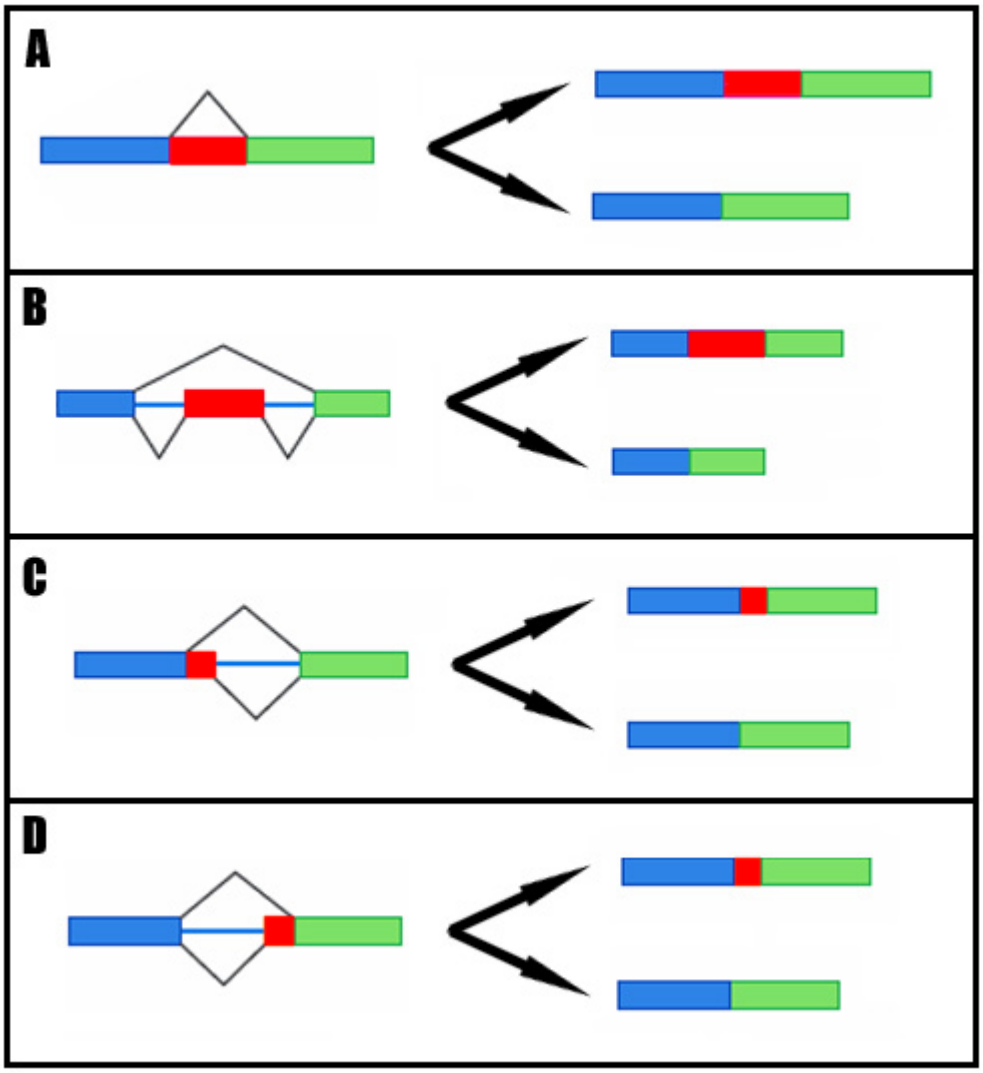


Figure 7.1. Types of alternative splicing. The four basic types of alternative splicing are illustrated. These consist of: Retained introns (**A**), cassette exons (**B**), competing 5' splice sites (**C**) and competing 3' splice sites (**D**). Pre-mRNAs are shown on the left, with the differentially spliced region highlighted in red, and black lines spanning regions which can be spliced out. The products of differential splicing are shown on the right. For retained introns (**A**), the red region is considered an intron if skipped and an exon if retained.

7.2. Analysis of alternative splicing events identified by whole transcriptome RNA-seq in *A. nidulans*

Visual analysis of whole transcriptome RNA-seq data and associated splice junctions identified by Tophat alignment was initially used to gauge the extent of alternative splicing in *A. nidulans*. A number of putative alternative splicing events were observed, and the most common form of alternative splicing appeared to be retention of introns. Potential alternative splicing events involving cassette exons and variable intron length due to competing 5' or 3' start sites were also observed, but with much lower frequency.

Further investigation into the extent of alternative splicing in *A. nidulans* utilised transcripts assembled by Cufflinks as previously described. Cufflinks identifies multiple isoforms of transcripts produced from the same loci, and outputs these with the same initial cufflinks identification number (ID), followed by a decimal point and a unique identification number for each transcript (e.g. CUFF01.1, CUFF01.2… etc). The number of loci across which transcripts had been assembled could therefore be found by searching the list of cufflinks transcript IDs for those that end in ".1". Of the 16065 total transcripts identified by Cufflinks, 15666 were shown to have unique IDs using this method. Subtracting this number from the total gave 399 transcripts which represent alternative isoforms. As some transcripts may have more than two isoforms, this number did not necessarily represent the number of alternatively spliced genes in *A. nidulans*. The search of cufflinks IDs was therefore repeated to identify the number of transcripts with IDs ending in ".2" as this would identify the number of loci at which these transcripts could be located. 374 distinct loci were identified by this search, representing the number of genes predicted by cufflinks to produce alternatively spliced transcripts.

Six genes indicated by Total RNA-sequencing to produce multiple transcript isoforms were chosen for further investigation (Figs. 7.2-7.7). Gene selection was by visual interrogation of

Cufflinks-assembled transcripts and included as many different alternative splicing methods as could be identified. As previously discussed, the major form of alternate splicing in *A. nidulans* is intron retention, and therefore this was the most significant and most closely examined, with transcripts from three of the six genes fall into this category. Individual examples of exon skipping and competing 5' splice site usage were also identified. Finally, an example of differential transcription start site usage resulting in transcripts with different first exons was investigated. Although this itself is not a form of alternative splicing, it is similar in that it allows multiple distinct transcripts to be produced from the same genetic locus, and the chosen example exhibits two very distinct 5' splice sites, both of which are associated with the same 3' splice site, meaning that alternative splicing does occur as a direct result of this modification in transcription start site.

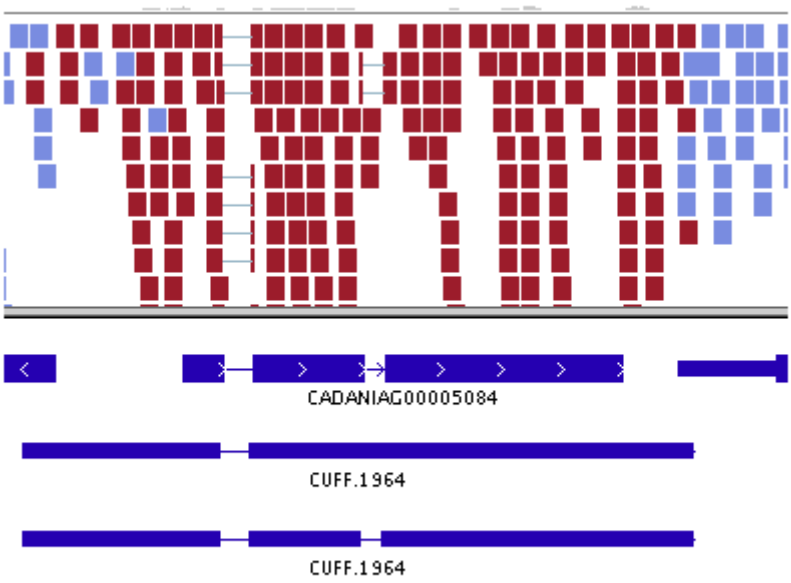


Figure 7.2. First example of alternative splicing by intron retention. The top track shows reads from whole-transcriptome RNA-seq mapped with Tophat to the CADRE2.5 *A. nidulans* genome. The second track shows the gene model annotation for this locus (CADANIAG00005084). The bottom two tracks show the modified transcripts predicted to originate from this gene. A single intron is predicted in the first transcript, while the lower transcript has two predicted introns. The reason for this can be seen from the RNA-seq reads, some of which have been split by Tophat and confirm the presence of this intron, while others appear to read through this region, suggesting that this sequence is retained in some transcripts.

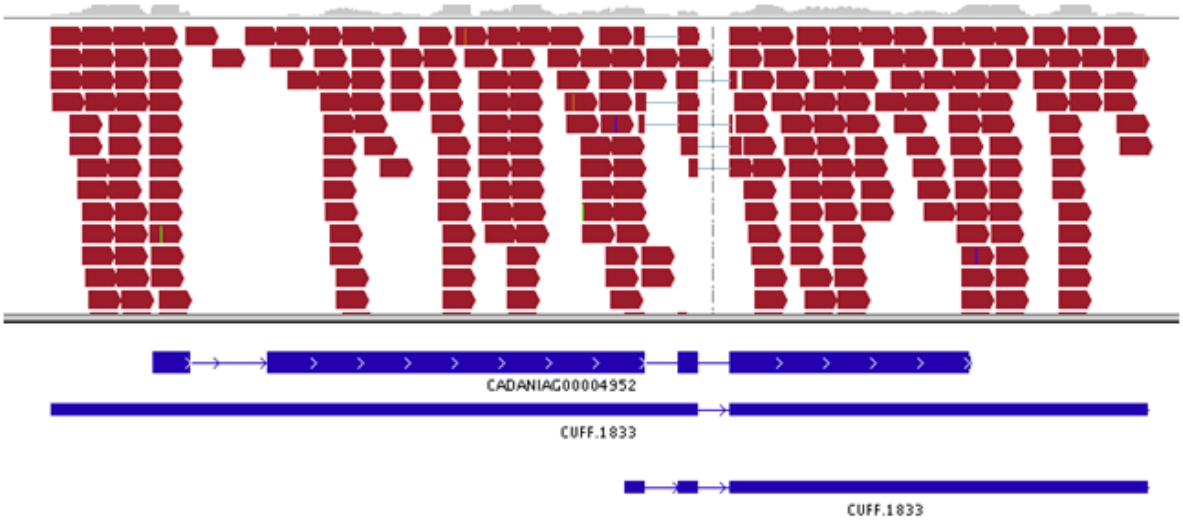


Figure 7.3. Second example of alternative splicing by intron retention. The top track shows reads from whole-transcriptome RNA-seq mapped with Tophat to the CADRE2.5 *A. nidulans* genome. The second track shows the gene model annotation for this locus (CADANIAG00004952). The bottom two tracks show transcripts assembled by Cufflinks. While the rightmost intron is predicted in both transcripts, the left intron is only found in the shorter transcript, suggesting alternative splicing by intron retention. It is also worth noting that the lower transcript is significantly shorter, suggesting that alternative 5' start sites are used in addition to alternative splicing to further modify the transcripts produced from this locus. This alternative start site may also facilitate the splicing out of the sequence in this region.

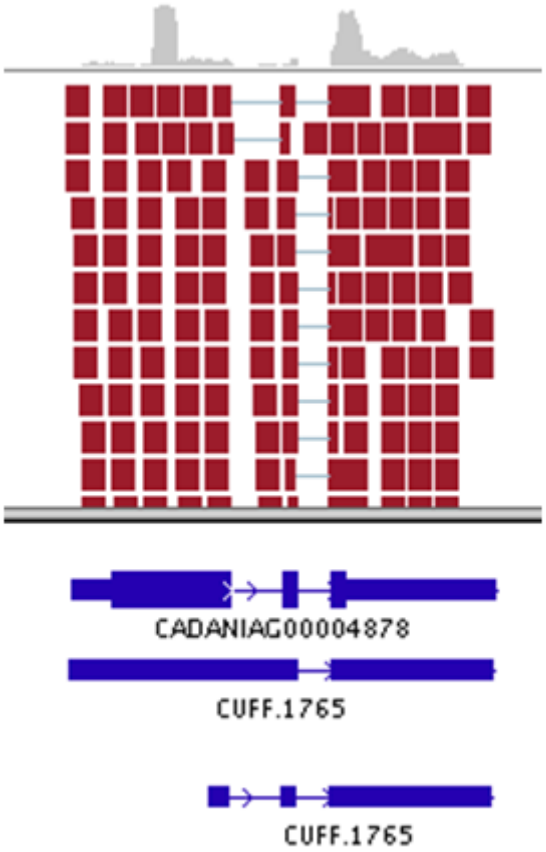


Figure 7.4. Third example of alternative splicing by intron retention. The top track shows reads from whole-transcriptome RNA-seq mapped with Tophat to the CADRE2.5 *A. nidulans* genome. The second track shows the gene model annotation for this locus (CADANIAG00004878). The bottom two tracks show transcripts assembled by Cufflinks. While the lower transcript appears to have 2 introns, only the second intron is present in the upper transcript. It is also worth noting that the lower transcript is significantly shorter, suggesting that alternative 5' start sites are used in addition to alternative splicing to further modify the transcripts produced from this locus. This alternative start site may also facilitate the splicing out of the sequence in this region.

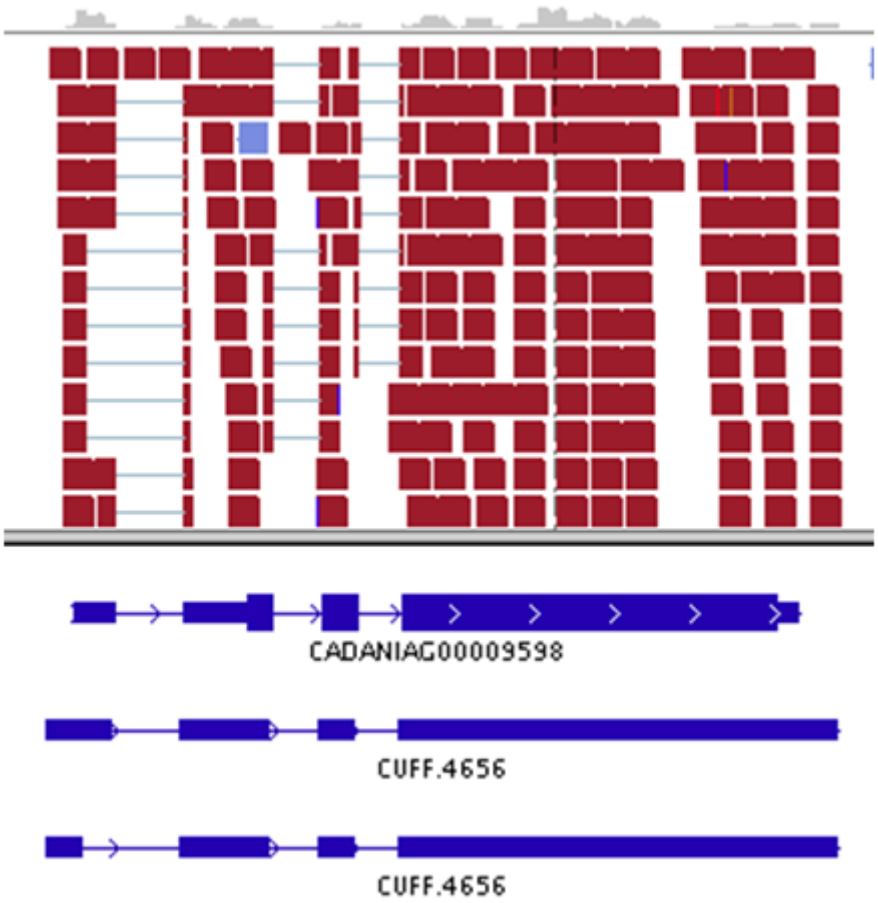


Figure 7.5. Example of alternative splicing utilising competing 5' splice sites. The top track shows reads from whole-transcriptome RNA-seq mapped with Tophat to the CADRE2.5 *A. nidulans* genome. The second track shows the gene model annotation for this locus (CADANIAG00009598). The bottom two tracks show transcripts assembled by Cufflinks. While these two transcripts are nearly identical, the first intron is shown to originate at two different 5' splice sites, resulting in a shorter first exon in the lower transcript.

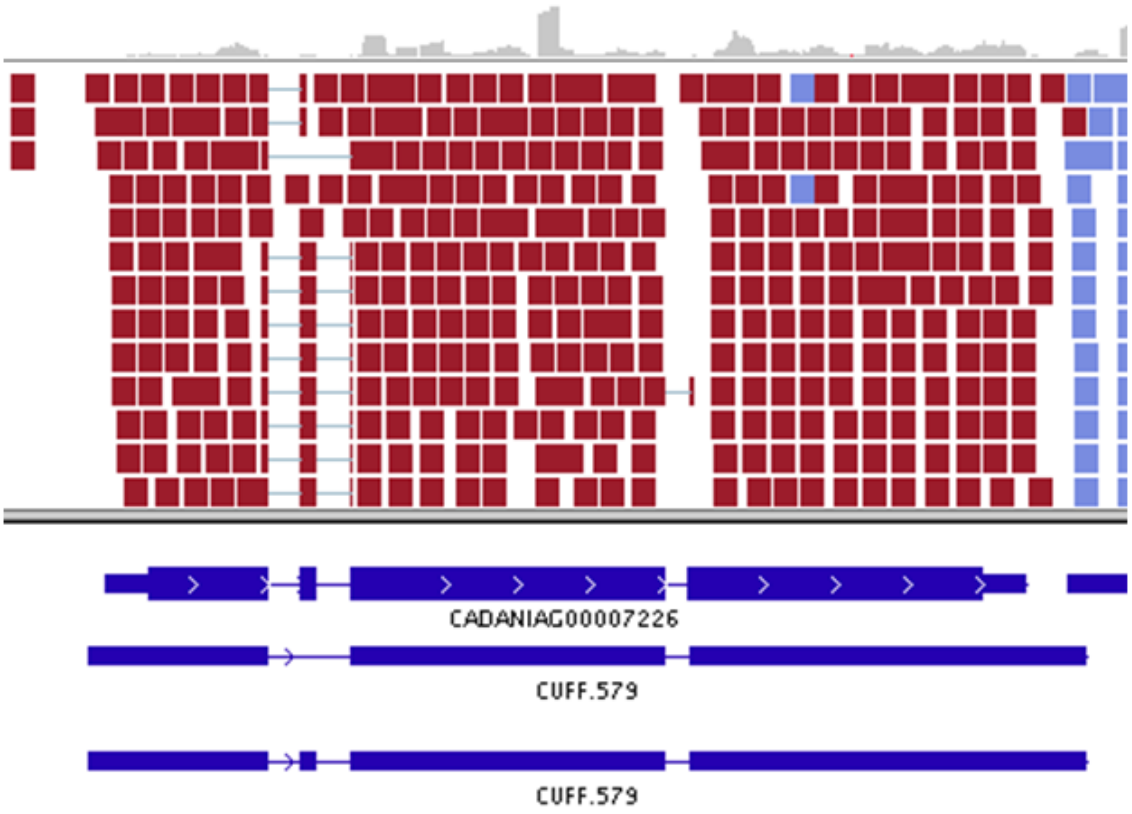


Figure 7.6. Example of alternative splicing utilising a cassette exon. The top track shows reads from whole-transcriptome RNA-seq mapped with Tophat to the CADRE2.5 *A. nidulans* genome. The second track shows the gene model annotation for this locus (CADANIAG00007226). The bottom two tracks show transcripts assembled by Cufflinks. The small second exon predicted in the gene annotation is present in only one of the two predicted transcripts, with RNA-seq reads indicating that in some instances the cassette exon is spliced out.

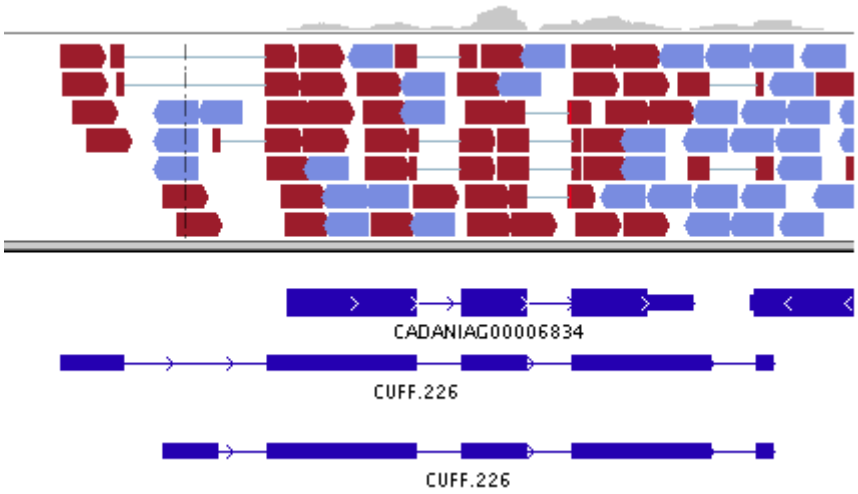


Figure 7.7. Example of alternative 5' start site utilisation resulting in alternative transcripts from a single locus. The top track shows reads from whole-transcriptome RNA-seq mapped with Tophat to the CADRE2.5 *A. nidulans* genome. The second track shows the gene model annotation for this locus (CADANIAG00006834). The bottom two tracks show transcripts assembled by Cufflinks. RNA-seq data indicates that this transcript can originate from two alternative 5' transcription start sites, each producing a different exon at the start of the transcript, which is then spliced together with a downstream sequence which remains constant. While this is not considered a method of alternative splicing, it is another way in which variation in transcripts can be achieved from a single locus. The use of alternative 5' start sites which are in a different frame may also contribute to alternative splicing events downstream by bringing motifs responsible for determining splice site location intro frame.

Primers were designed by first using IGV to visualise the regions where alternative splicing had been identified by RNA-seq data. Sequences corresponding to the regions of interest were extracted from the appropriate CADRE2.5 FASTA reference file and checked against the aligned sequence in IGV to ensure the correct sequence had been obtained. Sequence features such as TSS and splice sites were marked to facilitate manual design of PCR oligos. Individual forward and reverse PCR oligos were designed with a length of 18 nucleotides and GC content of 10 (56 %), to provide an approximate melting temperature of >50 °C and therefore an appropriate annealing temperature for successful PCR. The position of each oligo was also carefully chosen, as sequences being spliced out in each example were relatively small (35-100 bp), meaning that the two PCR products produced in each case would be similar in size. Oligos were therefore designed to produce small PCR fragments (approximately 150-200 bp in size), which allowed greater resolution of the products when analysis by electrophoresis. The positions of these oligonucleotides are shown in Fig. 7.8 and their sequences listed in Appendix 2.1. (Splicing A-E).

To confirm these events experimentally, PCR oligos were designed to capture the region in which alternative splicing occurs. If transcripts of different lengths are produced due to differential splicing in these regions, this would result in multiple PCR product lengths. The RNA-seq data further facilitates this experiment as it precisely defines splice junctions. This allows for accurate prediction of PCR products resulting from each splicing event and thus facilitates rapid validation of alternative splicing.

To maximise the chance of producing both splice variants, RT PCR was performed with RNA grown on both nutrient rich and highly nutrient deficient media. Total RNA was extracted from wild type G00 *A. nidulans* grown on both minimal medium supplemented with NO_3^- and on a severely nutrient limited medium (oat) as an extreme alternative condition where metabolic stress should promote alternative splicing of transcripts (see Chapter 2.2.4

for detailed growth conditions). All RNA underwent DNase I treatment and cleanup by phenol/chloroform precipitation to prevent DNA contamination of RT-PCR. To produce full length cDNA for PCR, first round RT-PCR was performed using oligo dT (25mer) and Superscript III reverse transcriptase (Invitrogen). The appropriate annealing temperature for each set of primers was determined by finding the precise melting temperature of each oligo using the online software "OligoCalc" (Kibbe, 2007). Subsequent PCR reactions were performed using specifically designed primers as previously described. PCR products were visualised under UV next to a 50 base pair ladder on 2.2 % agarose gel stained with ethidium bromide.

PCR of cDNA with oligos designed to isolate the example of a cassette exon produced no product in the expected size ranges. This PCR was repeated using a range of annealing temperatures to ensure that the failure of the reaction was not due to experimental error, however no product was observed for any of the repeats with either RNA sample. For all other samples, two distinct PCR products were identified in the expected size ranges, thus confirming the presence of each of these transcripts in *A. nidulans* cells grown in both conditions (Fig.7.8).

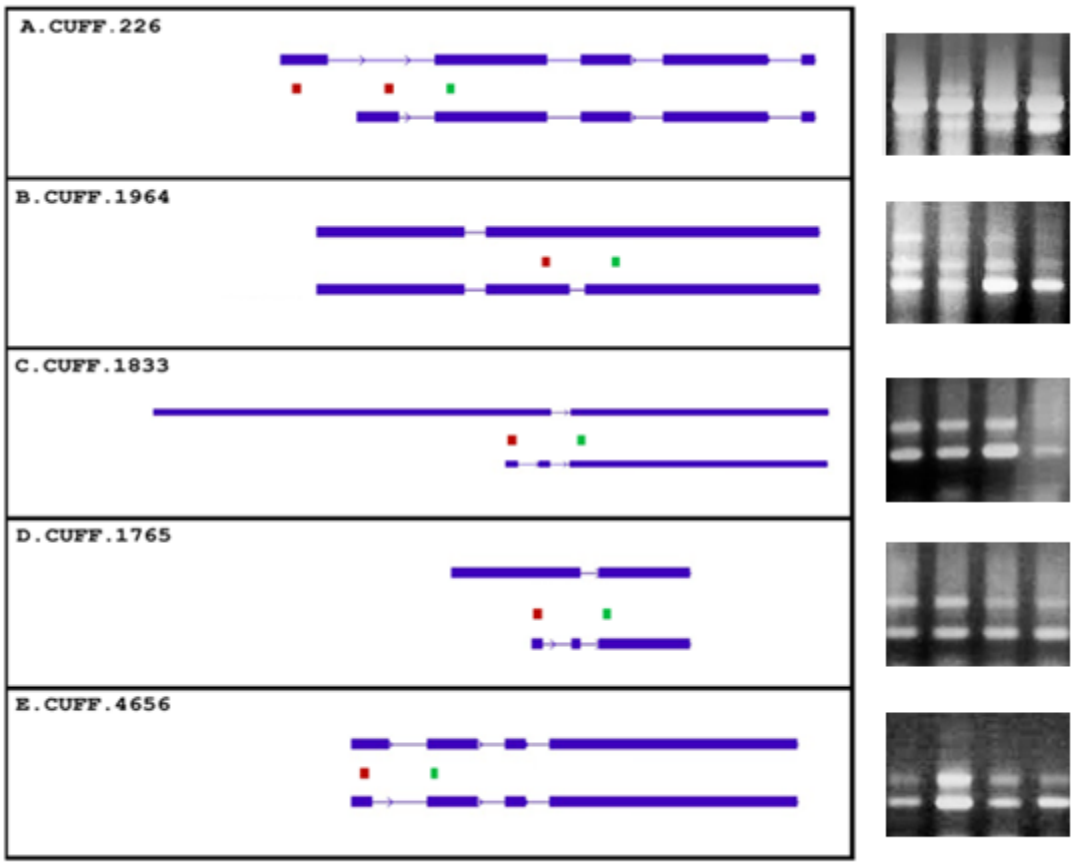


Figure 7.8. PCR oligo design and products for investigating 5 sets of differentially spliced transcripts. Forward (red) and reverse (green) PCR oligos are shown aligned to their positions on 2 differentially spliced transcripts assembled from total RNA-seq data. These transcripts indicate multiple forms of alternative splicing. In the case of A, 2 different transcript start sites result in the first exon being different depending on the point of transcriptional initiation. B, C and D show differential inclusion/exclusion of an intron, while E demonstrates two different intron start positions resulting in two distinct lengths of the first exon. For A, we utilised three primers, with a forward primer in each of the two starting exons and a reverse primer in the second exon. Using the RNA-seq data the length of each splice junction was found and the two forward primers positioned to give significantly different sized products when used in combination with the same reverse primer. For each of the other examples, primers were designed to fall across the splice site in question, resulting

in different length products dependant on the length or exclusion/inclusion of splice junction. Gels for each PCR are shown alongside their respective transcripts, with the two left lanes showing the products of PCR reactions using RNA from culture grown on minimal media $+NO_3^-$, and the right two lanes showing the products of PCR reactions with RNA from culture grown on oat. Two different size bands were produced in both conditions for all 5 genes, the sizes consistent with alternative splicing events predicted by RNA-seq and Cufflinks assembly. This indicates that two different lengths of transcripts are being produced as a result of alternative splicing in these regions.

7.3. Discussion

Alternative splicing in *A. nidulans* has been investigated through whole transcriptome RNA-seq and transcripts predicted using the Cufflinks transcript assembly software. 399 splice variants have been identified across 374 unique loci. This represents approximately 3.6% of known protein coding genes in *A. nidulans* being alternatively spliced. While this figure is extremely low compared to humans where up to 100 % of genes are expected to produce multiple transcript isoforms (Pan *et al*., 2008; Wang *et al*., 2008), it is a significant advance on previous studies in *A. nidulans*, which predicted only 100 splice variants from publicly available ESTs. Of those previously predicted alternative splicing events, the 74% were shown to fall into the category of intron retention, while 25% utilised alternative 5' or 3' splice sites and only 1% involved cassette exons **(**McGuire *et al*., 2008). Visual analysis of the alternative splicing events predicted by Cufflinks assembly of Tophat-aligned whole transcriptome RNA-seq reads indicated that high-throughput sequencing data supports these previous findings, although the proportions of each category were not recorded.

Of the 399 alternative transcripts identified in *A. nidulans*, six were investigated by PCR using custom designed oligos to produce cDNA fragments of different lengths based on the differences in transcript structure. The cassette exon event which was investigated did not produce any PCR products in the expected size ranges, indicating that this alternative splicing event was either wrongly predicted by Cufflinks, or that it does not occur in either of the conditions used. Examples of retained introns, alternative splice sites and alternative transcription start sites were all proven experimentally. While alternative transcription start site usage is not technically classified as alternative splicing, it is interesting to note that two of the three examples of alternative splicing by retained introns were also shown to also demonstrate alternative transcription start sites. In each case, the intron which was alternatively spliced was the one closest to the 5' end. This suggests that alternative transcription start sites not only produce different lengths of transcripts, but can have an effect on splice sites. Conserved sequence features at each end of splice sites were identified shortly after the discovery of split genes in 1977 (Breathnach, 1978), therefore the effect of alternative transcription start site position on intron retention is most likely due to a frame shift, bringing otherwise unused splicing motifs into frame. This is supported by the example in Fig. 7.7, which shows two very distinct exons produced by alternative transcription start sites, each with its own distinctive 5' splice site, both of which are associated with the same 3' splice site.

The previously described analysis was performed using whole transcriptome RNA-seq data from a single condition. Utilising RNA-seq data produced with *A. nidulans* RNA from multiple conditions where alternative splicing of transcripts may be up-regulated, and comparing transcripts assembled between each condition would provide a more comprehensive list of transcripts which can be alternatively spliced.

Although a definite bias for intron retention was observed as the main method of alternative splicing in *A. nidulans,* the exact number of transcripts in each class was not recorded. It is also worth noting that some examples of alternative transcripts may not fall into these categories, being instead produced by modifications at the 5' or 3' end due to alternative start sites or non-splicing post translational modifications. Classification of identified alternative transcripts into the four previously defined groups would therefore provide a more accurate representation of the number of alternatively spliced transcripts. This analysis would also facilitate a direct comparison of the frequency of occurrence and proportions of each group with those reported by previous works in *A. nidulans*.

Chapter 8

8. Further work

8.1. Availability of data as a community resource

Collaboration with AspGD has made RNA-seq data produced for this thesis available to members of the *Aspergillus nidulans* community. These data are currently being used to assist reannotation of the *A. nidulans* genome.

Through collaboration with CADRE, RNA-seq data produced from *A. nidulans* grown on minimal media with nitrate as sole nitrogen source has been uploaded to the CADRE website (http://www.cadre-genomes.org.uk) and can be viewed through the Ensembl genome browser. To display the RNA-seq data, select "*Aspergillus nidulans"* from the CADRE homepage, then use the "Sample entry points" menu to navigate to the genome browser. The RNA-seq data can then be added to the display by clicking "Configure this page", then selecting "Caddick, RNA-seq, NO_3" from the Functional genomics submenu.

The availability of these data in this format provides a high-quality visual resource to aid the interrogation of gene structure and transcription in *A. nidulans*. It is hoped that these data will be used by other members of the *A. nidulans* research community to aid their research and further our understanding of the *A. nidulans* transcriptome.

Further collaboration with CADRE is being undertaken to make all the RNA-seq data produced during this thesis available online. While aligned read data can be easily presented in a visual format, many members of the *A. nidulans* research community have expressed interest in utilising the raw read data from SOLiD sequencing. The sheer size of RNA-seq data files has proven to be a prohibitive factor in making these data publicly available. Therefore, while collaboration with CADRE is important to make RNA-seq data available for online visualisation as quickly as possible, the continued production of RNA-seq data and

desire to make these data fully available may mean that a dedicated website will be required in the long term.

8.2. Improving the analysis of novel protein coding genes

This thesis has identified a large number of putative novel protein coding genes from whole transcriptome RNA-seq. These could be further investigated using the results of work that is currently being undertaken to perform ribosomal profiling in *A. nidulans*. The strategy of ribosomal profiling is based on deep sequencing of ribosome-protected mRNA fragments, and enables genome-wide investigation of translated mRNA sequences with subcodon resolution (Ingolia *et al*., 2009; Ingolia, 2010). This approach would therefore provide direct evidence for which transcripts are translated and the region of the transcript that is coding. This would particularly useful for the confirmation of putative protein coding genes with non-canonical translation start sites. Ingolia *et al*., (2011) have shown that while a substantial number of intergenic ncRNAs in mammalian systems did not engage ribosomes, many putative intergenic ncRNAs contained short, successive segments that were translated at similar rates to classical protein coding sequences. The coding nature of these RNAs would generally not be detected by traditional methods which rely on *in silico* prediction. This work therefore has the potential to greatly expand our understanding of the *A. nidulans* proteome and the functional annotation of putative genes and transcripts (Weiss & Atkins, 2011).

8.3. Improving the analysis of non-coding and antisense transcripts

Examples of ncRNAs including NATs with and without poly(A) tails have been observed in other fungi, suggesting functional classes with vary in RNA structure, stability and/or

subcellular localisation (Donaldson & Saville, 2012). Identification of these RNAs in *A. nidulans* could therefore be aided by the production of RNA-seq libraries using ribosomal depletion to enrich for mRNAs, rather than the poly(A) selection used for the libraries in this thesis. These libraries have already been produced, as well as additional libraries from RNA extracted from *A. nidulans* grown in a range of different conditions. Future transcriptome analysis will combine RNA-seq data from these libraries to provide the widest possible expression profile, producing a far greater depth of coverage and a more robust identification of ncRNAs and antisense transcripts on a global scale.

8.4. *Aspergillus nidulans* RNA-seq in the future

Aspergillus nidulans RNA-seq for this thesis was conducted on the ABI SOLiD platform. This presented numerous bioinformatic issues, primarily due to the colour space format of SOLiD data which was not compatible with many of the commonly used systems for RNA-seq analysis. Combined with the poor reliability of the SOLiD system reported by users at the Liverpool Centre for Genomic Research, lack of support from ABI, and a general preference for Illumina sequencing within the *Aspergillus* community (e.g. Wang *et al*., 2010; Gibbons *et al*., 2012), this has resulted in a shift away from the SOLiD system. It is important to stress that this is currently a very dynamic field of study. Technology that is constantly evolving and improving, with each system having its own advantages and disadvantages, and techniques such as 5' specific RNA-seq continually being developed to provide greater accuracy and explore specific features. I therefore propose that future RNA-seq studies in *A. nidulans* may benefit from a combined strategy using data from alternative NGS systems and techniques to supplement current data.

References

Abler, M.L. & Green, P.J. (1996) Control of mRNA stability in higher plants. *Plant Mol. Biol*., 32, pp. 63-78.

Alt, F. W., Bothwell, A. L. M., Knapp, M., Siden, E., Mather, E., Koshland, M. & Baltimore, D. (1980) Synthesis of secreted and membrane-bound immunoglobulin μ heavy chains is directed by mRNAs that differ at their 3' ends. *Cell*, 20, pp. 293–301.

Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E. & Mattick, J.S. (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucl. Acids Res.*, 39, D146–D151.

Applied Biosystems (2008) SOLiD System Accuracy. Application Fact Sheet – SOLiD System. Available online at: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057511.pdf [Accessed 3/08/2012]

Arnaud, M. B., Chibucos, M. C., Costanzo, M. C., Crabtree, J., Inglis, D. O., Lotia, A., Orvis, J., Shah, P., Skrzypek, M. S., Binkley, G., Miyasato, S. R., Wortman, J. R. & Sherlock, G. (2010) The *Aspergillus* Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the *Aspergillus* research community. *Nucl. Acids Res.*, 38(Database issue), D420-427.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J.T., Harris, M.A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25, pp. 25–29.

Ast, G. (2004) How did alternative splicing evolve? *Nat. Rev. Genet*., 5, pp. 773-782.

Arst, H. N., Jr. & Cove, D. J. (1973) Nitrogen metabolite repression in *Aspergillus nidulans*. *Mol. Gen. Genet.*, 126**, pp.** 111-41.

Bailey, T. L., Bodén, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucl. Acids Research*, 37, W202-W208.

Bailey T. L. & Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2, pp. 28-36.

Ballario, P., Vittorioso, P., Magrelli, A., Talora, C., Cabibbo, A. & Macino, G. (1996) White collar-1, a central regulator of blue light responses in *Neurospora*, is a zinc finger protein. *EMBO J*., 15, pp. 1650-1657.

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

Bartel,D.P. & Chen,C.Z. (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat. Rev. Genet.*, 5, pp. 396–400.

Beiter, T., Reich, E., Williams, R.W. & Simon, P. (2009) Antisense transcription: a critical look in both directions. *Cell Mol. Life. Sci.,* 66, pp. 94–112.

Belasco, J. & Brewerman, G. (1993) Control of Messenger RNA Stability. New York: Academic Press.

Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E. & Plasterk, R. H. A.(2006). Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet*., 38, pp. 1375–1377.

Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein E. O. & Schreiber S. L. (2004) Global nucleosome occupancy in yeast. *Genome Biology,* 5.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tingprasit, W., Samanta, M., Weissman, S., Gerstein, M. & Snyder, M. (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science,* 306, pp. 2242–2246.

Birney, E., Clamp, M. & Durbin, R., (2004) GeneWise and Genomewise. *Genome Research*, 14, pp. 988–995.

Black, D.L. (2000) Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell,* 103, pp. 367–370.

Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.,* 72, pp. 291-336.

Blanchette, M. & Sinha, S. (2001) Separating real motifs from their artifacts. *Bioinformatics*, 17, S30-8.

Blaszczyk, J., Tropea, J. E., Bubunenko, M., Routzahn, K. M., Waugh, D. S., Court, D. L. & Ji, X. (2001) Crystallographic and Modeling Studies of RNase III Suggest a Mechanism for Double-Stranded RNA Cleavage. *Structure*, 9, pp. 1225–1236.

Boeger, H., Grisenbeck, J., Strattan J. S. & Kornberg R. D. (2003) Nucleosomes unfold completely at a transcriptionally active promoter. *Molecular Cell,* 11, pp. 1587-1598.

Borsuk, P. A., Nagięć, M. M., Stępień, P. P. & Bartnik, E. (1982) Organization of the ribosomal RNA gene cluster in *Aspergillus nidulans. Gene*, 17(2), pp. 147-52.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc. Natl. Acad. Sci. USA*, 75, pp. 4853–4857.

Bregman, A., Avraham-Kelbert, M., Barkai, O., Duek, Lea, Guterman, A. & Choder, M. (2011) Promoter elements regulate cytoplasmic mRNA decay. *Cell*, 147, pp.1473-1483.

Breslow, R. & Huang, D.-L. (1991) Effects of metal ions, including Mg^{2+} and lanthanides, on the cleavage of ribonucleotides and RNA model compounds. *Proc. Natl. Acad. Sci. USA* 88, pp. 4080–4083.

Browne, K. A. (2002) Metal ion-catalyzed nucleic acid alkylation and fragmentation. *J. Am. Chem. Soc.*,124(27), pp. 7950–7962.

Bruce, S. J., Gardiner, B. B., Burke, L. J., Gongora, M., Grimmond, S. M. & Perkins, A. C. (2007) Dynamic transcription programs during ES cell differentiation towards mesoderm in serum versus serum-free (BMP4) culture. *BMC Genomics*, 8, p. 365.

Burge, C. B., Tuschl, T. & Sharp, P. A. (1999) Splicing of Precursors to mRNAs by the Spliceosomes. In The RNAworld Second edition (ed. Gesteland *et al*.). *Cold Spring Harbor Laboratory Press*, Cold Spring Harbor, New York., pp. 525–560.

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J., Windsor, D., Sazanov, A., Fries, R. & Waddington, D. (1999) The dynamics of chromosome evolution in birds and mammals. *Nature*, 402, pp. 411-413.

Butler, J. E. & Kadonaga, J. T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev*., 16, pp. 2583–2592.

Butler, J. S. (2002) The yin and yang of the exosome. *Trends Cell Biol.* 12, pp. 90-96.

Caddick, M. X., Peters, D. & Platt, A. (1994) Nitrogen regulation in fungi. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology*., 65, pp. 169-177.

Cameron, V. & Uhlenbeck, O. C. (1977) 3'-Phosphatase activity in T4 polynucleotide kinase. *Biochemistry.* 16(23), pp. 5120-6.

Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. & Buell, C. R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics*, 7, pp. 327.

Caponigro, G. & Parker, R. (1995) Multiple functions for the poly (A)-binding protein in mRNA decapping and deadenylation in yeast. *Genes Dev*., 9, pp. 2421-2432.

Caponigro, G. & Parker, R. (1996) Mechanisms and control of mRNA turnover in *Saccharomyces cerevisiae. Microbiol. Rev*., 60, pp. 233-249

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple,

B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, T., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H. Kawai, J. & Hayashizaki, Y. (2005) The transcriptional landscape of the mammalian genome. *Science*, 309, pp. 1559-1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A. & Hayashizaki, Y. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet*., 38, pp. 626-635.

Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. (2012) Artemis; an integrated platform for visualisation and analysis of high-throughput sequence-based experimental data. *Bioinformatics,* 28(4), pp. 464-469.

Cereghino, G.P. & Schaffler, I.E. (1996) Genetic analysis of glucose regulation in *Saccharomyces cerevisiae*: control of transcription versus mRNA turnover. *EMBO J*., 15, pp. 363-374.

Chae, K. S., Kim, J. H., Choi, Y., Han, D. M. & Jahng, k. Y. (1995) Isolation and characterization of a genomic DNA fragment complementing an *nsdD* mutation of *Aspergillus nidulans. Mol. Cells*. 5, pp. 146-150.

Chambeyron, S. & Bickmore, S. C. (2004) Chromatin decondesnation and nuclear reorganisation of the HoxB locus upon induction of transcription. *Genes Dev.* 18, pp. 1119-1130.

Chamilos, G. & Kontoyiannis, D. P. (2006) Update on the antifungal drug resistance mechanisms of *Aspergillus fumigatus*. *Drug Resistance Updates*, 8, pp. 344–358.

Chen, C. & Duan, X. (2011) Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol. Biol.*, 733, pp. 93-103.

Chen, D. & Patton, J. T. (2001) Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5′-RACE and primer extension. *Biotechniques,* 30, pp. 574–582.

Chen, D., Toone, W. M., Mata, J., Lyne, R., Burns, G., Kivinen, K., Brazma, A., Jones, N. & Bähler, J. (2003) Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell*, 14, pp. 214–229.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S. & Gingeras T. R. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308, pp. 1149–1154.

Chomczynski, P. & Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.*, 162, pp. 156–159.

Clark, M.B. & Mattick, J.S. (2011) Long noncoding RNAs in cell biology. *Semin. Cell Dev. Biol.*, 22, pp. 366–376.

Cloonan, N., Forrest, A. R R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J. & Grimmond, S. M. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5, pp. 613-619.

Clutterbuck, A. J. (1974) *Aspergillus nidulans*, vol. 1. *Plenum Press*, New York, NY.

Cocquet, J., Chong, A., Zhang, G. & Veitia, R. A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, 88, pp. 127–131.

Coghalan, A. & Wolfe, K. H. (2002) Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Res.*, 12, pp. 857-867.

Coleman, J. E. (1992) Structure and mechanism of alkaline phosphatase. *Annu. Rev. Biophys. Biomol. Struct.,* 21, pp. 441–483.

Collins, L. & Penny, D. (2006) Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Investigating the intron recognition mechanism in eukaryotes. *Mol. Biol. Evol*., 23, pp. 901-910.

Conlon, H., Zadra, I., Haas, H., Arst, H.N. Jr, Jones, M.G. & Caddick, M.X. (2001) The *Aspergillus nidulans* GATA transcription factor gene areB encodes at least three proteins and features three classes of mutation. *Mol. Microbiol.,* 402(2), pp. 361-375.

Couttet, P., M. Fromont-Racine, D. Steel, R. Pictet, & T. Grange. (1997) Messenger RNA deadenylylation precedes decapping in mammalian cells. *Proc. Natl. Acad. Sci. USA,* 94, pp. 5628-5633.

Cove, D. J. (1966) Induction and repression of nitrate reductase in fungus *Aspergillus nidulans. Biochim. Biophys. Acta.,* 113, pp. 51-56.

Cowley, G. (2012) "Bioinformatic and functional characterization of promoter motifs in *Aspergillus nidulans"*. MSc Thesis, University of Liverpool.

Crawford., J. E., Guelbeogo, W. M., Sanou, A., Traore, A., Vernick, K. D., Sagnon, N. & Lazzaro, B. P. (2010) *De novo* transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. *PLOS one*, 5.

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S.E. (2004). WebLogo: A sequence logo generator. *Genome Res,*, 14, pp. 1188-1190.

Dang, Y., Yang, Q., Xue, Z. & Liu, Y. (2011) RNA interference in fungi: pathways, functions, and applications. *Eukaryot. Cell,* 10, pp. 1148–1155.

David, H., Ozcelik, I.S., Hofmann, G. & Nielsen, J. (2008) Analysis of *Aspergillus nidulans* metabolism at the genome-scale. *BMC Genomics,* 9(163).

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. & Steinmetz, L. M. (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U S A,* 103(14), pp. 5320-5325.

De Groot, P. W., Brandt, B. W., Horiuchi, H., Ram, A. F., de Koster, C. G. & Klis, F. M. (2009) Comprehensive genomic analysis of cell wall genes in *Aspergillus nidulans. Fungal Genet. Bio.*, suppl., 1, S72-81.

Decker, C. J. & R. Parker, R. (1993) A turnover pathway for both stable and unstable messenger-RNAs in yeast—evidence for a requirement for deadenylation. *Genes Dev.* 7, pp. 1632-1643.

Deonier, R., Tavaré, S. & Waterman, M. (2005) Computational Genome Analysis: an introduction. *Springer-Verlag,* p. 25.

D'haeseleer, P. (2006) What are DNA sequence motifs? *Nat. Biotech*., 24, pp. 423-425.

Dietzel, S., Schiebel, K., Little, G., Edelmann, P., Rappold, G.A., Eils, R., Cremer, C. & Cremer, T. (1999) The 3D positioning of ANT2 and ANT3 genes within female X chromosome territories correlates with gene activity. *Exp. Cell. Res*., 252, pp. 363–375.

Dilworth, R. P. (1950) A Decomposition Theorem for Partially Ordered Sets. *Annals of Mathematics*, 51(1), pp. 161-166.

Dohm, J. C., Lottaz, C., Borodina, T & Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucl. Acids. Res.*, 36(16), e105.

Dominski, Z. & Kole, R. (1991) Selection of splice sites in pre-mRNA with short internal exons. *Mol. Cell. Biol.*, 11, pp. 6075-6083.

Donaldson, M. E. & Saville, B. J. (2012) Natural antisense transcription in fungi. *Mol. Microbiol.*, 85(3), pp. 405-417.

Early, P., Rogers, J., Davis, M., Calame, K., Bond. M., Wall, R. & Hood. L. (1980) Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell*, 20, pp. 313–319.

ENCODE project consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, pp. 799-816.

Endo, Y., Yokoyama, M., Morimoto, M., Shirai, K., Chikimatsu, G., Kato, N., Tsukagoshi, N., Kato, M. & Kobayashi, T. (2008) Novel Promoter Sequence Required for Inductive Expression of the *Aspergillus nidulans* Endoglucanase Gene *egl. Bioscience, Biotechnology, and Biochemistry,* 72 (2), pp. 312-320.

Espeso, E. A. & Penalva, M. A. (1996) Three binding sites for the *Aspergillus nidulans* PacC zinc-finger transcription factor are necessary and sufficient for regulation by ambient pH of the isopenicillin N synthase gene promoter. *J. Biol. Chem*. 271, pp. 28825-28830.

Extebeste, O., Ugalde, U. & Espeso, E. A. (2010) Adaptive and developmental responses to stress in Aspergillus nidulans. *Curr. Protein Pept. Sci*., 11(8), pp. 704-18.

Faghihi, M.A. & Wahlestedt, C. (2009) Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.*, 10, pp. 637–643.

Fischle, W., Wang, Y. & Allis, C.D. (2003) Histone and chromatin cross-talk. *Curr. Opin. Cell. Biol.*, 15, pp. 172–183.

Fraser, J. A., Davis, M. A. & Hynes, M. J. (2001) The formamidase gene of *Aspergillus nidulans*: Regulation by nitrogen metabolite repression and transcriptional interference by an overlapping upstream gene. *Genetics*, 157, pp. 119-131.

Frith, M. C., Fu, Y., Yu, L., Chen, J. F., Hansen, U. & Weng, Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucl. Acids Res*., 32(4), pp.1372-1381.

Fritz, J. D., Greaser, M.L. & Wolff, J.A. (1991) A novel 3' extension technique using random primers in RNA-PCR. *Nucl. Acids Res.*, 19(13), p. 3747.

Fu, Y. H. & Marzluf, G. A. (1990) *nit-2*, the major nitrogen regulatory gene of *Neuropora crassa*, encodes a protein with a putative zinc finger DNA-binding domain. *Mol. Cell Biol.* 10, pp. 1055-1065.

Galagan, J. E, Calvo, S. E, Borkovich, K., Selker, E., Read, N.D., FitzHugh,W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang. S., Nielsen, C.B., Butler, J., Jaffe, D., Endrizzi, M., Qui, D., PIanakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., Jedd, G., Mewes. W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R.L., Perkins, D.D., Kroken, S., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P.C., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S, Lander, E.S, Nusbaum, C. & Birren, B. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, 422, pp. 859-868.

Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.J., Wortman, J.R., Batzoglou, S., Lee, S.I., Baştürkmen, M., Spevak, C.C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G.H., Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G.H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J.H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E.U., Archer, D.B., Peñalva, M.A., Oakley, B.R., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W.C., Denning, D.W., Caddick, M.X., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M.S., Osmani, S.A. & Birren, B.W. (2005) Sequencing of Aspergillus nidulans and comparative analysis with *A. fumigatus* and *A. oryzae. Nature*, 438(7017), pp. 1105-1115.

Gan, J., Tropea, J. E., Austin, B. P., Court, D. L., Waugh, D. S. & Ji, X. (2005) Intermediate States of Ribonuclease III in Complex with Double-Stranded RNA. *Structure*, 13, pp.1435-1442.

Ganley, A. R. D. & Kobayashi, T. (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res*., 17(2), pp. 184–191.

Gibbons, J. G., Beauvais, A., Beau, R., McGary, K. L., Latge, J. & Rokas, A. (2012) Global transcriptome changes underlying colony growth in the opportunistic human pathogen Aspergillus fumigatus. *Eukaryotic Cell*, 11, pp. 68-78.

Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P. & Rokas, A. (2009) Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol. Biol. Evol.*, 26(12), pp. 2731-2744.

Gilad, Y., Rifkin, S. A. & Pritchard, J. K. (2008) Revealing the architechture of gene regulation: the promise of eQTL studies. *Trands Genet*., 24(8), pp. 408-415.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996) Life with 6000 genes. *Science*, 274(546), pp. 563–567.

Gowda, M., Li, H., Alessi, J., Chen, F., Pratt, R. & Wang, G. (2006) Robust analysis of 5′-transcript ends (5′-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucl. Acids Res.*, 34(19), e126.

Gowda, M., Li, H. & Wang, G. (2007) Robust analysis of 5′-transcript ends: a high-throughput protocol for characterization of sequence diversity of transcription start sites. *Nature Protocols*, 2, pp. 1622 – 1632.

Grabowski, P.J. & Black, D. L. (2001) Alternative RNA splicing in the nervous system. *Prog. Neurobiol*., 65, pp. 289-308.

Graveley, B. R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*. 17, pp.100-7.

Guhaniyogi, J. & Brewer, G. (2001) Regulation of mRNA stability in mammalian cells. *Gene*, 265, pp. 11-23.

Gullerova, M. & Proudfoot, N.J. (2010) Transcriptional interference and gene orientation in yeast: noncoding RNA connections. *Cold Spring Harb. Symp. Quant. Biol.*, 75, pp. 299–311.

Gutierrez, S., Diez, B., Montenegro, E. & Martin, J.F. (1993) Characterization of the *Cephalosporium acremonium* pcbAB gene encoding alpha-aminoadipyl-cysteinyl-valine synthetase, a large multidomain peptide synthetase: linkage to the pcbC gene as a cluster of early cephalosporin biosynthetic genes and evidence of multiple functional domains. *J. Bacteriol.*, 173, pp. 2354–2365.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S. & Regav, A. (2010) *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol*. 28, pp. 503–510.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J. L., Root, D. E. & Lander, E. S. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*,477(7364), pp. 295-300.

Haas, H., Bauer, B., Redl, B., Stoffler, G. & Marzluf, G. A. (1995) Molecular cloning and analysis of *nre*, the major nitrogen regulatory gene of *Penicillium chrysogenum. Curr. Genet*. 27, pp. 150-158.

Hall, J., Husken, D. & Haner, R. (1996) Towards artificial ribonucleases: The sequencespecific cleavage of RNA in a duplex. *Nucl. Acids Res*., 24, pp. 3522–3526.

Hansen, K. D., Brenner, S. E. & Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucl. Acids Res.*, 38, e131.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S. & Frazer, K. A. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, 10, R32.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. & Matsushima, K. (2004) 5'-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.*, 22, pp. 1146-1149.

Heger, A. & Holm, L.U. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, 328(3), pp.749-67.

Hoffman, B. E. & Grabowski, P. J. (1992) U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev*., 6, pp. 2554-2568.

Holste, D. & Ohler, U. (2008) Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. *PLoS Comp. Biol.*, 4(1).

Hunt, A. G., Zu, R, Addepalli, B., Rao, S., Forbes, K. P., Meeks, L. R., Xing, D., Mo, M., Zhao, H., Bandyopadhyay, A., Dampanaboina, L., Marion, A., Von Lanken, C. & Li, Q. Q. (2008) *Arabidopsis* mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling. *BMC Genomics*, 9, pp. 220.

Husken, D., Goodall, G., Blommers, M. J., Jahnke, W., Hall, J., Haner, R. & Moser, H. E. (1996) Creating RNA bulges: Cleavage of RNA in RNA/DNA duplexes by metal ion catalysis. *Biochemistry* 35, pp. 16591–16600.

Ichikawa, J. K., Taft, B., Yang, J., Dimalanta, E., Yang, C., Chu, Y., Fu, H., Fantin, N., Menchen, S., Benson, S., Foster, E., Siddiqui, A., Suri, P., Ball, J., Sheridan, A., Laptewicz, M., Kasinskas, R., Shenasa, M., Tom, W., Berena, J. & Costa, G. (2010) Enhanced throughput and accuracy on the SOLiD 4 system. Life Technologies

Corporation. Poster presented at Advances in Genome Biology and Technology. Available online at: http://tools.invitrogen.com/content/sfs/posters/cms_079203.pdf

Ingolia, N.T. (2010) Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.* 470, pp. 119–142.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, pp. 218–223.

Ingolia, N. T., Lareau, L. F. & Weissman, J. S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147, pp. 789-802.

Ivanov, I. P., Loughran, G., Sachs, M. S. & Atkins, J. F. (2010) Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc. Natl. Acad. Sci. USA*, 107(42), pp. 18056-60.

Jacquenet, S., Ropers, D., Bilodeau, P. S., Damier, L., Mougin, A., Stoltzfus, C. M. & Branlant, C. (2001) Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3′ splice site and its cis-regulatory element: possible involvement in RNA splicing. *Nucl. Acids Res.*, 29, pp. 464-478.

Ji, X. (2008) The mechanism of RNase III action: how dicer dices. *Curr. Top Microbiol. Immunol.*, 320, pp. 99-116.

Jin, Y., Mead, J., Li, T., Wolberger C. & Vershon A. K. (1995) Altered DNA recognition and bending by insertions in the α2 tail of the yeast a1/α2 homeodomain heterodimer. *Science*, 270, pp. 290-292.

Jinks, J. L., Caten, C. E., Simchen, G. & Croft, J. H. (1966). Heterokaryon incompatibility and variation in wild populations of *Aspergillus nidulans*. *Heredity*, 21, pp. 227–239.

Jöchl, C., Rederstorff, M., Hertel, J., Stadler, P. F., Hofacker, I. L., Schretti, M., Hass, H. & Huttenhofer, A. (2008) Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein synthesis. *Nucl. Acids Res.*, 36(8), pp. 2677-2689.

Juven-Gershon, T. & Kadonaga, J. T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol*., 339, pp. 225-229.

Kafer, E. (1977) Meiotic and mitotic recombination in Aspergillus nidulans and its chromosomal aberrations. *Adv. Genet*., 19, pp. 33-131.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucl. Acids Res.*, 32(Database issue), D277–80.

Kankainen, M. & Löytynoja, A. (2007). MATLIGN: a motif clustering, comparison and matching tool. *BMC Bioinformatics* 8, pp.189.

Kawaji, H., Frith, M. C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P. & Hayashizaki, Y. (2006) Dynamic usage of transcription start sites within core promoters. *Genome Biol.*, 7, R118.

Keegan, L., Gill, G. & Ptashne, M. (1986) Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science*, 231, pp. 699–704.

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

Kibbe, W. A. (2007) OligoCalc: an online oligonucleotide properties calculator'. *Nucl. Acids Res.*, 35.

Kim, H. & Lipscomb, W. N. (1990). Crystal structure of the complex of carboxypeptidase A with a strongly bound phosphonate in a new crystalline form: comparison with structures of other complexes. *Biochemistry*, 29(23), pp. 5546–5555.

Klausner, R.D., Rouault, T.A. & Harford, J.B. (1993) Regulating the fate of mRNA: the control of cellular iron metabolism. *Cell*, 72, pp. 19-28.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y. & Carninci, P. (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, 3, pp. 211-222.

Kolasa, K. A., Morrow, J. R. & Sharma, A. P. (1993) Trivalent lanthanide ions do not cleave RNA in DNA–RNA hybrids. *Inorg. Chem.*, 32, pp. 3983–3984.

Kudla, B., Caddick, M. X., Langdon, T., Martinez-Rossi, N. M., Bennett, C. F., Sibley, S., Davies, R. W. & Arst, H. N. Jr, (1990) The regulatory gene *areA* mediating nitrogen metabolite repression in *Aspergillus nidulans*. Mutations affecting specificity of gene activation alter a loop residue of a putative zinc finger. *EMBO J.* 9(5), pp. 1355-1364.

Kulmburg, P., Mathieu, M., Dowzer, C., Kelly, J. & Felenbok, B. (2006) Specific binding sites in the alcR and alcA promoters of the ethanol regulon for the CREA repressor mediating carbon catabolite repression in *Aspergillus nidulans*. *Mol. Microbiol.* 7(6), pp. 847-857.

Kunitz, M. (1950) Crystalline desoxyribonuclease; isolation and general properties; spectrophotometric method for the measurement of desoxyribonuclease activity. *J. Gen. Physiol.*, 33, pp. 349-362.

Kupfer, D. M., Drabenstot, S. D., Buchanan, K. L., Lai, H., Zhu, H., Dyer, D. W., Roe, B. A. & Murphy, J. W. (2004) Introns and splicing elements of five diverse fungi. *Eukaryot. Cell*, 3, pp.1088-1100.

Kurz, A., Lampel, S., Nickolenko, J.E., Bradl, J., Benner, A., Zirbel, R.M., Cremer, T. & Lichter, P. (1996) Active and inactive genes localize preferentially in the periphery of chromosome territories. *J. Cell. Biol.* 135, pp. 1195–1205.

Konarska, M. M., Padgett, R. A. & Sharp, P. A. (1984) Recognition of cap structure in splicing *in vitro* of mRNA precursors. *Cell,* 38(3), pp. 731-736.

Kutach, A. K. & Kadonaga, J. T. (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol. Cell. Biol*., 20, pp. 4754-4764.

Lachner, M., O'Sullivan, R. J. & Jenuwein, T. (2003) An epigenetic road map for histone lysine methylation. *J. Cell. Sci.*, 116, pp. 2117–2124.

Lallena, M. J., Chalmers, K. J., Llamazares, S., Lamond, A. I. & Valcarcel, J. (2002) Splicing Regulation at the Second Catalytic Step by Sex-lethal Involves 3' Splice Site Recognition by SPF45. *Cell,* 109, pp. 285–96.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10, R25.

Lapidot, M. & Pilpel, Y. (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep.*, 7, pp. 1216–1222.

Lee, H., Li, L., Gu, W., Xue, Z., Crosthwaite, S. K., Pertsemlidis, A., Lewis, Z. A., Freitag, M., Selker, E. U., Mello, C. C. & Liu, Y. (2010) Diverse pathways generate microRNA-like RNAs and dicer-independent small interfering RNAs in fungi. *Molecular cell,* 38(6), pp. 803–814.

Lee, T. I. & Young, R. A. (2000) "Transcription of eukaryotic protein-coding genes". *Annu. Rev. Genet.*, 34, pp. 77–137.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, pp. 2078-9.

Li, L., Chang, S.S. & Liu, Y. (2010) RNA interference pathways in filamentous fungi. *Cell Mol. Life Sci.*, 67, pp. 3849-3863.

Libri, D., Piseri, A. & Fiszman, M. Y. (1991) Tissue-specific splicing in vivo of the beta-tropomyosin gene: dependence on an RNA secondary structure. *Science,* 252, pp. 1842-1845.

Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., Thompson, J. F., Letovsky, S., Milos, P & Causey, M. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.*, 27, pp. 652–658.

Lockard, R. E., Rieser, L. & Vournakis J.N. (1981) Labeling of eukaryotic messenger RNA 5' terminus with phosphorus -32: use of tobacco acid pyrophosphatase for removal of cap structures. *Gene Amplif. Anal*., 2, pp. 229–51.

Loftus, B. J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I. J., Fraser, J. A., Allen, J. E., Bosdet, I. E., Brent, M. R., Chiu, R., Doering, T. L., Donlin, M. J., D'Souza, C. A., Fox, D. S., Grinberg, V., Fu, J., Fukushima, M., Haas, B. J., Huang, J. C., Janbon, G., Jones, S. J., Koo, H. L., Krzywinski, M. I., Kwon-Chung, J. K., Lengeler, K. B., Maiti, R., Marra, M. A., Marra, R. E., Mathewson, C. A., Mitchell, T. G., Pertea, M., Riggs, F. R., Salzberg, S.L., Schein, J. E., Shvarsbeyn, A., Shin, H., Shumway, M., Specht, C. A., Suh, B. B., Tenney, A., Utterback, T. R., Wickes, B. L., Wortman, J. R., Wye, N. H., Kronstad, J. W., Lodge, J. K., Heitman, J., Davis, R. W., Fraser, C. M. & Hyman, R. W. (2005) The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans. Science*, 307, pp. 1321-1324.

Lyne, R.,. Burns, G., Mata, J., Penkett, C. J., Rustici, G., Chen, D., Langford, C., Vetrie, D. & Bähler, J. (2003) Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics*, 4, p. 27.

Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K., Arima, T., Akita, O., Kashiwagi, Y., Abe, K., Gomi, K., Horiuchi, H., Kitamoto, K., Kobayashi, T., Takeuchi, M., Denning, D. W., Galagan, J. E., Nierman, W. C., Yu, J., Archer, D. B., Bennett, J. W., Bhatnagar, D., Cleveland, T. E., Fedorova, N. D., Gotoh, O., Horikawa, H., Hosoyama, A., Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P. R., Kato, M., Kato, Y., Kin, T., Kokubun, A., Maeda, H., Maeyama, N., Maruyama, J., Nagasaki, H., Nakajima, T., Oda, K., Okada, K., Paulsen, I., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K., Terabayashi, Y., Wortman, J. R., Yamada, O., Yamagata, Y., Anazawa, H., Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka, A., Isono, K., Kuhara, S., Ogasawara, N. & Kikuchi, H. (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*, 438, pp. 1157–1161.

Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet., 24*(3), pp. 113-141.

Martinelli, S. D. & Kinghorn J. R. (1994) *Aspergillus*: 50 years on. *Elsevier.*

Maruyama, J., Juvvadi, P. R., Ishi, K. & Kitamoto, K. (2005) Three-dimensional image analysis of plugging at the septal pore by Woronin body during hypotonic shock inducing hyphal tip bursting in the filamentous fungus *Aspergillus oryzae*. *Biochem. Biophys. Res. Commun.* 331, pp. 1081–1088.

Mata, J. & Bähler, J. (2003) Correlations between gene expression and gene conservation in fission yeast. *Genome Res.*, 13, pp. 2686–2690.

Matsui, T., Segall, J., Weil, A. & Roeder, R.G. (1980) Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J. Biol. Chem.,* 255, pp. 11992-11996.

McComb, R. B., Bowers, G. N. Jr. & Posen, S. (1979) Alkaline Phosphatase. New York: Plenum.

McGuire, A. M. & Church, G. M. (2000) Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucl. Acids. Res*., 28(22) pp. 4523-4530.

McGuire, A. M., Pearson, M. D., Neafsey, D. E. & Galagan, J. E. (2008) Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biol.***,** 9,R50.

Meader, S., Hillier, L. W., Locke, D., Ponting, C. P. & Lunter, G. (2010) Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Res.* 20, pp. 675–84.

Mehlmanna, M., Townsenda, M. B., Stearsb, R. L., Kuchtaa, R. D. & Rowlen, K. L. (2005) Optimization of fragmentation conditions for microarray analysis of viral RNA. *Analytical Biochemistry,* 347(2), pp. 316–323.

Millan, J.L. (2006) Alkaline Phosphatases: Structure, substrate specificity and functional relatedness to other members of a large superfamily of enzymes. *Purinergic Signal*, 2, pp. 335–341.

Mitchell, P. & Tollervey, D. (2000) Musing on the structural organization of the exosome complex. *Nat. Struct. Biol.,* 7, pp. 843-846.

Mizuguchi, H., Nakatsuji, M., Fujiwara, S., Takagi, M., and Imanaka, T. (1999) Characterization and application to hot start PCR of neutralizing monoclonal antibodies against KOD DNA polymerase. *J. Biochem.,* 126, pp. 762–768.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. & Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucl.Acids Res*., 35, W182-W185.

Morozov, I. Y. & Caddick, M. (2012) Cytoplasmic mRNA 3' tagging in eukaryotes: Does it spell the end? *Biochem. Soc Trans*. 40(4), pp. 810-814.

Morozov, I. Y., Jones, M. G., Razak, A. A, Rigden, D. J. & Caddick, M. X. (2010) CUCU modification of mRNA promotes decapping and transcript degradation in *Aspergillus nidulans*. *Mol. Cell. Biol.,* 30, pp. 460-469.

Morozov, I. Y., Martinez, M. G., Jones, M. G. & Caddick, M. X. (2000). A defined sequence within the 3' UTR of the areA transcript is sufficient to mediate nitrogen metabolism signalling via accelerated deadenylation. *Mol. Microbiol*., 37(5), pp. 1248-1257

Morris, N.R. (1976a) A temperature-sensitive mutant of *Aspergillus nidulans* reversible blocked in nuclear division. *Experimental Cell Research*, 98(1),pp. 204-209.

Morris, N. R. (1976b) Mitotic mutants of Aspergillus nidulans. *Genet. Res. Camb*., 26, pp. 237-254.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420, pp. 520–562

Muhlrad, D., Decker, C. J. & Parker, R. (1994) Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5' ->3' digestion of the transcript. *Genes Dev*. 8, pp. 855-866.

Munroe, S.H. & Zhu, J. (2006) Overlapping transcripts, double-stranded RNA and antisense regulation: a genomic perspective. *Cell Mol. Life Sci.*, 63, pp. 2102–2118.

Muro-Pastor, M. I.,Gonzalez, R., Strauss, J., Narendja, F. & Scazzocchio,' C. (1999) The GATA factor AreA is essential for chromatin remodelling in a eukaryotic bidirectional promoter. *The EMBO Journal*, 18, pp. 1584 – 1597.

Murphy, J. E., Tibbitts, T. T. & Kantrowitz, E. R. (1995) Mutations at positions 153 and 328 in *Escherichia coli* alkaline phosphatase provide insight towards the structure and function of mammalian and yeast alkaline phosphatases. *J. Mol. Biol.*, 253(4), pp. 604–617.

Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R. & Fluhr, R. (2004) Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J.*, 39, pp. 877-885.

Nierman, W. C., Pain, A., Anderson, M. J., Wortman, J. R., Kim, H. S., Arroyo, J., Berriman, M., Abe, K., Archer, D. B., Bermejo, C., Bennett, J., Bowyer, P., Chen, D., Collins, M., Coulsen, R., Davies, R., Dyer, P. S., Farman, M., Fedorova, N., Fedorova, N., Feldblyum, T. V., Fischer, R., Fosker, N., Fraser, A., Garcia, J. L., Garcia, M. J., Goble, A., Goldman, G. H., Gomi, K., Griffith-Jones, S., Gwilliam, R., Haas, B., Haas, H., Harris, D., Horiuchi, H., Huang, J. Q., Humphray, S., Jimenez, J., Keller, N., Khouri, H., Kitamoto, K., Kobayashi, T., Konzack, S., Kulkarni, R., Kumagai, T., Lafon, A., Latge, J. P., Li, W. X., Lord, A., Lu, C., Majoros, W. H., May, G. S., Miller, B. L., Mohamoud, Y., Molina, M., Monod, M., Mouyna, I., Mulligan, S., Murphy, L., O'Neil, S., Paulsen, I., Penalva, M. A., Pertea, M., Price, C., Pritchard, B. L., Quail, M. A., Rabbinowitsch, E., Rawlins, N., Rajandream, M. A., Reichard, U., Renauld, H., Robson, G. D., de Cordoba, S. R., Rodriguez-Pena, J. M., Ronning, C. M., Rutter, S., Salzberg, S. L., Sanchez, M., Sanchez-Ferrero, J. C., Saunders, D., Seeger, K., Squares, R., Squares, S., Takeuchi, M., Tekaia, F., Turner, G., de Aldana, C. R. V., Weidman, J., White, O., Woodward, J., Yu, J. H., Fraser, C., Galagan, J. E., Asai, K., Machida, M., Hall, N., Barrell, B. & Denning, D.W. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, 438, pp.1151–1156.

Nunes, C.C., Sailsbery, J.K. & Dean, R.A. (2011) Characterization and application of small RNAs and RNA silencing mechanisms in fungi. *Fungal Biol. Rev.*, 25, pp. 172–180.

Orphanides, G., Lagrange, T. & Reinberg, D. (1996) The general transcription factors of RNA polymerase II. *Genes Dev.*, 10, pp. 2657-2683.

Ozsolak, F., Goren, A., Gymrek, M. A., Guttman, M., Regev, A., Bernstein, B. E. & Milos, P. M. (2010) Digital transcriptome profiling from attomole-level RNA samples. *Genome Res.*, 20, pp. 519–525.

Ozsolak, F. & Milos, P. M. (2011) RNA sequencing: advantages, challenges and opportunities. *Nat. Rev. Genet.*, 12(2), pp. 87-98.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genet.*, 40, pp. 1413–1415.

Paoletti, M., Rydholm, C., Schwier, E. U., Anderson, M. J., Szakacs, G., Lutzoni, F., Debeaupuis, J. P., Latge, J. P., Denning, D. W. & Dyer, P. S. (2005) Evidence for sexuality in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Curr. Biol.*, 15, pp. 1242-1248.

Platt, A., Ravagnani, A., Arst, H., Kirk, D., Langdon, T. & Caddick, M. X. (1996).Mutational analysis of the C-terminal region of AREA, the transcription factor mediating nitrogen metabolite repression in *Aspergillus nidulans*. *Molecular & General Genetics,* 250, pp.106-114.

Pontecorvo, G., Roper, J. A., Hemmons, L. M., Macdonald, K. D. & Bufton, A. W. J. (1953) The genetics of *Aspergillus nidulans*. *Advances in Genetics,* 5, pp. 141–238.

Ponting, C.P., Oliver, P.L. & Reik, W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, 136, pp. 629–641.

Prasanth, K.V. & Spector, D.L. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.*, 21, pp. 11–42.

Proudfoot, N. J., Furger, A. & Dye, M. J. (2002) Integrating mRNA processing with transcription. *Cell*, 108(4), pp. 501-512.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Stonnhammer, E. L. L., Eddy, S. R., Bateman, A. & Finn, R. D. (2012). The Pfam protein families database. *Nucl. Acids Res.*, 40:D290-D301.

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Ravagnani, A., Gorfinkiel, L., Langdon, T., Diallinas, G., Adjaji, E., Demais, S., Gorton, D., Arst, H. N. Jr, & Scazzocchio, C. (1997) Subtle hydrophobic interactions between the seventh residue of the zinc finger loop and the first base of an HGATAR sequence determine promoter-specific recognition by the *Aspergillus nidulans* GATA factor AreA. *EMBO J.*, 16, pp. 3974-3986.

Reinke, H. & Horz, W. (2003) Histones are first hyperacetylated and then lose contact with the activated *PHO5* promoter. *Molecular Cell,* 11, pp. 1599-1607.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. (2011a) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12, R22.

Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. (2011b) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*.

Robertson, H. D., Webster, R. E. & Zinder, N. D. (1968). Purification and properties of ribonuclease III from *Escherichia coli. J. Biol. Chem.,* 243(1), pp. 82.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. (2011) Integrative Genomics Viewer. *Nat. Biotech.*, 29, pp. 24-26.

Romfo, C. M., Alvarez, C. J., van Heeckeren, W. J., Webb, C. J. & Wise, J. A (2000) Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. *Mol. Cell. Biol.*, 20, pp. 7955-7970.

Ross, J. (1995) mRNA stability in mammalian cells. *Microbiol Rev*., 59, pp. 423-450

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 16, pp. *939*–945.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10), pp. 944-5.

Salamoy, A. A. & Solovyev, V. V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research*, 10, pp. 516–522.

Salzberg, S. L. & Yorke, J. A. (2005) Beware of mis-assembled genomes. *Bioinformatics,* 21, pp. 4320–4321.

Scazzocchio, C. (2000) The fungal GATA factors. *Curr. Opin. Microbiol.,* 3, pp. 126-131.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E. & Zipursky, S. L. L. (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101, pp. 671–68.

Schneider, T. D., Stephens, R. M. (1990) Sequence Logos: A new way to display consensus sequences. *Nucl. Acids Res., 18*, pp. 6097-6100.

Schumann, U., Ayliffe, M., Kazan, K. & Wang, M.B. (2010) RNA silencing in fungi. *Front. Biol.,* 5, pp. 478–494.

Shen, Y., Ji, G., Haas, B. J., Wu, X., Zheng, J., Reese, G. J. & Li, Q. Q. (2008) Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucl. Acids Res.,* 36(9), pp. 3150–61.

Sharakhov, I. V., Serazin, A. C., Grushko, O. G., Dana, A., Lobo, N., Hillenmeyer, M. E., Westerman, W., Romero-Severson, J., Costantini, C., Sagnon, N., Collins, F. H. & Besansky, N. J. (2002) Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestus. Science,* 298, pp. 182-185.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kaiai, J., Carninci, P. & Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA,* 100, pp. 15776-15781.

Sinha, S. & Tompa, M. (2000) A statistical method for finding transcription factor binding sites. *Eighth International Conference on Intelligent Systems for Molecular Biology*, San Diego, CA, pp. 344-354.

Sinha, S. & Tompa, M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucl. Acids Res.,* 30, pp. 5549-60.

Sinha, S. & Tompa, M. (2003) YMF, a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucl. Acids Res.,* 31(13), pp. 3586-3588.

Smale, S. T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.,* 15(19), pp. 2503-2508.

Smale, S. T. & Kadonaga, J. T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.,* 72, pp. 449-479.

Smith, C.A., Robertson, D., Yates, B., Nielsen, D.M., Brown, D., Dean, R.A. & Payne, G.A. (2008) The effect of temperature on natural antisense transcript (NAT) expression in *Aspergillus flavus*. *Curr. Genet.,* 54, pp. 241–269.

Smith, C. W. & Nadal-Ginard, B. (1989) Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell,* 56, pp. 749-758.

Spanu, P. D. & Doyle, K. (2009) Analysis of 5′ transcript heterogeneity by high-throughput sequencing of cDNA. *Nat. Methods,* 6.

Stormo, G. D. (2000) DNA binding sites: representation and discovery. *Bioinformatics,* 16, 16–23.

Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98, pp. 1-4.

Su, W.Y., Xiong, H. & Fang, J.Y. (2010) Natural antisense transcripts regulate gene expression in an epigenetic manner. *Biochem. Biophys. Res. Commun.,* 396, pp. 177–181.

Sugnet, C. W., Kent, W. J., Ares, M Jr. & Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput*., pp. 66-77.

Sullivan, D., Fahey, B & Titus, D. (2006) Fast PCR: General considerations for minimizing run times and maximizing throughput. *Cold Spring Harb. Protoc.*

Surosky, R.T. & Esposito, R.E. (1992) Early meiotic transcripts are highly unstable in *Saccharomyces cerevisiae. Mol. Cell. Biol.,* 12, pp. 3948-3958.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K. & Sugano, S. (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res*., 11, pp. 677-84.

Swanson, M. J., Qiu, H., Sumibcay, L., Krueger, A., Kim, S. J., Natarajan, K., Yoon, S. & Hinnebusch, A. G. (2003) A multiplicity of coactivators is required by Gcn4p at individual promoter *in vivo*. *Mol. Cell. Biol*, 23, pp. 2800-2820.

Takagi, T., Cho, E., Janoo, R. T. K., Polodny, V., Takase, Y., Keogh, M. C., Woo, S., Fresco-Cohen, L. D., Hoffman, C. S. & Buratowski, S. (2002) Divergent subunit interactions among fungal mRNA 5'-capping machineries. *Eukaryot. Cell*, 1(3), pp. 448-457.

Tan, S. & Richmond, T. J. (1998) Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature,* 391, pp. 660-666.

Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M. & Mattick, J.S. (2010) Non-coding RNAs: regulators of disease. *J. Pathol.,* 220, pp. 126–139.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) Systematic determination of genetic network architecture. *Nat. Genet*., 22, pp. 281–285.

Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J.J., Le Texier, V. & Muilu, J. (2004) ASD: the alternative splicing database. *Nucl. Acids Res*., 32 (Database), D64-D69.

The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1), pp. 25-9.

The FANTOM Consortium (2005) The transcriptional landscape of the mammalian genome. *Science*, 309, pp. 1559–1563.

Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics.*

The SAM Format Specification Working Group (2011). The SAM format specification (v1.4-r985). Available online at http://samtools.sourceforge.net/SAM1.pdf [Accessed 05/08/2012]

Timberlake, W.E. (1980) Developmental gene regulation *in Aspergillus nidulans*. *Dev. Biol.,* 78(2), pp. 497-510.

Timberlake, W. E. & Marshall, M. A. (1989) Genetic engineering of filamentous fungi. *Science*, 244, pp. 1313–1317.

Tisseur, M., Kwapisz, M. & Morillon, A. (2011) Pervasive transcription – lessons from yeast. *Biochimie.* 93, pp. 1889–1896.

Tophat manual. Available online at: http://tophat.cbcb.umd.edu/manual.html [Accessed 14/08/2012]

Trcek, T., Larson, D.R., Moldón, A., Query, C.C. & Singer, R.H. (2011) Single-molecule mRNA decay measurements reveal promoter-regulated mRNA stability in yeast. *Cell*, 147, p p.1484-1497.

Trapnell, C. (2012) How Cufflinks Works. Available online at http://cufflinks.cbcb.umd.edu/howitworks.html [Accessed 13/06/2012]

Trapnell, C., Pachter, L. & Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25 (9), pp. 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A.M., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.b & Pachter L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.*, 28, pp. 511-515.

Trevisan, G. L., Oliveira, E. H., Peres, N. T., Cruz, A. H., Martinez-Rossi, N. M. & Rossi, A. (2011) Transcription of *Aspergillus nidulans* pacC is modulated by alternative RNA splicing of palB. *FEBS Lett.*, 585(21), pp. 3442-3445.

Tucker, M. & Parker, R. (2000) Mechanisms and control of mRNA decapping in *Saccharomyces cerevisiae*. *Annu. Rev. Biochem.,* 69, pp. 571-595.

Vanecko, S. & Laskowski, M., Sr. (1961) Studies of the specificity of deoxyribonuclease I. III. Hydrolysis of chains carrying a monoesterified phosphate on carbon 5'. *J. Biol. Chem.,* 236, pp. 3312-3316.

Vardhanabhuti, S., Wang, J. & Hannenhalli, S. (2007) Position and distance specificity are important determinants of *cis*-regulatory motifs in addition to evolutionary conservation. *Nucl. Acids Res.,* 35(10), pp. 3203-3213.

Vitale, D., (2001) Optimizing cRNA fragmentation for microarray experiments using the Agilent 2100 bioanalyzer. *Agilent technologies.*, 5988-3119EN.

Vivancos, A. P., Güell, M., Dohm, J. C., Serrano, L. & Himmelbauer, H. (2010) Strand-specific deep sequencing of the transcriptome. *Genome Res.,* 20, pp. 989-999.

Volpi, E.V., Chevret, E., Jones, T., Vatcheva, R., Williamson, J., Beck, S., Campbell, R.D., Goldsworthy, M., Powis, S.H., Ragoussis, J., Trowsdale, J. & Sheer, D. (2000) Large-scale chromatin organisation of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell. Sci*., 113, pp. 1565–1576.

Vorobjev, P. E. & Zarytova , V. F. (2000) Cleavage of RNA in hybrid duplexes by ribonuclease H of *E. coli* : I. Substrate properties of complexes formed by RNA and a tandem of short oligodeoxyribonucleotides. *Russian Journal of Bioorganic Chemistry*, 26, pp. 656-66.

Wagner, D., Schmeinck, A., Mos, M., Morozov, I. Y., Caddick, M. X. & Tudzynski, B. (2010) The bZIP transcription factor MeaB mediates nitrogen metabolite repression at specific loci. *Eukaryot. Cell*, 9(10), pp.1588-601.

Wall, L. (1999) Perl, the first postmodern computer language. Available online at: http://www.perl.com/pub/1999/03/pm.html [Accessed: 05/09/2012]

Wang, B. B. & Brendel, V. (2006) Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. USA*, 103, pp.7175-7180.

Wang, B., Guo, G., Wang, C., Lin, Y., Wang, X., Zhao, M., Guo, Y, He, M., Zhang, Y. & Pan, L. (2010) Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. *Nucl. Acids Res.,* 38(15), pp. 5075-5087.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Shroth, G. P. & Burge, C. B (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456, pp. 470–476.

Ward, O. P., Qin, W. M., Dhanjoon, J., Ye, J. & Sign, A. (2006) Physiology and biotechnology of *Aspergillus. Adv. Appl. Microbiol*., 58; pp. 1-75.

Weiss, R. B. & Atkins, J. F. (2011) Translation goes global. *Science*, 334, pp. 1509-1510.

Werner, A., Carlile, M. & Swan, D. (2009) What do natural antisense transcripts regulate? *RNA Biol.* 6, pp. 43–48.

Wilhelm, B. T. & Landry, J.R. (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48, pp. 249-257.

Wiame, J. M., Grenson, M. & Arst, H. N. (1985) Nitrogen Catabolite Repression in Yeasts and Filamentous Fungi. *Adv. Micro. Phys.*, 26**,** pp.1-88

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J. & Bähler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453, pp. 1239-1243.

Williams, G. (2000). Emboss: Getorf. Available online at http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html [Accessed 09/06/2012]

Williams R. R. E., Broad, S., Sheer, D. & Ragoussis, J. (2002) Subchromosomal positioning of the epidermal differentiation complex (EDC) in keratinocyte and lymphoblast interphase nuclei. *Exp. Cell. Res*., 272, pp. 163–175.

Wilusz, J.E., Sunwoo, H. & Spector, D.L. (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, 23, pp. 1494–1504.

Wortman, J. R., Gilsenan, J. M., Joardar, V., Deegan, J., Clutterbuck, J., Andersen, M. R., Archer, D., Bencina, M., Braus, G., Coutinho, P., von Döhren, H., Doonan, J., Driessen, A. J., Durek, P., Espeso, E., Fekete, E., Flipphi, M., Estrada, C. G., Geysens, S., Goldman, G., de Groot, P. W., Hansen, K., Harris, S. D., Heinekamp, T., Helmstaedt, K., Henrissat, B., Hofmann, G., Homan, T., Horio, T., Horiuchi, H., James, S., Jones, M., Karaffa, L., Karányi, Z., Kato, M., Keller, N., Kelly, D. E., Kiel, J. A., Kim, J. M., van der Klei, I. J., Klis, F. M., Kovalchuk, A., Krasevec, N., Kubicek, C. P., Liu, B., Maccabe, A., Meyer, V., Mirabito, P., Miskei, M., Mos, M., Mullins, J., Nelson, D. R., Nielsen, J., Oakley, B. R., Osmani, S. A., Pakula, T., Paszewski, A., Paulsen, I., Pilsyk, S., Pócsi, I., Punt, P. J., Ram, A. F., Ren, Q., Robellet, X., Robson, G., Seiboth, B., van Solingen, P., Specht, T., Sun, J., Taheri-Talesh, N., Takeshita, N., Ussery, D., vanKuyk, P. A., Visser, H., van de Vondervoort, P. J., de Vries, R. P., Walton, J., Xiang, X., Xiong, Y., Zeng, A. P., Brandt, B. W., Cornell, M. J., van den Hondel, C. A., Visser, J., Oliver, S. G. & Turner, G. (2008) The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort. *Fungal Genet. Biol*., 46 Suppl. 1:S2-13.

Wright, M.W. & Bruford, E.A. (2011) Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics*, 5, pp. 90–98.

Yamamoto, Y. Y., Yoshioka, Y., Hyakumachi, M. & Obokata, J.,(2011) Characteristics of core promoter types with respect to gene structure and expression in *Arabidopsis thaliana. DNA Res.*, 18(5), pp.333-42.

Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K. & Rotman, G. (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, 21(4), pp. 379-86.

Zagorowska, I., Kuusela, S. & Lonnberg, H. (1998) Metal ion-dependent hydrolysis of RNA phosphodiester bonds within hairpin loops. A comparative kinetic study on chimeric ribo/2′-O-methylribo oligonucleotides. *Nucl. Acids Res.*, 26, pp. 3392–3396.

Zeugin, J. A. & Hartley, J. L. (1985). Ethanol precipitation of DNA. Focus, 7(4), pp. 1-2.

Zhang, Y., Liu, X. S., Liu, Q. & Wei, L. (2006) Genome-wide *in silico* identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. *Nucl. Acids Res.*, 34(12), pp. 3465–3475.

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G. J., Wang, X. & Qi, Y (2007). A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii. Genes Dev,*. 21, pp. 1190–1203.

Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L. & Lander, E. S. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458, pp. 223–227.

Appendix 1. Solutions and media

1.1. Buffers and solutions for general molecular biology

0.5 M EDTA (pH 8.0), per 1 litre:

168.1 g EDTA (pH 8.0)

1 x Tris EDTA (TE) buffer, per 1 litre;

3.72 g EDTA, 12.11 g Tris-HCL (pH 7.5). Steralised by autoclaving.

50 x Tris Acetate (TAE) buffer, per 1 litre;

242 g Tris base, 57.1 ml glacial acid, 100 ml 0.5 M EDTA (pH 8.0)

10 x gel loading buffer

50 ml glycerol, 25 ml 1.0 M EDTA (pH 8.0), 100 mg bromophenol blue.

Phenol:

Pre-made aqua phenol [Q-Biogene] was used, to which 1/10 volumes of 1 M sodium acetate (pH 5.0) was added for use in RNA extraction.

1.2. *Aspergillus nidulans* solutions and media

Salts solutions, per 1 litre;

KCL	26 g
$MgSO_4$ $7H_2O$	26g
Kh_2PO_4	76 g
trace elements solution	50 ml
sodium tetraborate	0.04 g
cupric sulphate	0.4 g
ferric orthophosphate	0.8 g
manganese sulphate	0.8 g
sodium molybdate	0.8 g
zinc sulphate	8.0 g

solution stored at 4°C

Vitamin solution, per litre;

p-aminobenzoic acid	0.4 g
inositol	0.4 g
nicotine acid	0.1 g
calcium pantothenate	0.6 g
pyridoxine	0.25 g
choline chloride	1.4 g
riboflavin	0.1 g
putrecine	2.0 g

D-biotin	10 ml
pyridoxine (calcium salt)	0.25 g

solution stored at 4°C

Trace elements solution, per litre;

Sodium tetraborate	0.04 g
Cupric sulphate	0.4 g
Ferric orthophosphate	0.8g
Manganese sulphate	0.8 g
Sodium molybdate	0.8 g
Zinc sulphate	8.0 g

Complete medium (CM), per litre;

Glucose	10 g
Aspergillus salts solution	20 ml
Vitamin solution	10 ml
Yeast extract	1 g
Peptone	2 g
Casamino acids	1 g
Adenine	75 mg

Adjusted to pH 6.5.

Minimal media (MM), per 1 litre;

Glucose	10 g
Aspergillus salts solution	20 ml

Adjusted to pH 6.5

1.3. Sterilisation of media

All media were autoclaved for 20 minutes at 15 psi and stored at 4°C.

1.4. Solid media

Where required, solid media was created by the addition of 2 % Agarose to MM or CM.

Appendix 2. Software and data

Appendix 2 is supplied as a DVD attached to this thesis. Files were produced on a Mac operating system, for best results in a Windows environment, it is recommended that text files be opened in wordpad.