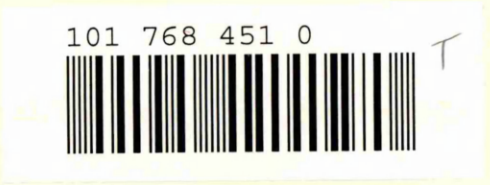


The Genomics of *Aspergillus fumigatus*

John Robert Woodward

A thesis submitted in partial fulfilment of the requirements of
Sheffield Hallam University

For the Degree of Master of Philosophy

DECEMBER 2003

"Lay down, lie back, shut up, submit"!

Zodiac Mindwarp and the Love Reaction
1988


SHEFFIELD HALLAM UNIVERSITY
TH
ADSETTS CENTRE

Acknowledgments

There are many people throughout my time working on this project that have helped me. Each individual has contributed in their own way, be it either negative or positive, but in all I feel that I am stronger because of it. Many events have also occurred during my time that have shaped the way I see and perceive what is around me in what has become one of the most challenging episodes of my life. Each individual and each event in turn has made me the person I am now.

The one beacon that has kept me going is the memory of my mother. I started this project not long after her passing in an attempt to better myself and throw myself into something that both her and I could be proud of. Her influence on my life gave me strength, in her likeness, and if it weren't for that, I would not be submitting this thesis right now. There is little to say about her in that I can only thank her for what she gave to me and continues to give to me. And for that I am eternally grateful.

One person who has also been my rock is Liz. I know I must have been hell to live with, yet she is still with me and has not muttered a word of indignation or scathing when I, on the other hand, have been like a bear with a sore head! Rather unfairly, she has been my emotional punch bag and I realise that I might not have been the nicest guy to be around on a number of occasions when things were going badly. But she is still there and I can only try and make this up to her now this whole thing is over. Thank you Liz. You have meant everything.

I have to also thank the people that made this whole thing possible. Thanks go to Dr. Jenny Shelton, who had the balls to take me on (again!) after knowing from my undergraduate days what I was like! Thanks also to Dr. Mike Quail, who has put up with my incessant questions and pestering, moaning and whinging. I wouldn't have

put up with me! To Dr. Neil Hall, Marie-Adele Rajandreem and Dr. Bart Barrell for allowing me to use their resources, allowing me the time to start and complete the project and paying me in the process! To Dr. David Denning, for it is he that got the project up and running in the first place. And to all the people I have worked with in the various departments along the way. They have all helped in their own way.

Of course, family and friends are also of great help and their constant support is not forgotten. To them I give GREAT thanks. Each and all have been as encouraging as they can be and I am eternally grateful.

And last, but not least, there are the other influences on my life. Aston Villa Football Club, you kill me! But those 90mins on a Saturday are one of the greatest escapes still. I want to thank Bruce Lee, for his philosophy, his way and his art. A man of great discipline and an example to all. And I want to thank the passions in my life that give me thrills, spills and escape: Cars, sport, music.....the list can go on. Whatever has made its mark in these last four years, thank you. I hope to be like you one day.....

Abstract

In 1999, a proposal was put forward for a pilot project to sequence a region of the medically important pathogenic fungus *Aspergillus fumigatus* to determine whether a whole genome sequencing project was viable. This thesis reports on the generation of a Bacterial Artificial Chromosome library and the sequencing and annotation of a 1Mb region surrounding the *niaD* gene cluster. Identification of the BAC containing the *niaD* gene was achieved through hybridisation and the clones "walked" out from this. The final minimal tiling path contig used for sequencing consisted of 16 overlapping clones. The total sequence completed was 921,539bp which, after annotation, was shown to contain 341 predicted putative protein coding genes and 8 tRNA genes with a GC content of 50.6%. Of significance was the synteny found in the *qut/qa* gene cluster between *A. fumigatus* and *A. nidulans*. There was complete conservation of this cluster between the two, except for the inversion of two regions, *facC* and *trpC*. An *aroM* gene was also found, which has a known involvement in aromatic compound synthesis and could be a good candidate for possible drug targets.

Abbreviations

ABPA	Allergic Bronchopulmonary Aspergillosis
ACT	Artemis Comparison Tool
AIDS	Acquired Immune Deficiency Syndrome
AFAR	Aflatoxin B1-aldehyde reductase
AKR	Aldo-ketone Reductase
AmB	Amphotocerin B
ATPase	Adenosine Tri-phosphate-ase
BAC	Bacterial Artificial Chromosome
bp	Base pairs
BSA	Bovine Serum Albumin
CB	Consensus band
cDNA	Complementary Deoxyribonucleic Acid
CDS	Coding Sequence
CIAP	Calf Intestinal Alkaline Phosphatase
CsCl	Caesium Chloride
CSF	Cerebral Spinal Fluid
dCTP	di-Cytosine Tri-phosphate
DDW	Double Deionised Water
DNA	Deoxyribonucleic Acid
dNTP	di-Nucleotide Tri-phosphate
EAA	Extrinsic Allergic Alveolitis
EC	Enzyme Commission Number
EDTA	Ethylenediaminetetraacetic acid
EST	Expressed Sequenced Tags
EtBr	Ethidium Bromide
FGI	Fungal Genome Initiative
GO	Gene Ontology
GST	glutathione-S-transferase
HIV	Human Immunodeficiency Virus
Kb	Kilobase
LB	Luria Bertani Broth
MAPK	Mitogen-activated Protein Kinase
Mb	Megebase
mRNA	Messenger Robonucleic Acid
ORF	Open reading Frame
PABA	Para-aminobenzoic Acid
PAC	P1 Artificial Chromosome
PCR	Polymerase Chain Reaction
PEG	Polyethylene Glycol
PFGE	Pulsed Field Gel Electrophoresis
PMSF	Phenylmethylsulphonylfluoride
PNK	Polynucleotide Kinase
RFLP	Restriction Fragment Length Polymorphism
RH Map	Radiation Hybrid Map
RIP	Repeat-induced point mutation
rRNA	Ribosomal Ribonucleic Acid
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorhism

SSLP	Simple Sequence Length Polymorphisms
STM	Signature Tagged Mutagenesis
STS	Sequenced Tag Sites
TAP	Tandem Affinity Purification
TAE	Tris-acetate EDTA
TBE	Tris-Borate EDTA
TE	Tris-EDTA
tRNA	Transfer Ribonucleic Acid
U/V	Ultraviolet
VNTR	Variable Number Tandem Repeat
YAC	Yeast Artificial Chromosome

Table of Contents

1. Introduction

The Mycology of pathogenic fungi was largely a science based on observation. This is a laborious process but with the advance of molecular biological techniques this has changed. There are indications that the identification, diagnosis and treatment of fungal infections are relying more on modern molecular systems. This research is aimed at analysing the DNA sequence of the mould *Aspergillus fumigatus*, a clinically important Ascomycete.

Classification of fungi has traditionally been through the particular unique sexual structure of that organism. Therefore, there are many fungi in the Aspergillus group that are related only because of the lack of an identifiable sexual stage. *A. fumigatus* is a filamentous fungus and the hyphae specialise to form conidiophores, which produce the organism's spores, termed conidia. Most fungi grow exclusively as moulds or yeasts thus *Aspergillus* will only grow as hyphae and *Candida glabrata* will only grow as yeast. Fungi can show both morphological forms in different stages of their life cycle and this may be influenced by the medium.

An increasing number of fungi sequencing projects are underway, with several complete such as the sequence for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Projects on *Candida albicans, Neurospora crassa, Candida glabrata, Cryptococcus neoformans, Pneumocystis carinii, Aspergillus fumigatus* and *Magnaportha grisea* are all underway or first drafts have been published. A resource for the fungal genomics can be found in a recent review by Yoder and Turgeon, published in 2001. There is a proposal by Fink with the collaboration of a number of eminent fungal researchers, to sequence the

genome of 15 medically and agriculturally important fungi at the rate of approximately one a month. Known as the FGI, or Fungal Genome Initiative, it will give excellent insight into eukaryotic evolution as well a massive resource on fungal genomes. The FGI realise that the increase of fungal infections is having a very significant impact on human disease numbers, as well as being highly damaging in agriculture. If this proposal is to be funded, then a vast amount of information will be available within two years (http://www-genome.wi.mit.edu/seq/fgi)

There are more than 200,000 identified species of fungi and of these about 200 have pathogenic properties in humans. Pathogenicity of these fungi can be separated into two types: 1) Opportunistic which tend to cause disease in patients that already have abnormal immune systems and cannot fight against the fungal invasion. Due to increased incidences of immunosupression in patients for varying reasons such as AIDS, transplantation, neutropenia and leukaemia, fungi that were never previously seen as pathogenic are being isolated from patients with varying degrees of immunosupression. Opportunistic fungi do not usually cause disease in immunocompetent patients, but there are exceptions. *A. fumigatus* is one of these exceptions. 2) Dimorphic fungi can also cause disease in immunocompetent patients. They exist in two morphological forms. In the wild, they produce infectious conidia and form hyphae. Yet, in the body, with increased temperature, the conidia proliferate as yeast-like structures. The conidia can easily overcome the host's defences. Most dimorphic fungi also have geographical restrictions e.g. *Coccidioides immitis* is found only in southwest America and Mexico. This thesis focuses on the pilot-sequencing project of *Aspergillus fumigatus*, an increasingly important pathogen, especially in immunocompromised patients.

1.1 Project background

Dr David Denning of the University of Manchester put forward a proposal to the Sanger Institute for a pilot project to study the feasibility of sequencing the genome *Aspergillus fumigatus* in1999. This proposal was accepted and I was available to carry out some of the research. Relatively little research had been done on *A. fumigatus* at a genetic level and there were few sequenced genes in the public databases to access and study. At the time of writing, there were 34 *A. fumigatus* sequenced genes in the Swiss-PROT database (http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz), (compared to 478 sequenced genes of its relative, *A. nidulans*). *A.fumigatus* is a significant clinical problem due mainly to the increase in opportunistic fungal infections. Complications arise in HIV infections, post transplant immunosupressant treatment, severe burns and various other conditions that decrease immune response. There are a number of fungi that take advantage of these conditions to invade and cause damage to varying degrees, but there are no other fungi that are as prevalent, virile and damaging as that of *Aspergillus fumigatus*.

1.2 An introduction to *A. fumigatus*

Aspergillus fumigatus is now known to be the most common mould causing invasive infection worldwide. It is a thermophilic fungus and can grow well at over 40°C and up to 50°C on decomposing organic material. It was first described by Micheli in 1729. He described the conidial head of the fungus, with the spore heads around the central

structure, and named it Aspergillus after what he thought resembled aspergillum, the perforated tool for the sprinkling of holy water. Various attempts have been made to classify the whole genus, which is closely related to the *Penicillium* genus. The naming of the first species, *A. flavus*, was in 1809 by Link. There are a large number of the *Aspergilli* that are pathogenic and cause aspergillosis. Aspergillosis is a general term, which covers the infection by any *Aspergillus* species. *A. flavus, A. niger, A. terreus* and *A.nidulans* are all highly prevalent and pathogenic, but it is *A.fumigatus* that is the most predominant in pathogenicity, accounting for over 90% of invasive aspergillosis (Denning, 1998). The most common and effective drugs administered to patients with Aspergillus infection are the azoles. There are significant species differences in response to drug therapies, so species identification is critical for effective treatment (Denning, unpublished data) and any molecular approach to *Aspergillus* clarification will be beneficial.

1.2.1 History of infection

The first infection was described in 1815 by Mayer, who saw the infection in the air-sack of a jackdaw. Later in 1842, Bennett in 1842 described an aspergilloma in man in a hospital in Edinburgh. An aspergilloma is a fungal ball that establishes itself typically in pre-existing cavities in the lungs. There were recorded cases by Rayer (1842) and by Gairdner (1856) of similar infections, but the organism was unidentified. Virchow then described a number of cases of aspergillosis in 1856. Many more cases have been reported between 1890 (Wheaton) and 1947 (Cawley). The first allergic reaction due to

Aspergillus was recorded in London in 1952 (Hinson *et al*,1952) and the first fatal invasive infection of a known immunocompromised patient was recorded in the British Medical Journal in 1953 for a patient found in Gloucester (Rankin, 1953). Appendix 1 is a table of dates that mark out the history of aspergillosis.

1.3 Biology

1.3.1 Reproduction

Asexual sporulation is the most widespread form of reproduction for fungi. Reproduction is carried out by conidiation, a complicated process of temporal and spatial regulation of gene expression, intercellular communication and cell specialization. Detailed studies on the mechanisms of this process have not been carried out on *A. fumigatus*, but have been studied in other fungi. In general terms, the reproductive cycle has three observed stages: 1) A growth phase that allows the cells to acclimatise, adapt and to respond to induction signals; 2) initiation of the development pathway; 3) processes leading to sporulation.

1.3.1.1 Formation and establishment of the colony

Growth in the initial stage is vegetative leading to germination of a spore and subsequently hyphal formation, which grow by apical extension in a polar form. This then forms a network of multiple hyphae called a mycelium. Although seemingly vegetative in nature, it is clear that the hyphae must communicate to produce an ordered

network, where each hypha has a definite role in the uptake of nutrients from its environment. There also has to be some kind of interaction to help time the formation of the various reproductive tools that the fungus will need. Sixteen hours after germination, the centre of the colony starts to form hyphal branches that form conidiophores, the structures that produce the spores. The formation of the hyphae to the formation of the first spore (conidium) takes between 6 and 8 hours, with a total of approximately 24 hours from first spore germination to reinitiation of the asexual cycle. The colony continues to form with the older conidiophores in the centre and newer formations moving to the edge of the colony.

1.3.1.2 Conidiophore formation

This is a complicated process that has several distinct steps. The formation begins with the growth of a stalk that elongates by apical extension (Figure 1).

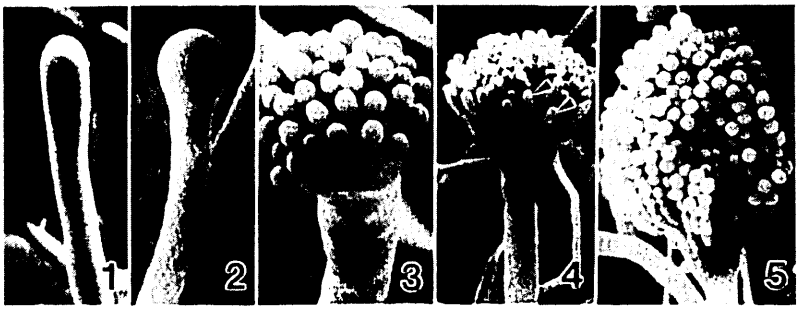


Figure. 1 Scanning electron micrographs of the different stages of conidia formation. 1) Early conidiophore stalk 2) Vesicle formation 3) Metulae development 4) Phialide development 5) Mature conidiophores with chains of conidia (Adams and Wieser, 1999)

This is different to the vegetative hyphae in three ways. There is a thick walled foot cell that anchors the stalk to its growth medium. It is also wider than the vegetative aerial hyphae and the hyphae rarely branch, with the growth length being relatively uniform. This conidiophore stalk then stops growing and the tip swells to a "vesicle" of about 10μm. Nuclei align around this vesicle in large numbers and divide in unison, producing buds at the same time. This layer of buds is called the metula. There are approximately 60 metula on the conidiophore head. The metula then bud again and produce another uninucleate layer and these are known as phialides, which then form the uninucleate spores called conidia. Each phialide forms 100+ spores and generates over 10,000 spores in total on one conidiophore.

Key developmental regulatory pathways in conidiation have been studied extensively in *A. nidulans*. The null mutants of the *brlA* gene are known as "bristle" as these mutants get blocked in early development and do not make it to the polar growth stage of the conidiophore stalk to the swelling of the conidiophore vesicle. The mutants control growth of the conidiophore stalks in an erratic way, where the stalk grows 20-30 times taller than the wild type and the whole colony has a "bristly" appearance. *brlA* is thought to activate expression of development specific genes at the very start of conidiophore vesicle formation, including *abaA* and *wetA* (Marshall and Timberlake, 1991).

Although the processes of function of conidia formation are thought to be similar to those of *A. fumigatus*, the timescales and conditions in *A. fumigatus* may differ. Further study

of the genetic mechanisms and processes will be more freely obtainable by the genome project.

1.3.2 Pathogenesis

Depending on the medium used, *A.fumigatus* spores take between 5 and 12 hours to germinate at 37°C (Ng *et al* 1994). Most species of *Aspergillus* cannot grow at 37°C and this is an important characteristic that separates the pathogenic species from the non-pathogenic species. The growth rates of the various *Aspergillus* species also differ quite markedly, the most rapid grower being *A.fumigatus*. A major marker for the pathogenicity may well be this rapid onset of growth, shown vividly by experiments *in-vitro* with physiological and pharmacological concentrations of hydrocortisone, where *A.fumigatus* has been known to accelerate growth up to 40%. It has a doubling time of 48 mins and a hyphal extension rate of 1-2cm/h in these conditions (Ng *et al* 1994). Another factor in *Aspergillus* pathogenicity may be the small spore size, allowing the spore to penetrate deep into the lungs. The spores are hardy, being able to survive extreme atmospheric conditions. This is thought to be due to the conidial hydrophobic protein coat, which may also protect the spores from host defence mechanisms. The hydrophobic surface allows them to stay airborne, even in damp conditions and they are prevalent in the environment, accounting for 0.3% of spores (Schmitt *et al,* 1991). Surveys have shown that that most humans would inhale several hundred conidia from the environment per day (Chazalet *et al* 1998, Goodley *et al* 1994). *A.fumigatus* has the added factor that it also efficiently binds laminin (Tronchin *et al* 1993) and fibrinogen (Bouchara *et al* 1993)

more strongly than the other species facilitating adhesion of the spores to the airways to allow invasion.

There are other virulence determinants thought to be instrumental in the pathogenicity of *A.fumigatus*. Experimentally, these have shown to be various proteases (Monod *et al* 1995, Markaryan *et al,* 1994 and Reichard *et al,* 1997), ribotoxin (Arruda *et al* 1992, Latge *et al* 1991 and Smith *et al* 1993), phospholipases (Birch *et al* 1996), a hemolysin (Ebina *et al* 1994), gliotoxin (Mullbacher *et al* 1985, Sutton *et al* 1996 and Waring 1990), aflatoxin (Denning 1987), phthioic acid (Birch *et al* 1997) and various other toxins (Frisvad and Samson, 1990 and Moss 1994). *A. fumigatus* also produces at least four phospholipases (Birch *et al* 1996) may have a role in invasive aspergillosis. Certain *Clostridium* and *Pseudomonas* species produce phospholipases and their mechanisms of cell destruction are very similar to that of *A. fumigatus.* Gliotoxin is known to reduce phagocytosis by reducing macrophage and neutrophil numbers (Mullbacher *et al* 1985) and it can promote programmed cell death (Waring 1990). Granuloma formation may be due to the small amounts of phthioic acid produced by *A. fumigatus*, but it is not thought to be a major factor in virulence of the species. The organism itself also seems to be very well protected due to the mannitol (Birkinshaw *et al* 1931, Wong *et al* 1989 and Megson *et al* 1994) and catalases (Hearn *et al* 1992, Takasuka *et al* 1997, Calera *et al* 1997) that it produces. These may well protect the organism from free radicals produced by phagocytes that work in attacking "normal" invaders. The most infamous illness caused by *A. fumigatus* is Farmer's Lung, where the lung becomes infected after repeated exposure to the fungus from ploughing soil and moving hay bales. These symptoms

would also commonly be found in long term mine workers. Patients suffering from cystic fibrosis in rural regions tend to have a higher incidence of *Aspergillus* invasion. However, urban locations are also good environments for the prolification of *A. fumigatus*. Cellars, potted plants and spices (Staib 1984) are all good environments for the fungus as well as unfiltered marijuana smoke, which may well yield large quantities of the organism (Denning, 1998).

The first barrier against infection is the macrophage found in the lungs and nasal cavities. These are capable of destroying the spores. The hyphae are killed by neutrophils, which are dysfunctional in AIDS patients (Schaffner *et al* 1982, Diamond *et al* 1978) and also by monocytes (Roilides *et al* 1994). Hyphae are generally too large to be ingested by the defence mechanisms and killing occurs extracellularly. Histological responses range from a small, granulatomous response to an invasion of necrotised tissue. Neutropenia and, in AIDS patients, dysfunctional neutrophils are also important risks in invasive aspergillosis.

The pathogenesis of the organism is increased by the release of lytic enzymes and toxins. Research has been carried out into what individual factors relate to the pathogenicity of the organism, but so far, little has been discovered. What is important is the dosage of conidia taken in and how it relates to the damage of the defence mechanisms available to the host. In humans, the defence is primarily the removal of the conidia from the lung via ciliar transport and secondarily phagocytosis by the alveolar macrophages. If the primary defence is damaged or fails, then chronic colonisation can occur, such as in cystic fibrosis

(Wojnarowski *et al* 1997). If secondary defences are damaged, then invasive infection can become established if the phagocytic ability of the neutrophilic granulocytes is limited, for example in leukaemia (Schaffner *et al* 1982). Advanced colonisation results in invasion of areas such as blood vessel walls, organ cartilage borders and can lead to haemorrhaging. Aspergilli are among the most mycotoxic species known (Samson 1992). This pathogenicity, as mycotoxins, can be readily ingested with food, especially grain or dairy products. *A. fumigatus* produces a number of toxins, such as gliotoxins, which can act as an immunodepressants (Sutton *et al* 1996), aflatoxins, which are thought to have a role in hepatic cancer in patients with Hepatitis B (Pitt 1994), and ochratoxins, which are thought to have a role in renal failure (after large scale exposure to the fungus) (Di Paolo *et al* 1993).

1.3.3 Epidemiology

A recent study of invasive aspergillosis in a German hospital found that case frequency increased approximately 14-fold from 1978 to 1992 and in the last year of study 4% of the patients had invasive aspergillosis at autopsy (Groll *et al* 1996). It has taken over candidiasis as the most frequent fungal infection detected in patients after death. In Japan, a study of autopsies between 1970 and 1995 showed that the frequency of invasive aspergillosis had increased from 0.4% to 1.4%. The reasons for the increase include the worldwide increase in HIV infection (up to 40 million cases by the year 2000 with a potential estimated number of 1.4 million cases of aspergillosis), particularly in the third world, the development of newer and more sophisticated chemotherapy methods for

malignant tumours and the large increase in worldwide organ transplants (up to 500,000 each year). Linked to this is the large increase in immunosuppressant drugs for transplant anti-rejection and for the treatment of autoimmune diseases such as systemic lupus erythematosus, plus the ineffectiveness of Itraconazole and AmB. The patients at most risk are those with chronic granuloma disease, a defect in the white blood cell (25-40%), lung transplant patients (17-26%), allogeneic bone marrow patients (4-30%), neutropenic leukaemia patients (5-25%), heart transplant patients (2-13%), pancreatic transplant patients (1-4%), kidney transplant patients (~1%) and AIDS, multiple myeloma and combined immunodeficiency patients (~4%) (Denning, 1998). It can be clearly seen that the combat of invasive aspergillus infection is of a high priority to the medical world. The mortality of invasive aspergillosis, if untreated is approximately 85%, and falls to 50% with treatment (Denning, 1996).

Aspergillus also causes a number of other diseases in man. These include aspergilloma (a fungal ball that infects and proliferates in existing lung cavities), sinusitis, allergic bronchopulmonary and sinus infections, keratitis, which can present itself as blindness, and post-operative infections in normal, non-immunocompromised patients. Aspergilloma cases are predicted to rise because of the notorious difficulty of treatment and the widely reported increase in the number of cases of tuberculosis worldwide. There is also an increase in the number of cases of cystic fibrosis, which is also seen to be a further contributor to the increasing numbers of aspergilloma. The fact that 40% of patients with aspergilloma are not expected to live beyond 5 years from diagnosis is also

seen as another factor for the rapid increase of research into this extremely common and virile fungus.

1.3.4 Treatment

Until 2001, there were only two major treatments licensed for Aspergillosis, and these were the antibiotics Amphotericin B (AmB) and Itraconazole. Despite reasonable success of these two drugs in vitro, *in-vivo* success is low and hence the mortality rate from invasive aspergillosis is still high. After 30 years of AmB anti-fungal therapy, it is still not fully understood how the drug works, although it probably binds to membrane sterols, opening transmembrane channels and resulting in increased permeability to monovalent cations (Bolard 1986). It inhibits proton ATPase pumps, decreasing cell energy levels as well as inducing membrane fragility by increasing lipid peroxidation (Brajtburg *et al*, 1985). It is insoluble in water and to make it biologically active, the drug has to be solublised first. Bristol-Myers-Squibb were the first company to try and overcome this problem with Fungizone, the first marketed AmB formulation, wherein it was solublised in the detergent deoxycholate. The drug is highly toxic to the patient (Clements and Peacock, 1990). The side effects are severe, with increased nephrotoxicity, which is increased in the presence of cyclosporin. The AmB does damage the fungal cells, but also damages some of the cells surrounding the infection. This effect may be due preferential binding of the drug to ergosterol, which is found in fungal membranes rather than cholesterol found in mammalian cells. It is also thought that AmB has a role in binding to serum lipoprotein and then being internalised through lipoprotein receptors to promote

toxicity, but this has not been widely studied (Vertut-Doi *et al* 1994 and Wasan *et al* 1993).

The mode of action of the triazole itraconazoles are much better understood to those of AmB. It works by having its free azole nitrogen bind to the catalytic haem iron atom of a cytochrome P-450. Inhibition of the P-450 14α-demethylase prevents production of ergosterol in fungal membranes. This in turn alters membrane fluidity, the relationship with various membrane enzymes and thus an altered synthetic pathway predominates with increased accumulation of phospholipids and unsaturated fatty acids occuring within the fungal cells. Itraconazole only weakly binds to mammalian cytochrome P-450 and therefore the drug is not as toxic as AmB. Despite its reduced toxicity to that of AmB, Itraconazole still has major drawbacks. Firstly, no intravenous preparation is commercially available. This becomes a problem when treating patients that have swallowing difficulties or have impaired bowel absorption. For itraconazole to work effectively, it needs good absorption and distribution within the body. Secondly, the absorption rate varies markedly in different patients. Constant monitoring of the patient with respect to itraconazole concentrations needs to be undertaken. It is more effective in patients who have previously not responded favourably to AmB (Denning 1996, Denning *et al* 1994, Dupont 1990). Thirdly, resistance strains are now beginning to be identified (Denning *et al* 1997). Resistance mechanisms have not yet been fully researched, but it may be mediated by one of a number of mechanisms, such as energy-dependant efflux mechanisms, increased expression of the sterol 14α-demethylase and altered affinity of

the enzyme for the drug (Tobin *et al* 1997). There may be interference with metabolism of other drugs e.g. phenytoin, carbamazepine, via the P-450 system of the patient.

To summarise, anti-*Aspergillus* therapy still remains inadequate. Success of AmB in invasive aspergillosis is about 34% (Denning 1996). Not only that, infection can occur in patients who have been treated with AmB for other unrelated illnesses. This highlights the deficiency of drug treatment against *A. fumigatus* and shows that there is a desperate need for new, innovative therapies to be developed. Hence the *Aspergillus* genome sequencing project is of utmost importance as it should facilitate the development of novel drug therapies.

1.4 Genetics

There are five components to the fungal genome: chromosomal genes, mitochondrial genes, plasmids, mobile genetic elements and fungal virus genes.

1.4.1 Chromosomes and chromosomal genes

There are a number of factors that impinge on accurate chromosomal analysis. Firstly, fungi can alternate between a haploid and a diploid somatic phase. Some yeasts, such as *C. albicans,* are permanently diploid and some fungi are polyploid, for example *P. infestans*. Secondly, the haploid stage is not conducive to microscopic observation. It is impossible to observe chromosomes in their natural state using conventional

microscopy. Thirdly, fungal chromosomes are small, tightly wound and highly condensed and the nuclear membrane remains during most of the mitotic cycle, further reducing the capacity to observe the chromosomes with a conventional microscope. Some fungi have had chromosomal counts performed on them using a combination of techniques such as genetic linkage analysis, pulsed-field electrophoresis and cytology. Table 1 shows the haploid chromosome count for a number of fungi.

Table 1

ORGANISM	CHROMOSOME NUMBER
Basidiomycota	
Coprinus cinereus	13
Puccinia kraussianna	30-40
Shizophyllum commune	11
Chytridiomycota	
A.javanicus	14 (but can vary to polyploidy)
Allomyces arbuscula	16
Oomycota	
Saprolegnia spp.	8-12
Achlya spp	3,6,8
Pythium	Usually 10 or 20
Phytophthora spp	9-10
Ascomycota	
Aspergillus nidulans	8
Neurospora crassa	7
Saccharomyces cerevisiae	17

In *A. fumigatus* the confirmed number of chromosomes is seven. However, in unpublished reports, some laboratories have managed to separate a total of eight chromosomes, the same as the close relative, *A. nidulans.* This may be due to the

chromosomes running as "doublets". The only published report of note regarding *A. fumigatus* having eight chromosomes is by Amaar and Moore (1998), in which a comparison is made using the *niaD* gene and comparing chromosomal position to that of *A. nidulans* by hybridisation. Their estimate is done by chromosome size, not by visualisation; hence confirmed evidence is still not forthcoming.

Fungal nuclear genomes are smaller than other eukaryotes for instance, *Saccharomyces cerevisiae* is 13.5Mb, compared to that of *Drosophila* at 165Mb and the human genome at 3000Mb. One of the reasons is thought to be that there is very little reiterated DNA. It accounts for only 7% of the DNA in *A. nidulans*, and even then most of this codes mainly for cell components, which are required by the organism in large amounts, such as ribosomal RNA, transfer RNA and chromosomal proteins. One notable exception to this is *Bremia lactucae,* which is highly repetitive, where 65% of its 5Mb genome is repetition. *S. cerevisiae* is thought to transcribe as much as 50-60% of its genome. Therefore it can be deduced that fungal genomes have very little non-coding DNA. Moreover, fungal introns are very short and they are usually between 50-200bp compared to that of higher eukaryotes that have introns sizes that can be 10kb or more.

1.4.2 Mitochondrial genes

Fungal mitochondria contain a small, circular molecule of DNA. It is comparable in size to that of other eukaryotes, with a genome size between 19kb and 121kb.

Fungal mitochondrial DNA is known to be important in aging in several filamentous fungi, including *Aspergillus*. A single mutation in a single mitochondrion can be responsible for the premature aging (senescence) of an entire fungal colony. Eventually, the mutant gene will displace the mitochondrial DNA of the wild type (Esser 1990, Bertrand 1995). The fungi *Podospora* can be maintained with the senescence phenotype indefinitely with the right conditions and media. However, the cultures need to be kept young by repeated subculturing, as the cultures will die if they are kept growing continuously for more than 25 days. The mutants are brought about by an extra nuclear "infective factor". Non-senescent strains can "acquire" the ability to senesce when their hyphae anastomose (connect) with strains that already have the ability. It was shown that senescence could be switched off indefinitely by dosing growing strains with sub-lethal amounts of mitochondrial DNA or protein synthesis inhibitors.

1.4.3 Transposons and Plasmids

Plasmids have been found in a number of fungi e.g. *S. cerevisiae*. This plasmid is seen to have a 2μm length when viewed under an electron microscope. It has a length of 6.3kb, up to 100 copies and is found in the nucleus of the organism (Deacon, 1998). The plasmid has no known function, but has been manipulated extensively and forms the basis of many yeast vector systems. Fungal plasmids are usually found in the mitochondria of the organism. *Neurospora crassa* plasmids have been extensively studied (Galagan *et al* 2003). The plasmids themselves are a linear piece of DNA that shows a degree of sequence similarity to mitochondrial DNA. It is thought that these

plasmids, because of the homology, are defective, excised elements of the mitochondrial genes. *N. crassa* also has circular plasmids within the mitochondria. These elements have very little or no homology to that of the mitochondrial genome. Their length can vary between 3 and 5kb and sometimes occurs in tandem to make larger repeat elements. These fungal plasmids have no known function as yet. This is a marked contrast to that of the plasmids found in bacteria, where the majority have a function conferring to antibiotic resistance or pathogenicity determinants (Oliver and Schweizer, 2001).
Transposons are short elements of DNA that integrate within the chromosome and they replicate by producing RNA copies of themselves, which is translated to DNA via transposon encoded reverse transcriptase. These are rare in fungi and have had little research carried out on them.

S. cerevisiae have transposable elements that are involved with switching mating type. This is a process that ensures that there will always be a mixture of two mating types within the population. This mating type switching has not been found in *N. crassa*, so is not necessarily a general feature of ascomycetes.

1.4.4 Fungal genetic variation: Non-sexual variation and Heterokaryosis

Fungi are often haploid and are the only major eukaryotes to be so. Fungal hyphae have an unusual organisation and this tends to make it more difficult to study them in respect to other organisms. Haploid organisms tend to express all of their genes and this in turn brings continual selection pressures, with mutations that occur in any gene swinging the

fitness of the organism either to the detriment or the good of the organism. But another phenomenon, called heterokaryosis, allows fungi to exist with multiple nuclei that are all genetically different in their hyphal cytoplasm. This can arise by mutation in one of the multiple nuclei or from anastomosis. This phenomenon has significance for DNA sequencing and it is necessary to bear this in mind with interpreting the data.

1.4.5 Clinical Manifestations

An outline of the major diseases caused by *A. fumigatus* is shown in appendix 3. It describes the kind of diseases that can occur when *Aspergillus fumigatus* is taken in by the host and the effects it can have upon the host depending upon varying factors such immunocompetence, general health and mode of intake. It is interesting to note that most immunocompetent patients do not develop invasive aspergillosis, but this does not make them immune to the effects of the fungus. Immunocompetent patients can contract an invasive *aspergillus* related disease, but this is a much more rare occurrence. Because some of these diseases overlap in their effects, it is difficult to divide them into the immunocompromised and immunocompetent variants.

1.5 Genomics, Comparative Genomics and Post genomics

Genomics and comparative genomics is a growing science, which is going to prove invaluable in the fungal arena for gene analysis, the development of new drugs and interpretation of pathogenicity. This section outlines the main molecular background to

mapping and the methods used in the project. It additionally reviews some of the current methodology in comparing sequences and gene expression.

1.5.1 Mapping

Genome mapping is a tool that helps the determination of relative positions of genes on a DNA molecule. There two broad categories of genome mapping; genetic mapping and physical mapping. Both of these methods give a good indication of the likely order of genes along a chromosome and are used to "guide" the scientist to a certain gene. Physical maps show an accurate estimate of the distances between genes along a chromosome. A physical map will help the scientist "home in" on a particular gene of interest far more easily than the information from genetic mapping alone.

1.5.1.1 Genetic mapping

Genetic maps rely solely on genetic markers. "Marker" is a very broad term for describing any observable variation that is a result of mutation or alteration at a specified genetic locus. Commonly used genetic markers include 1) Restriction length polymorphisms (RFLPs), VNTRs, or variable number of tandem repeat polymorphisms, Microsatellite polymorphisms and single nucleotide polymorphisms (SNPs).

Genetic Linkage Analysis is a statistical based analysis, which infers likely crossover patterns of genes and therefore the order of the markers involved. The markers are often

seen to co-segregate with certain genes and it is this information that is used to infer gene order from the linkage analysis. Genetic maps are also used to form the scaffold that is required to help build more detailed genome maps. These detailed "physical" maps further clarify the DNA sequence between genetic markers, and are essential to the rapid identification of genes.

1.5.1.2 Physical Mapping

There are three main forms of physical maps; chromosomal maps, radiation hybrid maps (RH maps) and sequence maps. Each form has a varying degree of resolution. The lowest resolution of these is chromosomal mapping. RH and sequence maps are much more detailed. Breaks are induced using radiation to determine the distance between two markers. DNA is exposed to a calculated dose of radiation. This type of mapping is able to isolate almost any genetic marker, as well as other genomic fragments, to a highly accurate map position. Its usefulness is especially significant in positioning markers in regions where highly polymorphic genetic markers are scarce. RH mapping has also been used as a bridge between linkage maps and sequence maps.

Sequence mapping, or sequence tagged site (STS) mapping relies on the phenomenon of short sequences of DNA that are shown to be unique and the location of this unique site must already be identified. The unique site must be nowhere else on that particular chromosome or genome if the fragment happens to cover the entire genome. Expressed sequence tags, or ESTs, are short sequences of DNA that are identified by the analysis of

complementary, or cDNA clones. ESTs can be used as a marker only if the STS comes from a unique gene and not from a gene family member where the genes all have similar sequences. Simple Sequence Length Polymorphisms (SSLPs) are arrays of repeat sequences that vary in length. SSLPs that are polymorphic and have previously been mapped by linkage analysis are highly regarded as they provide a good link between genetic and physical maps. Random genomic sequences can be used either by sequencing random pieces of cloned genomic DNA, or by looking at and determining the use of sequence already deposited in a database.

To use STSs for mapping, the DNA is cut into smaller fragments using a restriction enzyme and this breaks down overlapping DNA fragments from a region of DNA of interest. To form the map from this, it must be known which fragments contain which STSs. To do this, copies of the DNA fragments are made using standard molecular cloning to form a library.

These copies are pieced back together in the correct order by looking at clones that contain overlapping DNA fragments via end sequencing and computer methods to identify overlaps. This is what is known as a contig. The clones are archived in freezers and the sequence data is stored on computers and used as starting information for generating continuous DNA sequence. The STSs serve to anchor the sequence onto a physical map.

STS-based mapping does have limitations. There may be gaps in clone coverage or clones may become lost or mapped to a wrong position. DNA fragments can break. Deletion of DNA fragments can also occur during the replication process. Clones that

have two distinctly different regions of DNA can be replicated. This would give DNA segments that are widely separated in the genome being mistakenly mapped to adjacent positions. There is also the risk of the cloned DNA getting contaminated with host DNA. To try and lessen the effects of these problems improve the mapping accuracy, comparison and integration of STS-based physical maps with genetic, radiation hybrid, and cytogenetic maps is frequently undertaken. This comparison, or "cross-referencing" helps enhance the usefulness of a given map, confirms the STS order, and helps orientation when building contigs. The comparison of genetic and physical maps can be very laborious, so computers and dedicated software align the maps and speed up the process.

The mapping for the *Aspergillus* project has been mainly to set out the framework from which the genome sequencing could commence. There has been input from both genetic and physical mapping but the physical mapping has taken priority, as initially a particular gene was the starting point. Once found, the information could be used to continue to find the next piece of the puzzle. Radiation hybrid mapping was used to find the *niaD* gene and a combination of some of the techniques, or variations of them, to continue the project.

1.5.2 Annotation

Gene annotation takes raw DNA sequence and adds useful information so that the downstream user can interpret the information. The annotator will always strive to be as correct and accurate as possible, but even with careful observation, it is extremely difficult to get everything 100% correct, so constant checks with other annotators and

other interpretations would be made continuously to maintain the high standards. Genome annotation falls broadly into two sections. The first is the prediction of the position and start site of all of the genes and the second is to discern the function of the gene products.

In a given genome, a number of software programs can predict the majority of genes in a genome with a relatively high degree of accuracy and consistency. However, the nature of a genome will always throw up regions within the genome where gene prediction is particularly difficult and the programs will therefore never become 100% efficient. It might be good for a program to be accurate to 97% and well understood in other working areas, but for a genome scientist, it is usually the other 3% of unknown quantity that hold the regions of interest. Therefore the outputs of the prediction programs still need to be closely looked at and the results interpreted according to the investigator's knowledge. Table 3 shows a few of the programs that are used for gene prediction.

Table 3. Gene prediction programs

Program	Source	Reference
Glimmer	http://www.tigr.org/softlab/glimmer/glimmer.html	Delcher *et al* (1999)
Orpheus	http://pedant.gsf.de/orpheus/	Frishman *et al* (1998)
GeneMark	http://opal.biology.gatech.edu/GeneMark/	Borodovsky and McIninch (1993), Lukashin and Borodovsky (1998)
GeneMark.hmm		
CRITICA	http://geta.life.uiuc.edu/~gary/programs/CRITICA.html	Badger and Olsen (1999)

(Parkhill, 2002)

Gene prediction is concerned with the determination of the open reading frames (ORFs) that begin with a start codon and ending in an in-frame stop codon, using software tools to predict these from the DNA sequence. What is essential with gene prediction is deciding which of the ORFs actually are coding regions of sequence. Also, it is obvious that all the genes in an organism will not be identical. There is a high likelihood that some sets of genes within a given genome will vary from the normal rules that those genes may adhere to in other organisms for example, genes that encode small basic ribosomal proteins or largely hydrophobic proteins or repetitive proteins. The gene finding programs will not always see these genes, as they do not always conform to the model of the genes within that organism. There are also other reasons, such as horizontal transfer of genes in recent evolution in a genome and there may be non-typical genes such as RNA genes, DNA repeats and psuedogenes that all contribute to gene prediction programs either not seeing genes or over-predicting genes in non-coding sequences. These all have to be removed or edited manually.

Following prediction, genes are given putative roles using databases such as FASTA (Pearson and Lipman, 1988) and BLAST (Alstchul *et al* 1990). It is likely the databases will give slightly different hits because of the differing calculation methods of the databases and it is therefore useful to run both searches to determine best hits according to the context of the sequence. The database searches are not 100% accurate or infallible though and therefore this step still requires further human intervention to ensure that there is a high degree of accuracy to function assignment. Comparison with biologically characterised function is always much more accurate than comparison with previous annotation as this can lead to mistakes and chains of wrongly annotated genes. To avoid

problems of wrongly assigned functions, there is a further step of looking at protein families or motifs. This is done by identifying sequences or patterns that are common to all members of a particular protein family and then this information can be used to identify new member of that family. This can be a more accurate way of assigning function or structure as proteins usually have a modular structure and this can sometimes confuse the similarity based searches like BLAST and FASTA as different domains can match different sets of database sequences. PROSITE (Bairoch, 1991) looks for motifs and residue patterns in a protein. Short functional motifs such as enzyme active sites as easily found by this program, but although the method is highly specific, the program lacks the ability to identify larger and more diffuse features such as entire protein domains and has recently been modified to include profile information (Falquet *et al* 2002). Further refinements of functional assignment do exist. The most useful of these is orthology and paralogy. These give very good insights into the mechanisms of functional transfer of information from one genome to another. Orthologues are equivalent genes in two organisms directly descended from the same gene in the two organisms direct ancestor. Paralogues are genes related through a gene duplication event in either the parent organism or in one of the descendants. Orthologues are usually seen to serve the same function yet paralogues have either usually perform the same function yet lead to functional redundancy or have diverged to perform a different function or act on different substrates. Therefore, if orthologues and paralogues of a given functionally characterised gene can be identified, the functional description of the characterised gene can be transferred to the novel gene (Figure. 2)

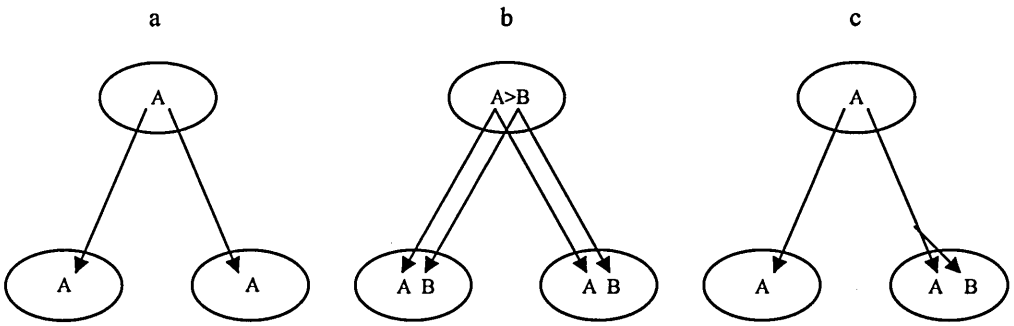


Figure. 2 The ovals at the top represent the ancestral species and the bottom ovals represent the descendant species. A and B represent given genes. A and A are orthologues in the descendant species and A and B are paralogues in the descendant species. In (a), this illustrates a simple descent. (b) shows gene duplication of A and B in ancestral species and (c) shows a gene duplication of A to B in one lineage only.

(Parkhill, 2002)

Looking for what is called reciprocal best matches between gene sets computationally helps the identification of orthologues. A gene is compared from one genome against all of the genes in a second genome. If a best match is found then this best match is then compared back against all the other genes in the original genome. If the second match identifies the original match from the first comparison, then it is known as the reciprocal best match. This can only be done to identify orthologues in two complete genomes of the organisms if the comparison is known. But this method is not always free from mistakes and some paralogous genes can be identified as orthologues. Careful manual checking has to be done at this process.

Orthology can be checked by looking at synteny in related genomes. Synteny is the conserved gene order and orientation in a genome. Orthologues will tend to be found in areas of high conservation and in the same context as other orthologues in compared genomes. Any differences found needs great care in assignment of orthology.
Similarity based methods such as those just described can assign function to a gene putatively with a relatively high degree of accuracy, but it is only in the lab that a definite function can be ascribed to a particular gene.

It is imperative that an annotator is aware of what they are working on and should be vigilant of any potential problems that may arise. The information that is gleaned from the sequence is going to be new and novel and therefore the annotator must have a good biological understanding of both the bioinformatic analysis and the biological processes within the organism. The programs and techniques are good and they facilitate the tedious groundwork, but the full understanding and interpretation of the genomic information is the purview of the annotator.

1.5.3 Comparative Genomics

The use of comparative genomics will be an essential tool in analysing the sequence generated by the *A. fumigatus* project, especially the data from the closely related *A. nidulans* species. To date, little data have been released from the *A. nidulans* project, although the sequencing is virtually complete. A target date of Spring 2003 has been set by the Whitehead Institute to release annotation data

(http://www-genome.wi.mit.edu/annotation/fungi/aspergillus/).

A number of important organisms, both related and unrelated to *A. fumigatus* have recently been sequenced and analysed and comparative genomic techniques applied to the findings. Here, I will review some of the more recent, relevant and more important aspects of current comparative analysis.

1.5.3.1 Characteristics in *Neurospora crassa* sequence

Neurospora crassa paved the way for modern genetics in the early 20th century and the sequence and annotation data was published by the Whitehead Institute in 2003 (Galagan *et al* 2003). It too, like the *Aspergilli*, is a multicellular filamentous fungus that has helped us understand concepts such as genome defence mechanisms, DNA methylation, circadian rhythms, DNA repair, post-transcriptional gene silencing and mitochondrial gene importation. It is also a good model for other eukaryotic organisms.

The assembled genome is approximately 40Mb in length. A total of 10,082 protein-coding genes were predicted, with 9,200 longer than 100 amino acids in length. This is approximately twice the number as predicted in *S. pombe* and almost as many as the *D. melanogaster*. The genes were found to cover over 44% of the genome and a gene could be found approximately every 3.7kb. Gene length was found to be slightly longer than that of *S. pombe* (1.67kb and 1.4kb respectively). Table 4 shows that *Neurospora* has a greater number of introns than *S. pombe* in its genes and this accounts for the longer gene length (on average, 1.7 introns per gene, each averaging at 134 nucleotides).

Feature	Value
General	
Size (bp)	38,639,769
Chromosomes	7
G+C content (%)	50
Protein-coding genes	10,082
Protein-coding genes >100 amino acids	9,200
tRNA genes	424
5S rRNA genes	74
% coding	44
% intronic	6
Average gene size (bp)	1,673 (481 Amino Acids)
Average intergenic distances (bp)	1,953
Predicted protein-coding sequences	
Identified by similarity to known sequences	1,336 (13%)
Conserved hypothetical proteins	4,606 (46%)
Predicted proteins (no similarity to known sequences)	4,140 (41%)

Table 4. *Neurospora crassa* genome features (Galagan *et al* 2003)

4,140 (41%) of *Neurospora* proteins do not match known proteins from public databases. 5,805 (57%) of the proteins have no match to any genes in that of either *S. pombe* or *S. cerevisiae*. Only 14% (1,421) genes match either plant or animal genes. This indicates that its biology is not likely to be homologous to yeasts, but may have some homology with other filamentous fungi and higher eukaryotes.

Additionally, studies on repeat sequences over 200bp long with 65% similarity showed that 10% of the genome consists of repeat sequences (Kelkar *et al* 2001). This emphasises a phenomenon that is unique to fungi called Repeat-induced point mutation

(RIP). This is a process that was originally detected in *Neurospora crassa* and is a highly efficient genome defence mechanism. Occurring during the haploid stage of the *Neurospora* sexual cycle, the method can detect and mutate both copies of a sequence duplication. Within the duplicated sequence, CG to TA is mutated at a 30% rate during the sexual cycle. It has been proposed as a mechanism for defence against mobile or selfish DNA. When the sequence was analysed, it was seen that 81% of the repetitive sequence in *Neurospora* had been mutated by RIP. Repetitions of over 400bp are most susceptible to RIP, with over 97% of the genomic repeats being RIP mutated. The hypothesis that RIP acts a defence mechanism is supported by the fact that there were no intact mobile elements identified in the sequence and 46% of repetitive nucleotides could be identified as mobile element relics.

Neurospora photobiology has been studied for a number of years because blue light has major implications in developmental circadian rhythms. The genome sequence has shown that there are a number of sequences that show similarity to blue light sensing genes and also homologous with *Aspergillus nidulans* velvet gene, which has been implicated in both red and blue light responses. This was unexpected, as no red light photobiology has been described for *Neurospora crassa* before. *Arabidopsis* has recently been shown to have phytochromes that associate with cryptochromes that have a role in blue light sensing and signalling (Devlin and Kay, 2001), therefore the two phytochromes and the velvet homologue may have a role in regulation of this aspect of *Neurospora* photobiology.

MAPK (Mitogen-activated protein kinase) pathways integrate signals from multiple receptor pathways including two-component signalling systems. This mechanism is conserved, as the nine MAPK proteins identified in *Neurospora* are also found in both *S. pombe* and *S. cerevisiae.* In addition, *Neurospora* has a complement of 11 histidine kinases, compared to one in *S. cerevisiae* and three in *S. pombe.* A third of these genes are similar to proteins found in *A. fumigatus* and *A. nidulans* that affect conidiation (unpublished data). The functions of the rest of the genes are unknown, although seven contain PAS/PAC domains, which would implicate them in light and oxygen responses. The number of histidine kinases is thought to show that they have a larger role than thought and shows that filamentous fungi are more similar to plants than animals, where two-component systems are less abundant.

The sequence has also shown that the pathway for macroconidiation differs in both *Neurospora* and *A. nidulans*. Components of the pathways have been found in both species and some signalling proteins are known to be conserved upstream. But in contrast, there is little conservation downstream between the two fungi. *Neurospora* requires the *acon-2, acon-3, fld* and *fl* genes for conidiation. *A. nidulans* requires the FlbC, FlbD, BrlA, AbaA and WetA gene products to perform the same function. The *Neurospora* sequence has not revealed any FlvC, BrlA or AbaA homologues, with only a weak similarity in one protein that has 100 amino acids of the carboxy terminus of the WetA gene. This shows that the mechanics of similar functions of the two fungi are very different.

Although *Neurospora* is a saprotroph, the sequence contained a number of genes required for plant pathogenesis, identified in other fungal pathogens. It contains a number of genes that code for a wide range of extracellular enzymes that digest plant material, although there is no observed cutinase homologue, which is the enzyme that breaks down cutin found in many other plant pathogens. It also contains a number of cytochrome P450 enzymes that help in the detoxification of plant anti-fungal compounds. The sequence also shows that all the known signal transduction components used in ascomycete pathogenesis are present in *Neurospora* even though *Neurospora* is not known to be a pathogen.

1.5.3.2 Comparative Genomic Analysis of *Vibrio cholerae*

In 2002, Dziejman *et al* published an analysis of *Vibrio cholerae*, the causative bacteria of the disease cholera. The majority of their studies used the technique of microarraying to carry out the comparisons of a number of strains relating to endemic and pandemic outbreaks of the disease. The study was originally set up to analyse the evolution of *V.cholerae* as a human pathogen and environmental organism. The complete sequence of the "El Tor" O1 Strain N16961 from the 7^{th} pandemic was used as a starting point for comparison. Another serogroup, O139, shares locality with the O1 strain in many areas and are thought to be closely related, with the O139 evolving from the O1 strain using various molecular typing methods such as ribotyping, RFLP analysis and sequence analysis and comparison of homologous housekeeping genes. The El Tor strains can be

distinguished from "classical" strains that have caused previous pandemics by their haemolytic properties and agglutination with erythrocytes, as well as antibiotic resistance. The comparison of the El Tor strain was done against a classical strain, pre-pandemic El Tor strain, another pandemic El Tor strain and two non-toxigenic strains to find unique genes in N16961.

The analyses found a high degree of genetic similarity between the strains that had been isolated over the past century, but some genes were found to be unique in the El Tor and O139 strains. It is these unique genes that may have contributed to the success of these two strains.

The array analyses identified 143 genes that were absent from the eight of the nine test strains compared to N16961 (Table 5).

Table 5.

Strain	Origin	Year Isolated	Biotype, serogroup	No. genes absent
N16961	Bangladesh	1971	El Tor, O1 (7th Pandemic)	N/A
2740-80	Gulf Coast, US	1980	Environmental, non-toxigenic El Tor O1	49
NCTC 8457	Saudi Arabia	1910	El Tor O1 (Pre-pandemic)	39
MAK 757	Celebes Islands	1937	El Tor O1 (pre-Pandemic)	49
569B	India	1948	Classical, O1	46
O395	India	1965	Classical, O1	36
NIH 41	India	1940	Classical, O1	48
HK1	Hong Kong	1961	El Tor, O1 (7th Pandemic)	0
C6709	Peru	1991	El Tor, O1 (7th Pandemic)	1
MO10	India	1992	El Tor, O139	47

(From Dziejman *et al* 2001)

Strain HK1 seemed to be identical to N16961. C6709 was missing only 1 gene. The remaining seven strains all were missing between 36 and 49 genes, an approximately 1% difference between the genomes of the test strains and N16961. In comparison to studies done on *S.aureus* and *Helicobacter pylori*, this is a marked difference. Approximately 12% of the genes were missing within different strains of *S.aureus* and 6% with *H. pylori*. This suggests high genomic conservation between strains over the past century, a remarkable feat. The identified genes fell into 4 main groups (Table 6).

Table 6.

Group	Strains where genes are absent	No. of genes absent*
Group I	O395	7
Genes present in El Tor but not classical	569B	7
strains	NIH41	7
Group II	2740-80	14
Genes present only in strains able to cause	NCTC8457	14
epidemic disease (absent from	MAK757	2
environmental and pandemic El Tor)		
Group III	2740-80	22
Genes present only in 7th pandemic strains	NCTC8457	22
HK1, C6709 and MO10	MAK757	22
	O395	22
	569B	22
	NIH41	22
Group IV	MO10	42
Genes uniquely absent from a single strain	NIH41	14
	MAK757	15

* Genes identified by array, PCR and/or Southern analysis

(from Dzeijman *et al* 2001)

Some genes could differentiate the classical biotype strains from those of the El Tor biotype strains. The genes would be in all strains except classical. The microarray technique would not allow the researchers to identify the genes only in classical strains as the microarray itself was constructed from N16961. The "non-classical" strains were all El Tor strains of diverse origin and all contained a pathogenicity island coding for the

TCP pili of the CTX bacteriophage that encodes the cholera toxin in the bacterium. Therefore, it could be deduced that they are all potentially pathogenic.
Genes were identified in only pandemic strains, including the classical strains (group 2). Other genes that were found to be specific to the 7th pandemic strains, including epidemic and endemic El Tor O1 and closely related O139 strain MO10.

1.5.3.3 The genome sequence of *Plasmodium falciparum*

The human malaria parasite *Plasmodium falciparum* is responsible for hundreds of millions of cases of malaria and kills over 2,700,000 people worldwide every year. Despite many years of research, it is becoming more and more prevalent in tropical and sub-tropical regions.

In 2002, a consortium consisting mainly of the Institute for Genome Research in the USA and The Sanger Institute published the genome sequence and its subsequent genomic findings (Gardner *et al* 2002). Here I will give a brief overview of the findings made.

Using whole genome shotgun technology, the sequence was assembled using YACs to assist contig ordering and gap closure and using sequenced tagged sites (STSs), microsatellite markers and HAPPY mapping to orientate the contigs. A very high A+T content made the gap closure very difficult as this makes the genome much more unstable and making it less conducive to standard sequencing techniques.

The genome consists of 22.8Mb along 14 chromosomes that range in size from 0.6Mb to 3.3Mb. This is nearly twice the size *Schizosaccharomyces pombe.* The A+T content is approximately 80% and was found to rise to over 90% in introns and intergenic regions. Various computer programs predicted protein encoding genes and these were then manually curated. There were approximately 5,300 protein encoding genes identified, which was approximately the same as that found in *S. pombe* (Table 7). This infers a gene density of one gene per 4,338bp. Introns were predicted in 54% of *P. falciparum*, which equates to being similar to that found in both *S. pombe* and *Dictyostelium discoideum*, but much higher than that found in *Saccharomyces cerevisiae*, which has only 5% of genes containing introns. The mean length of the gene, excluding introns, was 2.3Kb, much higher than that found in other organisms, which are usually between 1.3 and 1.6Kb. There were also a much higher proportion of genes that were over 4Kb in length (15.5%). *S. pombe* and *S. cerevisiae* have 3.0% and 3.6% respectively. Many of these larger genes encode for uncharacterised proteins that are thought to be cytosolic proteins as there are no recognisable signal peptides. There were also no transposable elements identified, so the reason for this increased length of these genes still a mystery.

Table 7.

Feature	*P. falciparum*	*S. pombe*	*S. cerevisiae*	*D. discoideum+*	*A.thaliana*
Size (bp)	22,852,764	12,462,637	12,495,682	8,100,000	115,409,949
G+C content (%)	19.4	36.0	38.3	22.2	34.9
No. of genes	5,268*	4,929	5,770	2,799	25,498
Mean gene length (bp)	2,283	1,426	1,424	1,626	1,310

Gene density (bp per gene)	4,338	2,528	2,088	2,600	4,526
% coding	52.6	57.5	70.5	56.3	28.8
Genes with introns (%)	53.9	43.0	5.0	68.0	79.0
Exons					
Number	12,674	ND	ND	6,389	132,982
No. per gene	2.39	ND	NA	2.29	5.18
G+C content (%)	23.7	39.6	28.0	28.0	ND
Mean length (bp)	949	ND	ND	711	170
Total length (bp)	12,028,250	ND	ND	4,548,978	33,249,250
Introns					
Number	7,406	4,730	272	3,587	107,784
G+C content (%)	13.5	ND	NA	13.0	ND
Mean length (%)	178.7	81.0	NA	177	170
Total length (bp)	1,323,509	383,130	ND	643,899	18,055,421
Intergenic regions					
G+C content (%)	13.6	ND	ND	14.0	ND
Mean length (bp)	1,694	952	515	786	ND
RNAs					
No. of tRNA genes	43	174	ND	73	ND
No. of 5S rRNA genes	3	30	ND	NA	ND
No. of 5.8S, 18S and 28S rRNA units	7	200-400	ND	NA	700-800

ND, not determined, NA, not applicable

*70% of genes matched ESTs or encoded proteins detected by proteomic analyses

+ "No. of genes" is figure for Chr 2 only. Genome still yet incomplete, therefore some figures extrapolated for whole genome. Other genomes stated here are complete.

(Gardner *et al*, 2002)

Over half of the genes predicted (52%) were detected in samples taken from several stages of the life cycle of the parasite. 49% of these genes with 97% identity over 100bp overlapped with expressed sequence tags (ESTs) which were also taken from different cell cycle stages. At the time of their writing, the group had not completed proteomic and EST studies to a depth enough to back some data up, but it could be assumed that the annotation process identified large proportions of most genes. But to fully predict correctly, the identification of the 5' ends of genes and genes with small exons is very difficult without the support of EST and protein evidence. It was suggested by the consortium that further studies into ESTs and full length complementary DNA sequences would be required to develop more intuitive training sets for gene finding programs to verify predicted genes.

60% of the 5,268 predicted proteins (3,208 hypothetical proteins) did not have the required similarity to proteins in other organisms to assign functions. It was therefore proposed that nearly two thirds of the proteins are unique in the organism, which is much higher than found in previously sequenced eukaryotes. This may be down to the greater evolutionary distance between *Plasmodium* and other sequenced eukaryotes. This is magnified by the much higher A+T content of the genome, which will decrease the likelihood of sequence similarity in other organisms. 5% of the proteins (257) had significant similarity to hypothetical proteins in other organisms. 1,631 (31%) predicted proteins contained one or more transmembrane protein domain and there were putative signal peptide and signal anchors found in 911 (17.3%) of the predicted genes.

The Gene Ontology (GO) database (Ashburner *et al* 2000) was used to describe the roles of genes and gene products in the *P. falciparum* genome. GO terms were assigned to 2,134 gene products (40%). A comparison from the paper between *P. falciparum* and *S. cerevisiae* can be seen in Figure 3. There are higher values for nearly all categories in *S. cerevisiae* than in *P. falciparum*, which is due to the larger proportion of the genome being characterised. Two exceptions to this are related to the parasite cell life cycle. It was found that at least 1.3% of the genes in *P. falciparum* are involved in cell-to-cell adhesion or invasion of the host cells. There are 208 (3.9%) genes in the genome that are known to be involved in host immune system evasion. This is indicated by the number of gene products designated "physiological processes" in *P. falciparum* compared to that of *S. cerevisiae.*

Highlighting the reasonable lack of knowledge of *P. falciparum* compared to other organisms such as *S. cerevisiae* is the number of areas that are under-represented in the *P. falciparum* genome. For instance, the GO terminology had very few assignments in the areas of sporulation and cell budding, which are included in "other cell growth and/or maintenance", but this was to be expected. But it was also found that very few genes in *P. falciparum* were designated in "cell organisation and biogenesis", the "cell cycle" or "transcription factor" compared to that of *S. cerevisiae.*

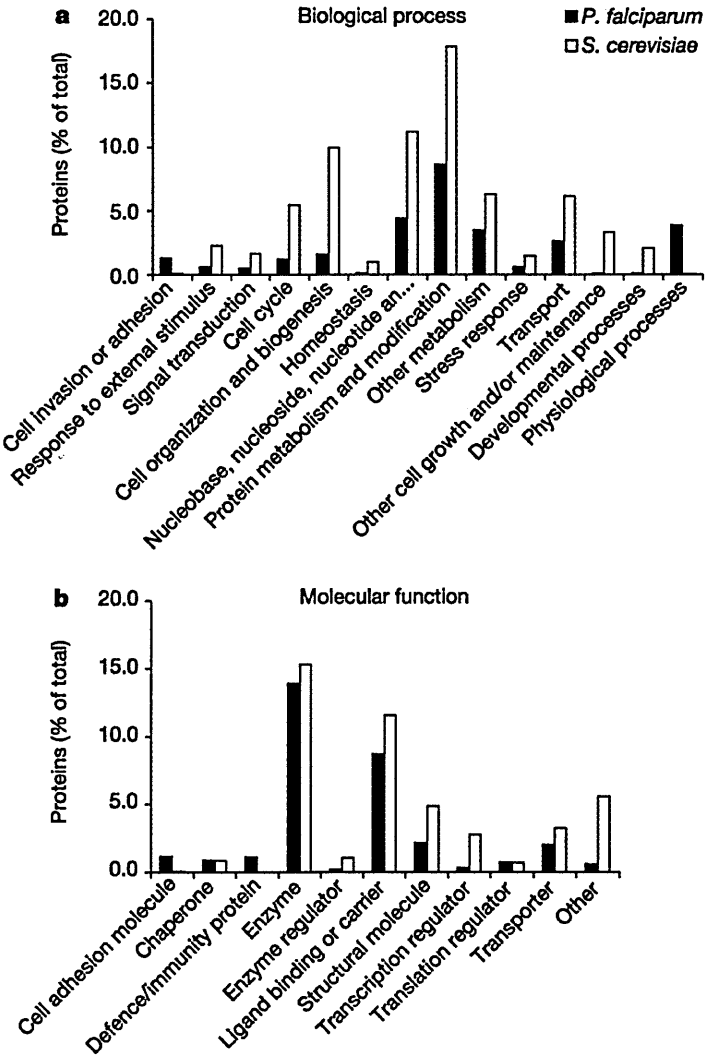


Figure. 3 Gene Ontology classifications. Classification of *P. falciparum* proteins according to the "biological process" (a) and "molecular function" (b) ontologies of the Gene Ontology system. (Gardner *et al*, 2002)

When subjected to comparative genome analysis with other sequenced eukaryotes, the research showed that in overall genome content, *P. falciparum* is most closely related to *Arabidopsis thaliana*. But the affinity that seems apparent between *Arabidopsis* and *Plasmodium* may not show the true phylogenetic history of the *P. falciparum* lineage.

Any vaccine produced from these studies must induce a protective immune response that is better than those provided naturally. Up to the point of the publication date, approximately 30 antigens had been identified using conventional techniques and some have been tested in clinical trials, with one giving partial protection in field settings. The genome publication will hopefully stimulate research into hundreds more potential antigens which could be scanned for properties such as surface expression or limited antigenic diversity. Combined with stage-specific expression data from microarrays and proteomics, potential antigens from one or more stages of the life cycle could be identified. But for the genome to have an impact on vaccine development, high throughput immunological assays that can identify novel candidate vaccine antigens that are targets of protective humoral and cellular immune responses in humans needs to be developed. Also, new methods in to maximise the longevity and quality of the immune response will have to be developed to produce effective malaria vaccines. These kind of studies will be carried out at the completion of the whole genome sequencing project, but the pilot project may give a small glimpse of what to expect. A more in depth analysis of the findings of this sort are discussed further in chapter 4.

1.5.3.4 Genomic approaches to fungal pathogenicity (and related studies)

The previous sections provide examples to illustrate that genomics will be invaluable in understanding pathogenicity as conventional genetic and biochemical methods are very limited, especially in pathogenic fungi.

Lorenz published a review in 2002 (Lorenz 2002) outlining the techniques that are used to characterise pathogenicity traits in fungi. Mutagenic techniques have been staple techniques for many years in genomic studies, with the isolation and analysis of mutants providing good information in many projects. Genomic approaches make identification of mutated loci and the creation of mutant pools much easier. Signature tagged mutagenesis (STM) for bacterial pathogens was developed by Holden (Hensel *et al*, 1995) and was used for *Aspergillus fumigatus*. Holden tested 4648 STM strains in a mouse model of aspergillosis and found that two insertions were reproducibly defective. One of these was upstream of the *pabaA* gene, which encodes for *para*-aminobenzoic acid (PABA) synthase. There were no known *in vitro* phenotypes but a growth defect was detected in the absence of PABA. Moreover, it was markedly avirulent unless the mice were fed PABA supplements. The presence of PABA had never before been associated with virulence in fungi.

Microarray analysis has been used successfully in *Vibrio cholorae* as described in 1.6.3.2. In *C. albicans*, exposure to itraconazole was used to induce expression changes in 296 genes on a slide from Incyte Genomics containing 6,600 Open Reading Frames (ORFs), ESTs and genomic fragments. The upregulated genes included most of the ergosterol biosynthesis pathway, which is a known target for triazole compounds.

A consortium, led by the laboratory of Braun published data (Braun *et al* 2001) on the construction and use of a 2000 gene filter based array. They used the array to study morphogenesis by profiling *C. albicans* strains in the TUP1 and NRG1 genes. ScTUP1 is

known to be a transcriptional repressor targeted to specific genes by ScNRG1. Mutations is *tup1* and *nrg1* code for filamentation. With this, CaTUP1 and CaNRG1 co-regulate a number of genes involved in yeast-hypha transition, including *HWP1*, *ECE1*, *ALS3* and *ALS8*. They also have promoter regions similar to those found to bind to NRG1 *in vitro*. CaTUP1 is also targeted to promoters by CaMIG1 protein, which co-regulate genes unrelated to morphogenesis.

A two-hybrid system, developed by Fields (1989) was used to good effect to study protein-protein interactions. Requiring in frame fusions with ORFs, hundreds of thousands of clones were screened for the small fungal genome. Uetz *et al* (2000) systematically cloned 6000 ORFs into both fusion constructs. Transformants were pooled and 962 interacting pairs were found. Alternatively, they arrayed 6000 strains, each expressing one fusion and screened against 192 selected "bait" proteins, which identified 281 pairs. It was simpler, but interference was higher; only 20% of initial interactions were confirmed. This needs to be optimised, but it shows that the genomic approaches have advantages.

Snyder *et al* designed protein arrays with 5800 yeast proteins, expressed in yeast and purified as glutathione-S-transferase (GST) fusions. Control experiments using an anti-GST antibody found 93.5% of the spots were detectable. The array was then probed with a biotinylated calmodulin which identified six known and 33 unknown calmodulin-binding proteins. The array was also used to identify liposome-binding proteins that contain various phosphotidylinositol compounds. It was also used to study kinase

specificity (Zhu *et al* 2000). As this shows, it would be easy to adapt this to pathogenic fungi, to find host factor binding proteins for example.

Pathogen information will be improved by the systematic examination of every gene in a pathogenic organism to uncover the overall expression patterns. Microarrays could also provide information on drug responses in those organisms. Gray (1998) used microarrays to measure the effects of kinase inhibitors on the whole of the yeast genome. This was done by measuring changes in mRNA levels before and after treatment.

Genomics will bypass the traditional challenges in fungal pathogens and present a good opportunity to look at virulence. Some technologies have not yet been applied to pathogenic fungi, as there are still a lot of fungal pathogens to be sequenced. When this occurs, we will see rapid and productive applications of genomics in this field.

With respect to transcriptomics and proteomics, Gavin *et al* (2002) performed a large-scale study of the multiprotein complexes in *Saccharomyces cerevisiae* using Tandem-affinity purification (TAP) and mass spectrometry. 1739 genes were tagged. 1143 of these had human orthologues. There were also 589 purified protein assemblies. The analysis of the assembly data showed 232 different multiprotein complexes and from those showed that 344 proteins had new, previously unknown cellular roles. From those, 231 had no previous functional annotation in the databases. There was seen to be conservation across species between human and yeast complexes, from single proteins to their molecular environment.

Patterns that could identify genetic regulatory networks or gene groups that are similar have been demonstrated on human cell lines and yeast (Tavozoie *et al* 1999, Spellman *et al* 1998). The Spellman group found 800 *S. cerevisiae* genes whose transcripts alternate through one peak per cell cycle. The 800 genes were identified using objective, empirical models of cell cycle regulation, but with an arbitrary threshold. Below this threshold, it was not known if there were genes whose expression was truly periodic or the periodicity might even have biological significance.

1.6 In conclusion

Aspergillus fumigatus is fast becoming a very important, medically significant organism. It is unusual in that it is an opportunistic pathogen. It can affect both immunocompromised and normal hosts in a variety of ways, from skin lesions to eye infections to full blown lethal lung and central nervous system invasion. Because of the highly opportunistic nature of *A. fumigatus*, immunocompromised host groups are at very high risk of infection. As the worldwide incidence of HIV/AIDS, organ transplants and other immunocompromising illnesses increase, the combat against *A. fumigatus* becomes more and more important. *A. fumigatus* infection is now the most common mould causing infection worldwide and is the most prevalent and virile of the known aspergillosis causing aspergilli. It is extremely resistant to virtually all known anti-fungal drugs, with many of them having little or no effect. Of the drugs that do have limited effect, these usually have severe and highly unpleasant side effects for the host.

So far, little study on the genetic mechanisms of *A. fumigatus* has been carried out. In fact, little is known about its genetic make-up at all. So far, there are only a small number of genes in the public databases that can be accessed. This is compared to *A. nidulans*, its closest relative, where there are over 400 genes in the databases. It is still not known what the definitive chromosome number is for *A. fumigatus*.

This study was initiated to determine the feasibility of sequencing the entire *A. fumigatus* genome on a small and manageable scale. The proposal was to sequence 1Mb and if this was successful, then the sequence would be annotated to determine what information the sequence carried. The thesis describes this sequencing, the strategies used and the background to the choice of vectors and technology. It is broadly divided into three parts: BAC library construction, physical mapping and annotation. Once this was complete and was determined then the whole genome sequence could be undertaken, with the process scaled up to achieve this. The sequence could be used in comparative and post genomic studies to determine factors such as the mechanisms to its pathogenicity and virulence, as well as its other biological mechanisms to give a much greater understanding of how the fungi works at a genetic level. This information would hopefully lead to more successful drug therapies and development of more specific host and organism-targeted drugs.

Genomics and post genomics are going to play a large part in the understanding of *A. fumigatus*. As in the examples of other organisms such as *Plasmodium falciparum*, as well as related organisms such as the yeasts *S. cerevisiae* and *S. pombe*, the genomic approach can determine large amounts of very useful information about the organism.

At the time of writing, the whole genome sequencing of *Aspergillus fumigatus* is well under way.

2. BAC library construction

2.1 Introduction

2.1.1 History of Vector development

Until the mid 1980's, genomic research relied on approaches that used DNA vectors such as plasmids, cosmids and bacteriophage λ. Stable, large inserts could not be generated. Cosmids were developed in 1978 by Collins and Hohn. These are plasmid/bacteriophage hybrids (ColE1) that carry cohesive ends (cos) of λ and can be packaged in vitro, using a system developed by Hohn and Murray (1977). They can be manipulated to accept inserts of 35-45kb and as they are multi copy, a high yield of DNA can be generated. However, the vector itself cannot carry inserts over the stated size the cosmid needs to be packaged into the Bacteriophage λ head with the physical size of the insert being the limiting factor.

In 1979, Sternberg published data on the development of the P1 vector. This worked in similar way to that of the cosmid, but could be packaged into a much larger Bacteriophage P1 head increasing the possible insert size from the 45kb of a cosmid to

100kb in the P1. P1 also has the added advantage that it was a single copy vector. This factor inherently gave the DNA much better stability than was found in cosmids. Cosmids and P1 became the vectors of choice for a number of years. This was because of their ease of use and no special laboratory requirements were needed. They also efficiently produced high yields of DNA. But genome research became more intense and larger regions of DNA from more complex organisms were of interest to scientists. The limitations of insert stability and size became more and more apparent.

In 1987, Burke *et al* developed the Yeast Artificial Chromosome (YAC). It promised the possibility of being able to routinely clone inserts of up to 1MB and could change the face of genome mapping and analysis. It was envisaged that even the largest genomes could be analysed with fewer clones covering regions of interest, genome wide. Whole genes or gene clusters could be studied without the need for manipulation to reconstruct the region of interest.

Unfortunately, as work geared up to make full use of this new cloning vector, it was seen that there were some serious drawbacks with this method. The construction of a YAC library was very labour intensive and complicated. New techniques had to be developed and learned, followed by protracted periods of careful production. This would involve duplicating procedures repeatedly to generate a large enough number of clones to create a library to cover a genome. YAC clones were therefore seen as expensive, time consuming and difficult to master. Only a few laboratories were able to invest in the technique. Moreover, YAC clones had high levels of chimerism and insert rearrangement (Burke 1990; Neil *et al* 1990; Green *et al* 1991; Venter *et al* 1996). These clones were thus considered unusable for sequencing large genomes and with too much effort targeted

into eliminating chimeras and rearranged inserts. What was required was a vector that would allow easy and efficient cloning of insert sizes that were big enough to allow relatively deep coverage of genomes and gene clusters, as well as the ability of the clones to be easily manipulated and stored. This would allow the distribution of clones and libraries to other laboratories and thus in turn widen the study and approach by the scientific community.

In the early 1990s three vectors were developed. Firstly, Kim *et al* (1992) developed the Fosmid. This was a modified form of a cosmid based on the F factor replicon of *E.*coli. It is a low copy vector and because of this it has inherent stability. One major advantage was the generation of libraries of certain DNA that previously had been unclonable in *E. coli* because of high copy number. Fosmids allowed the maintenance and propagation of unstable and unclonable genomic segments and allowed the construction of libraries with a more comprehensive representation of the genomes under study. But the disadvantage was that the insert size was not much larger than that of a cosmid. The development of the Bacterial Artificial Chromosome (BAC) (Shizuya *et al*,1992) from the fosmid and then the P1 Artificial Chromosome (PAC) (Ioannou *et al*,1994) has revolutionised cloning.

The vector of choice for large-scale genome analysis and mapping is the Bacterial Artificial Chromosome (BAC) (Cai *et al.* 1998). The basic structure of most BAC vectors is not strictly artificial but is based on the *E. coli* F factor. This incorporates some essential genes that provide stability and are also involved with copy number.

An important reason for using a BAC, as opposed to using P1 Artificial Chromosome (PAC) is the actual size of the vector. BAC cloning has superseded PAC cloning because

PAC vectors are often around 16kb long, whereas BAC vectors are usually no more than 8-9kb. Working with BACs therefore becomes much easier. Also, pBACe3.6 contains several unique cleavage and modification sites that other BAC vectors do not have.

The F-factor in *E. coli* is a ~100kb plasmid which codes for more than 60 proteins that are involved in replication, partition and conjugation. The F-factor is usually present in a closed circular, double stranded form, with maybe 1-2 copies per cell. It can insert randomly into 30 or more sites on the *E. coli* chromosome.

BAC vectors now utilise only minimal sequence needed for autonomous replication, partitioning of the plasmid and copy number control. The genes that code for the proteins to do this are derived from the F-factor and are listed as thus: *oriS,* is the origin of replication and is unidirectional; *repE* encodes for a protein called RepFIA protein E that is autoregulatory and is required in replication from *oriS*; *parA, parB* and *parC* are required for ParFIA partitioning and *parB* and *parC* are required for compatibility with other F factors.

2.1.2 Technical considerations for library production

The development of BAC and PAC vectors mentioned in 2.1 allowed the rapid analysis of large genomic regions, complete genes and provided the tools for large-scale genome analysis. Whole libraries with very deep coverage (greater than x10) of the human genome could be gridded out onto Nylon filter membranes and used in hybridisation experiments providing unlimited shelf life for storage. BACs allow insert sizes of up to 300kb, with most libraries regularly having average insert sizes of 80-200kb. Although

insert sizes are much bigger than that of cosmids, the inserts are still nowhere near the sizes of those cloned into YACs. However, BACs are easier to manipulate in culture and bacterial cultures tend to grow more rapidly than yeast. They are also easier to grid and handle. The growth is denser, which in turn allows better screening. But probably the most important factor is that pure DNA from bacterial clones can be extracted in high yield using simple plasmid preparation techniques.

The most important aspect of a genome study is to simplify the logistics of the project to save time and cost. The proportion of the genome represented in the library needs to be maximised. This can be done by increasing the overall depth of the library. Increasing the library size significantly increases the chance of creating contiguous pathways of overlapping clones. Whole genomic regions can be spanned in this way. It also maximises the chance of finding a clone that has a locus of interest within it. Determining overlap of clones is important in three ways. It allows correct orientation with respect to other clones, as they would be in the genome. It also helps identify the path of clones across a genomic region of interest and provides proof that the clone belongs to the genome of study. The larger the library size, the greater the chance of a sizeable overlap. It might be preferable to be able to make libraries as large as possible, but there are other constraints that have to be considered in financial and practical terms.

Secondly, creating a library with larger average insert sizes means fewer clones overall to cover a region under study. Insert size is an important part of genome analysis as it determines the genome coverage in the library. If markers are known in the genome at 100kb spacing, there is no point having a library of insert sizes of 50kb as it will rarely link two markers together. Other methods such as walking or fingerprinting would be

needed to complete the genome "map". Whereas a library with insert sizes of 150kb will allow contiguous overlapping of clones. But there are inherent problems with this approach. If the sizes of the clones themselves are too large then this could cause problems with the computer assembly.

Ideally when a library is constructed, the total amount of DNA in the library, represented by summing the sizes of all the inserts, is usually a factor bigger than the total amount of DNA in the whole genome of the organism of study. The ratio of DNA in a library to that of the DNA in the genome of study is known as the depth of coverage, or redundancy. There are a number of factors in determining the depth of coverage and quality of a library when constructing a library for the first time: genome size, desired number of clones, vector only clones or empty wells and the average insert size. Genome coverage of a library can be calculated using the following formula:

c (genome coverage) = (total number of clones-non contributing clones)x average insert size/haploid genome size.

The coverage value gives an indication of the number of clones that contain a gene or locus of interest that will be in the library. But it is only an indication. Some genomes contain loci that are very difficult to clone (high number of repeats, chimeric DNA etc). These regions may not be represented in the library. There may also be certain areas in the genome that are not represented well purely because of the uneven spread of clones across the library. Some regions may be over represented in the library. If it is assumed that the library is random, probabilities of loci being present in the library follows a

Poisson distribution if the coverage is between 0.5 and 10x. It also has to be assumed that the genome size is 100 or more times larger than the average insert size (see Table 8)

Table 8.

Probability of having one or more clones/loci with a library as a function of library size.

Library Size (genome coverage)	*P* (%)
0.5	39.3
1	63.2
2	86.4
3	95.0
4	98.2
5	99.3
6	99.75
7	99.91
8	99.97
9	99.99
10	99.995

The probability of finding x clones from a library of c coverage is calculated as:

$P(x) = C^x/x! \times e^{-c}$

(Dunham *et al*, 1997)

Although it is assumed with these calculations that the cloning has occurred randomly, the calculation is actually an overestimation. Some genomic regions are not equally well adapted as other regions for cloning. From the large number of factors that create cloning bias, two are most prominent. Firstly, the sites that restriction enzymes work on are not

evenly spread across the genome. Secondly, certain regions in genomic DNA are lethal to the *E. coli* host. To counteract this, library construction must have a depth that is larger than that predicted by the Poisson distribution. If a genome is small or if only a few clones per loci are required, then a 10x coverage library would be of good use, but would be economically unfounded. But if a walking or mapping strategy is to be utilised, where each clone needs to be verified, then there are going to be regions where there will be a greater probability of there being no clones present for the next step. So constructing larger libraries with greater depth will increase the probability that contigs can be contiguously built over large areas of the genome, but can also allow a much higher flexibility in selecting clones that will provide a minimal tiling path for sequencing (i.e. the coverage that takes the least number of clones to cover the region). Therefore, with these considerations in mind, the number of clones required in a library can be calculated.

When constructing libraries that have large depth of coverage, a large number of clones need to be generated and handled. Because of this, the logistics of handling, storage and manipulation become a lot more complicated. The table below shows numbers of clones required for a 7.5x coverage of organisms of differing genome sizes (see Table 9)

Table 9.

Haploid genome size of organism (Mb)	Average insert size of clones (Kb)	Library size	
		no. of clones	no. of 384 well plates
20	50	3000	8
1000	100	7500	20
3000	50	450,000	1172
3000	150	150,000	391

(Dunham *et al*, 1997)

This illustrates that there are a number of factors in combination that determine how large a library should be, clearly shown by the discrepancy between the genomes that are both 3000Mb long, yet with a larger insert, the number of plates to handle is vastly reduced.

2.1.3 The Vector

The vector that was used for the *Aspergillus fumigatus* library was pBACe3.6. It has a number of other features within its genotype that make it highly conducive to accepting large inserts of DNA and maintaining them with a high degree of stability. The wildtype *loxP* site is retained as well as including an additional mutant loxP511 site. *loxP* stands

for "locus of crossing over of P1 phage" and is involved with the Cre-*loxP* recombination. Cre is the enzyme that catalyses the reaction. An animated illustration of how this mechanism works can be seen at (http://www.clontech.com/products/families/creator/popups/creloxanimation.shtml). The mutation *loxP511* is a recombination proficient mutation with regards to other mutations, but is less proficient than the wild type *loxP*. Incorporated into the *loxP511* mutation is the PI-SceI site. PI-SceI is a "homing endonuclease" usually found in eukaryotes, prokaryotes and archaebacteria and is coded for by a kind of parasitic DNA element that is highly proficient at replicating itself. PI-SceI is produced by an autocatalytic protein splicing reaction from a precursor protein and works enzymatically on the cleavage of the DNA at a defined position. By doing this, it initiates the recombination of its own gene into a genome that does not have it (Frengen *et al* 1999). Close to this site is a Tn7att site, which is present to aid downstream manipulation such as gene knockouts. It retrofits DNA by utilising the specificity of the Tn7 within the genome and without affecting the insert in any way.

Positive selection for inserts containing BAC clones is achieved through inclusion of the sacBII gene from Sternberg's P1 vector (Pierce et al. 1992,). Possession of this gene is lethal to *E. coli* when it is grown on sucrose containing medium. This is because the gene codes for levansucrase, which converts sucrose to levan, which is highly toxic to the cells. If an insert disrupts this gene, then the cell will grow, as the gene has been disrupted and no levansucrase is expressed.

The *sac*B selection that is built into the genotype of the vector helps this enormously.

As shown in figure 4, the vector has a split mutli-cloning site either side of the stuffer fragment. This allows the insertion of inserts that are cut with a number of restriction enzymes. pBACe3.6 will accept inserts cut with *Bam*HI, *Sac*II, *Eco*RI, *Sac*I and *Mlu*I. This versatility allows the investigator much more freedom in the enzymes that they use. A full review of this vector can be read in the paper by Frengen *et al* (1999) and the full sequence can be found in GenBank Accession No: U80929.

The parental form of the vector has a 2.7kb pUC plasmid, or "stuffer fragment" inserted into the sacBII gene. The "stuffer fragment" is removed in the preparation of the vector. The stuffer fragment is inserted into the sacBII so to leave the cloning site within the gene when the stuffer fragment is removed. If an insert does not incorporate, then the gene is fully functional and therefore becomes lethal to the cell in the presence of sucrose. The "stuffer fragment" is there for another reason: it increases the copy number of the vector to ensure large quantities of DNA during propagation and preparation of the vector. It is preferable to have as few colonies that do not contain inserts as possible.

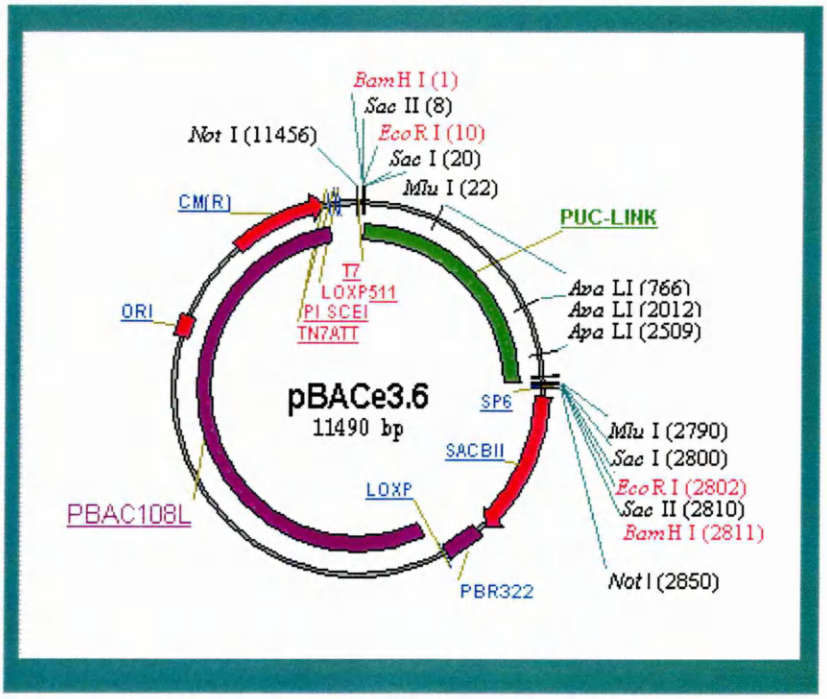


Figure 4. A diagrammatic map of pBACe3.6 (http://bacpac.chori.org/vectorsdet.htm)

2.1.4 DNA Source

The cells (Af293) that were used for this project are pathogenic and they come from a well-maintained cell line that was isolated from the lung of a 57 year old woman patient from Shrewsbury in 1993 and cultured at the Public Health Laboratory of The Royal Shrewsbury Hospital, Shrewsbury, England. High purity DNA was important to eliminate any contamination that could hamper the ligation process. Any small, spurious DNA fragments will compete with the larger, required inserts and bring the overall insert size and coverage down. Other contaminants may also inhibit enzymatic activity. For the pilot project, it was initially envisaged that 1MB would be sequenced since we were

working from a set starting target with an unknown genomic position. Therefore a library needed to be generated from the whole genome to include all regions of the genome. The integrity of the DNA was important, as this would make the job of creating large inserts easier. Low quality DNA in smaller fragments would make it difficult to control the size of the fragments during digestion and bias the library towards smaller inserts. There needed to be enough DNA to work with, so concentration was also important. Both a high and low concentration of DNA tends to bias the library towards smaller inserts but can allow a high enough concentration for the ligations to work, as we found out. For BAC library construction, the DNA is prepared in agarose plugs, allowing isolation of very large fragments of DNA that can be manipulated in situ following restriction enzyme digestion and Pulsed Field Gel Electrophoresis. Fragments hundreds of kilobases in size can be generated and isolated. The DNA can be "cleaned" by a "pre-electrophoresis" step which removes the cellular debris, as well as smaller fragments of DNA and some RNA that may be present and hinder the downstream processes used to generate the library.

2.1.5 Bacterial Host Strain

DH10B was selected, due to the presence of the *deoR* gene. *deoR* and *CytR* regulate the *deo* operon which is involved in the uptake and catabolism of nucleosides and deoxyribonucleosides. The operon regulates the cluster to allow the cell wall to "open" much more readily and accept large pieces of DNA. Without this mutation, plasmids >10kb are only transformed with very low efficiency. Some DNA loci are difficult to

clone or maintain. Whilst *A. fumigatus* has little repetitive DNA it is advantageous to use DH10B cells that can cope with DNA instability.

To prevent degradation of DNA with high levels of methylated cytosine and adenine, DH10B cells have mutations and deletions in the genes coding for the restriction systems McrA, McrB (which both restrict DNA containing methylated cytosine residues) and Mrr (which restricts DNA containing methylated adenine residues). The *mcrA* gene has been mutated and the *mrr-hsdRMS-mcrBC* region has been deleted entirely. The *mcrA* gene encodes for proteins that are important in the restriction systems. Not having these genes allows DH10B to accept foreign methylated DNA much more readily. Fungal DNA is not highly methylated, with *Neurospora crassa* having no more than 2-3% of its cytosines methylated. It can also be assumed that this is the case for *A. fumigatus*, although this is not clear, as little investigation has been done in this area. Therefore, DH10B cells were used to prevent any library bias if *A. fumigatus* DNA turned out to have a higher degree of methylation than expected.

DH10B cells have other characteristics that make them much more amenable to accepting and propagating foreign DNA; *recA1,* which increases the stability of the inserts by involvement in homologous recombination; *hsd* encodes for a protein subunit that has specificity to a number of proteins (if mutation occurs in this gene, the cell is unable to carry out methylation and mutation); *endA1* improves the quality of the DNA from mini-preps by keeping it stable. The strain also contains a Φ80d*lacZ*ΔM15 marker that provides a α-compliment to the β-galactosidase gene from pUC or similar vectors for

blue/white selection. With a combination of all of these factors, DH10B was our host strain of choice for the BAC library.

2.2 Materials and Methods

The construction of the BAC library for *A. fumigatus* took almost a year to perfect. The experimentation and methodology was based closely on that set out in the methods of Osoegawa *et al* (1998).

2.2.1 BAC vector preparation

In order to ensure that the vector preparation was sufficient for library preparation and high quality ligation, the following methods were used:

- A caesium chloride gradient to remove and purify the vector DNA from the cells it is supplied in (section 2.2.1.1)
- removal of the "stuffer" fragment to open up the cloning site, preparation of λ DNA for use as a control for the ligations (section 2.2.1.2)
- a cloning site check to ensure that the site is of a good integrity and allow ligations to occur (section 2.2.1.3)

2.2.1.1 Caesium chloride gradient

Materials and Equipment

- Sorval RB5C centrifuge with GSA or SS34 rotor

- Mistral 3000 centrifuge
- Beckman Optimal TL Ultracentrifuge with TL100 rotor
- Bench top microfuge capable of 13,000rpm
- Minigel equipment and power pack
- Chloramphenicol (20μg/ml in Ethanol) (Sigma)
- GTE (see appendix 2)
- LB Broth (see appendix 2)
- Sodium Dodecyl Sulphate/Sodium Hydroxide (SDS/NaOH) (see appendix)
- Isopropanol
- 3M Sodium Acetate
- TE (10:10) (see appendix 2)
- T0.1E (see appendix 2)
- Ethidium Bromide (Sigma) stock solution: 10mg/ml ethidium bromide in TE 10:10

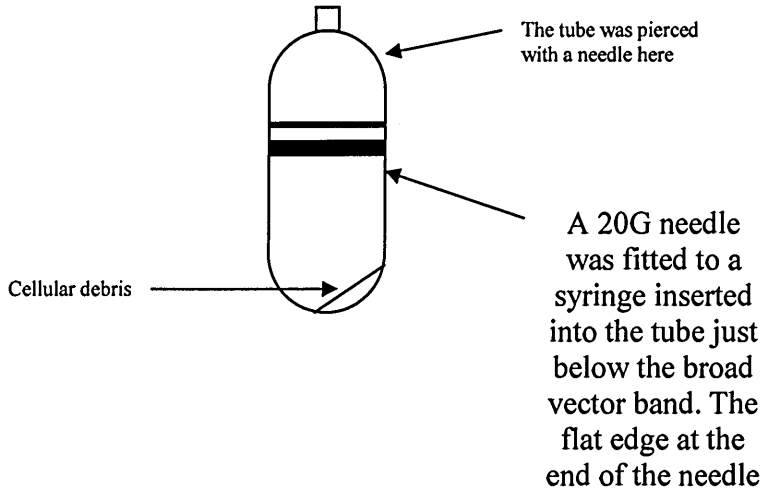
The vector was supplied in DH10B cells by Pieter de Jong (Children's Hospital Oakland Research Institute, USA) (http://bacpac.chori.org/pbace36.htm).

CsCl gradient purification is still the most efficient way of removing the host chromosomal DNA from the solution, leaving only the purified vector to be recovered and is described below.

1. *E. coli* cells containing the vector were spread from an agar stab onto an agar plate containing 20μg/ml chloramphenicol and grown at 37°C overnight.
2. A single colony was inoculated into 200ml of LB broth containing 20μg/ml chloramphenicol (x4) and grown 37°C overnight.
3. Cultures (4x200ml) were harvested in 250ml tubes and put into a GSA rotor and spun at 5000rpm for 10mins at 4°C.
4. Each pellet was then washed in 10ml of GTE. The suspensions were then combined and re-spun to pellet the cells.
5. The cells were then re-suspended in 10ml GTE and left at room temperature for 10mins.
6. The suspension was split (2x5ml) into Oakridge centrifuge tubes and 10ml of SDS/NaOH was added to each. The tubes were then mixed gently by inverting slowly and left at room temperature for 10mins.
7. Sodium Acetate (7.5ml) was added to each tube. The tubes were mixed and left on ice for 30mins.
8. The cell debris was pelleted by centrifuging the tubes at 10,000rpm for 20mins at 4°C in a SS34 rotor.
9. The supernatants were poured off gently and put into clean tubes. Isopropanol (12ml) was added and left for 15mins at room temperature.
10. The nucleic acid was pelleted by spinning at 15,000rpm for 30mins.
11. The supernatant was decanted off and the pellets washed in 70% ethanol.
12. The ethanol was decanted and the pellets left to air dry.
13. The pellets were resuspended in 2x4.6ml TE (10:10) for 2 hours.

14. Into two 50 ml Falcon tubes, 4.78g of Caesium Chloride (CsCl) were weighed out. The DNA solution was then added to each tube and the tubes swirled gently to dissolve the CsCl.
15. To each tube, 462µl of ethidium bromide stock solution was added.
16. The tubes were spun at 3000rpm at RT for 12mins to pellet the debris.
17. A 16G needle was attached to a 5ml syringe. The supernatant was recovered using the syringe and the contents emptied into a 5.1ml Beckman ultracentrifuge tube. The meniscus had to reach the base of the spout of the tube.
18. A metal sealer was placed on the top of the tube and a heated Beckman adaptor tool was placed on the top of this. With pressure and heat, the tube would seal by melting down the collar of the neck. Once the neck was sealed, a cooling tool would be placed on top of the metal sealer for a few seconds. All tools should be either provided with the centrifuge or the tubes.
19. The tubes were then places into a rotor with spacers and the rotor was placed into an ultracentrifuge.
20. The tubes were then spun at 70,000rpm for 24 hours at 20°C.
21. When the spin was finished, the tubes were carefully removed from the rotor and placed in a rack.
22. The tubes were then illuminated with long wavelength U/V light. This would show up the bands that contained the vector and the cellular debris. Figure 5 describes how the vector was removed:

Figure 5.



23. The removed vector was then placed back into a fresh, clean Beckman tube and the liquid topped up with stock CsCl and ethidium bromide/TE solution and the tubes sealed again.
24. The tubes were then spun overnight at 20°C and 70,000rpm.
25. The DNA was removed as before, but then split between two eppendorf tubes. There should be approximately 2x600µl.
26. An equal volume of isobutanol was added to each tube. The tubes were then vortexed and then spun in a microfuge for 1min at 13,000rpm. This was repeated by multiple isobutanol extraction until both layers colourless.
27. The lower layer containing the vector DNA was removed with a pipette and placed into a dialysis bag.
28. The bag was dialysed against 250ml T0.1E at 4°C, changing the TE every hour 5 times.

29. The vector DNA was recovered from the bag and placed into an eppendorf tube.
30. Added to this was 1/10 volume of sodium acetate then 2 volumes of ethanol.
31. This was then incubated overnight at -20°C.
32. The tube was then spun at 13,000rpm in a microfuge for 30mins to pellet the DNA. The supernatant was removed and the pellet allowed to dry.
33. The DNA was then resuspended in 120µl T0.1E.
34. A small amount of this was then run on a agarose minigel (0.8%) to check the yield.

2.2.1.2 Stuffer fragment removal

Materials and Equipment

- Calf Intestinal Alkaline Phosphatase (CIAP) (Boehringer Mannheim, 0.03U/µg)
- *Bam*HI restriction enzyme (New England Biolabs, 20U/µl)
- *Bam*HI enzyme specific buffer
- 10xBovine Serum Albumin (New England Biolabs, 10mg/ml)
- 0.1M EDTA
- Proteinase K (Roche, 14mg/ml)
- TBE of various concentrations (see appendix 2)
- High Gelling Temperature Agarose (Invitrogen)
- Ficoll loading dye SeaKem GTG Agarose (BMA Products)
- Low Range Pulsed Field Gel Marker (New England Biolabs, 25µg/ml)
- ¾ Dialysis tubing (Invitrogen)
- Polyethylene Glycol (PEG) 8000 (Sigma)

- CHEF gel apparatus (see appendix 2)

1. A reaction was set up to remove the 2.8kb pUC stuffer fragment that was inserted in the vector. The reaction was set up in bulk as follows:

CIAP (final concentration 0.03U/μg DNA)	7.5μl
DNA (5xD)	20μl
10xBamH1 Buffer	20μl
DDW	135μl
BamH1	5μl
10xBSA	20μl

2. This was incubated at 37°C for 1 hour. The reaction was stopped by adding 26μl of 0.1M EDTA and 5μl of 14mg/ml Proteinase K and incubated at 37°C for 1 hour.
3. A 1μl amount was run on a small check gel, 1xTBE to determine if digestion had worked.
4. The whole vector digest was then loaded with loading dye into a 1% CHEF gel in 0.5xTBE and run at 14°C for 16 hours, 6V/cm and a 0.1 to 40secs switching time. Also loaded onto the gel, in the side lanes was a Low Range Pulsed Field Gel Marker. The gel was removed from the apparatus and the flanking marker lanes were removed, taking a small, but minimal amount of the vector portion. These lanes were then stained in Ethidium Bromide for 15-20 mins to detect the position

of the vector and stuff fragment. The lanes were visualised under U/V light and small "nicks" were made where the position of the vector band was. The lanes were then reassembled with the main body of the gel containing the unstained vector and the gel slices cut from the main portion, using the flanking, stained portions as a guideline.

5. The gel slice was then placed into a ¾ inch dialysis tubing and the tubing was clipped at one end. The "bag" was then filled with sterile 0.5xTBE and then clipped at the other end to seal. The bag was then placed into a small electrophoresis tank filled with 0.5xTBE and electroeluted at 3V/cm for 3 hours. The current was reversed for 45 secs at the end of the run to detach any DNA from the walls of the dialysis bag.
6. The bag still containing the gel slice was dialysed over night at 4°C against 0.5xTE. The solution was then concentrated down to approximately 600µl by dialysing in 0.5xTE (containing 30% PEG). The vector was recovered and concentrated by ethanol precipitation. After centrifugation the pellet was resuspended in 100µl of T0.1E.

2.2.1.3 λ DNA for use as a control in ligations

Materials and Equipment

- λ DNA (New England Biolabs, 250µg)
- *Bam*HI restriction enzyme (20U/µl)
- *Bam*HI restriction enzyme buffer
- EDTA

- Proteinase K (14mg/ml)
- T0.1E
- CHEF gel apparatus (see appendix 2)

To check the vector and ligation conditions a positive control insert was prepared from bacteriophage λ. λ DNA is 48,502bp in length and a *Bam*HI digestion produces fragments of 5.5kb (x2), 6.5kb, 6.8kb,7.2kb and 17.3kb. It was the largest of these fragments that was purified and used as the positive control. The digestion was set up as follows:

λ DNA (500ng/μl)	10μl
10x *Bam*HI buffer	10μl
DDW	75μl
*Bam*HI	5μl

1. This was incubated at 37°C for 3 hours and the reaction was stopped by adding 13μl of 0.1M EDTA and 2.5μl of 14mg/ml Proteinase K and incubated at 37°C for 1 hour. This was run on a CHEF gel and the 17.3kb top band removed and extracted as described in section 2.2.1.2.
2. The λ DNA was resuspended in 60μl T0.1E. This could be then used as the positive control insert DNA.

2.2.1.4 Cloning site check

Materials and Equipment

- Polyethylene Glycol (PEG) 8000
- 1M $MgCl_2$
- T4 DNA ligase (Roche, 5U/µl)
- T4 DNA ligase buffer
- T4 polynucletide kinase (Roche, 5U/µl)
- 0.5M EDTA
- Proteinase K (Roche, 14mg/ml)

The cloning site needed to be examined to ensure that the dephosphorylation had been efficient and the cloning site had been undamaged. The DNA had previously been treated with Calf Intestinal Alkaline Phosphatase (CIAP) (described in section 2.2.1.2) to remove the phosphate group from the 5' ends of the DNA and prevent it from re-ligating. Kinase treatment replaces the phosphate and allows the linear molecule to re-ligate. When run on a gel, linear and circular DNA have different migration properties, so is a good indicator of whether or not the vector has been prepared correctly and is usable.

1. The test was done by setting up two ligation reactions, one with T4 Polynucleotide Kinase (PNK) (Roche 5U/µl) and one without. The two ligations were set up as follows:

Vector DNA	0.5µl	Vector DNA	0.5µl
20% (w/v) PEG 8000	5.0µl	20% (w/v) PEG 8000	5.0µl
1M $MgCl_2$	0.1µl	1M $MgCl_2$	0.1µl
10x ligation buffer	2.0µl	10x ligation buffer	2.0µl
T4 Ligase	0.2µl	T4 Ligase	0.2µl
T4 PNK	0.1µl	DDW	12.2µl
DDW	12.1µl		

The ligations were incubated at 37°C for 2 hours.

2. EDTA (2µl of 0.5M) and 0.5µl of Proteinase K were added and incubated at 37°C for 15mins to stop the reaction
3. A small amount (1µl) was checked on a 0.8% agarose gel to determine whether or not most of the ends could still be ligated depending on kinase treatment (Larin *et al* 1996).

2.2.1.5 Self ligation

Materials and Equipment

As the previous section, but without the T4 Polynucleotide Kinase

- Bench top microfuge capable of 13,000rpm
- 1M Sodium Chloride
- 1:1 Phenol/Chlorophorm (Applied Biosystems)
- 96% Ethanol

- Loading mix (see appendix 2)

1. Once it was determined that the ends could be ligated, a bulk ligation was set up to remove a minor, yet significant background of non-dephosphorylated vector ends. This was set up as follows:

Vector	50µl
20% (w/v) PEG 8000	100µl
0.1M $MgCl_2$	100µl
10x ligation buffer	40µl
T4 Ligase	20µl
DDW	90µl

This was incubated overnight at 16°C.

2. NaCl (50µl) was added and the ligation extracted with an equal volume of phenol/chloroform.
3. After centrifugation at 13,000rpm in a microfuge for 1min to separate the layers, the top layer containing the vector was removed and placed into another eppendorf tube.
4. Ethanol (1ml) was then added and incubated at either -70°C for 2 hours or at -20°C overnight. The tube was spun at 13,000rpm for 30mins and the ethanol discarded and the pellet dried.

5. The pellet was then resuspended in 100µl loading dye. The vector was then purified in the same way by CHEF gel electrophoresis and electroelution, as described in 2.2.1.2.

2.2.1.6 Control ligations

Materials and Equipment

The same as section 2.2.1.4 but without the T4 Polynucleotide kinase.

- Phenylmethylsulphonylfluoride (PMSF, 40mg/ml in isopropanol)
- Microfiltration discs (Millipore, 50mm, 0.025µm pore size)
- 0.5xTE
- ElectroMax DH10B cells (Life Technologies)
- Life Technologies CellPorator equipped with a voltage booster
- Chloramphenicol (Sigma)

1. A series of ligations (designated A, B and C) were set up to ensure that the ligations were working correctly. The basis of the ligations were set up as follows:

Vector	2.5µl
DDW	8.25µl
Positive control insert	20µl
30% PEG 8000	12.5µl
0.1M $MgCl_2$	1.25µl
10x Buffer	5µl
T4 DNA Ligase	0.5µl

The A ligation contained no control insert DNA

The B ligation contained no control insert DNA and no ligase

The C ligation contained control insert DNA from the large *Bam*HI digested λ fragment prepared in 2.2.1.3

2. These were incubated at 16°C for 5 hours. The ligation reaction was stopped by adding 2.5µl 0.5M EDTA and 1µl Proteinase K.
3. This was then incubated at 37°C for 1 hour and then 1µl of PMSF was added to kill the Proteinase K.
4. The ligations were then dialysed by placing on the surface of microfiltration discs that were floating on 30ml 0.5xTE in a Petri dish at 4°C. These were left for 4 hours. The ligation was then recovered and stored at 4°C until transformation.
5. The transformations were done using 2µl of the ligation with 20µl of cells and the electroporator set to 400V, 25µF and 1000Ω, fast charge rate.

6. The cells were then plated out onto agar plates containing 20μg/ml chloramphenicol and left at 37°C overnight.

2.2.1.7 Miniprep

Materials and Equipment

- Bench top centrifuge capable of 2000g
- Bench top microfuge capable of 13,000rpm
- New Brunswick Scientific Innova 4300 shaking incubator
- LB broth (see appendix 2)
- Chloramphenicol (12.5μg/ml, Sigma)
- GTE (see appendix 2)
- 3M Potassium Acetate
- Isopropanol
- 70% ethanol
- T0.1E (see appendix 2)

1. Colonies from the plates were inoculated into 5ml of LB broth containing chloramphenicol and incubated with shaking at 37°C for 18-20 hours.
2. The culture was centrifuged at 2000g at 4°C for 15mins using a bench top centrifuge.
3. The supernatant was poured off and the pellet was resuspended in 0.2ml of ice-cold GTE.

4. The suspension was transferred to a 1.5ml microfuge tube and 0.4ml (freshly prepared) 0.2N NaOH/1% SDS was added. The solution was mixed by inversion gently several times. This was left at room temperature for 5mins.
5. Potassium acetate was added and the tube inverted gently to mix. The mixture was then spun at 13,000rpm on a bench top microfuge for 15mins.
6. Supernatant was removed (0.75ml) without disturbing the pellet and transferred to a clean microfuge tube. Ice-cold isopropanol (0.45ml) was added and the tube was spun at 13,000rpm for 15mins to pellet the DNA.
7. The supernatant was removed and the pellet rinsed with 1ml of cold 70% ethanol. The ethanol was removed and the pellet was dried in the tube upside down until the pellet became transparent.
8. The pellet was then resuspended in 10µl of T0.1E.

2.2.1.8 Ligation check

Materials and Equipment

- Bovine Serum Albumin (100x 10mg/ml, New England Biolabs)
- 10x NEB Buffer 3 (New England Biolabs)
- *Not*I restriction enzyme (10U/µl, New England Biolabs)

1. A *Not*I enzyme digest was set up to determine if the insert had been incorporated into the vector and to ensure that there were no problems with the cloning site. The digestion was setup as follows:

Vector/Insert DNA	4µl
Buffer 3	1µl
10xBSA	1µl
*Not*I	0.5µl
DDW	3.5µl

2. This was incubated at 37°C for 1 hour and the reaction stopped by incubating at 65°C for 20mins. 1µl was run on a 0.8% agarose gel to determine if the insert had been excised from the vector.

2.2.2 Insert DNA preparation

Dr Michael Anderson from Manchester University supplied *Aspergillus fumigatus* Af293 DNA in agarose plug form. This is because the organism is highly virulent and the Sanger Institute does not have the facilities or the capabilities to harvest the fungus in a safe environment, so the DNA was extracted and supplied on our behalf.

2.2.2.1 Size fractionation

Materials and Equipment

- *Sau*3AI restriction enzyme (10U/µl, New England Biolabs)
- 10x*Sau*3AI restriction enzyme buffer (New England Biolabs)
- Bovine Serum Albumin (100x 10mg/ml, New England Biolabs)
- Proteinase K (14mg/ml, Roche)

- 1M EDTA
- T0.1E (see appendix 2)
- Ethidium Bromide (10mg/ml aqueous solution, Sigma)
- CHEF apparatus (see appendix 2)

A high concentration of *Aspergillus* DNA was required in the correct size ranges for the BAC library to be successful. Partial digestion with an enzyme and CHEF gel separation was used to achieve this. Firstly, a concentration of enzyme had to be determined to provide the correct amount of partial digestion of the DNA in the plug without the digestion going to completion. The enzyme chosen was *Sau*3AI, because its cut site was compatible with that of the cut site of *Bam*HI.

1. An enzyme concentration series was set up in order to determine the amount of enzyme required to meet the parameters stated above. The DNA plugs were digested in 1.5ml microfuge tubes as follows:

	1	2	3	4	5	6	7
*Sau*3AI buffer (µL)	50	50	50	50	50	50	50
*Sau*3AI (µl)	5µl 1/100	1µl 1/10	2µl 1/10	3µl 1/10	0.5µl	0.7µl	1µl
1M Spermidine (µL)	1.3	1.3	1.3	1.3	1.3	1.3	1.3
BSA 100µg/ml (µL)	2.5	2.5	2.5	2.5	2.5	2.5	2.5
DDW (µL)	441.2	445.2	444.2	443.2	445.7	445.5	445.2
Units	0.5	1	2	3	5	7	10

These were incubated on ice for 2hrs and then at 37°C for 20mins.

2. The digest was stopped by incubating plugs with Proteinase K and EDTA at 37°C for 1 hour, as described in 2.2.1.6. The plugs were then removed from the reaction mixture and washed in T0.1E to remove any excess enzyme. The plugs were then loaded onto a 1.0% CHEF gel and run under the same conditions as those stated in section 2.2.1.2.
3. From this, the gel was visualised by staining with Ethidium Bromide and viewing under U/V light. From these digests, an optimal enzyme concentration for partial digestion was chosen. This was done by visually deciding which of the plugs had given the highest, cleanest yield of DNA fragments in the correct, required size ranges.
4. Once the correct and most suitable enzyme concentration had been determined, the conditions for this digestion were then used to digest a large number of plugs (approx. 6-8 plugs). These plugs, once digested were then placed into another CHEF gel and a pseudo double sizing procedure was used to help remove the smaller fragments of DNA in two steps and using the same gel. Smaller fragments need to be removed to improve the library, as smaller fragments tend to ligate much more easily and would compete with the larger fragments in the ligations, reducing the overall average insert size of the library. This is difficult to achieve as the smaller fragments can stay in the gel due to a phenomenon called "trapping". This is where much smaller fragments of DNA are "trapped" by larger fragments of DNA as the DNA migrates through the gel. When the required size DNA is extracted from the gel, the smaller, unwanted fragments are also removed

with them and cause problems previously mentioned and gives an overall poor representation of the genome. Figure 6 describes the process in detail.

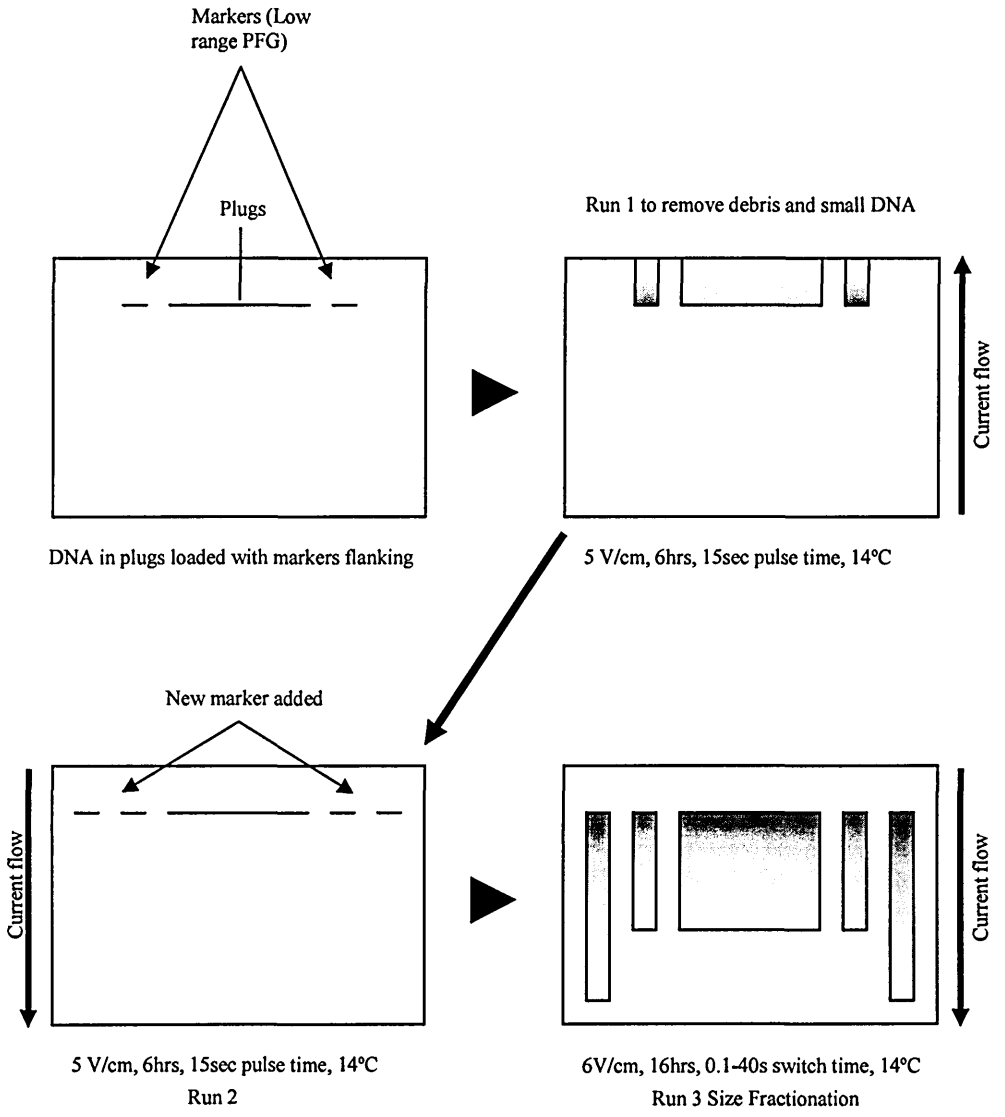


Figure 6. Diagrammatic representation of the pseudo double size fractionation of the partially digested DNA. The flanking marker lanes are removed when the fractionation has finished and stained separately to measure the marker distances, as not to damage the DNA in any way.

5. The gel was initially orientated so that the DNA ran towards the nearest edge of the gel. Using the conditions in the Osoegawa paper as guidelines, the smaller fragments of digested DNA and cellular debris, along with the smaller fragments of marker, were run off the edge of the gel without removing the larger, more useful size range fragments. The gel was then turned 180° and run under the same conditions as the first run to bring back the remainder of the DNA to the well from where the DNA started. The third run was run under normal conditions to separate the fragments more evenly for easier extraction of the correct sizes. As a guideline, a further marker lane was added to the gel to provide indication that the DNA had correctly run and that there was no debris or smaller fragments.
6. Once the pseudo double sizing had been completed, the flanking marker lanes were removed and stained separately and then visualised under U/V light to determine where the DNA had run. The size ranges required were marked and the flanking regions then placed back at the main body of the gel and the size ranges cut out.

2.2.2.2 DNA Extraction: Electroelution

Materials and Equipment

- ¾ inch dialysis tubing (Invitrogen)
- Small electrophoresis tank and power pack
- 0.5xTE (see appendix 2)
- 30% Polyethylene glycol (PEG) (Sigma)

1. The DNA was removed from the gel slices by a process called electroelution. Similar to electrophoresis, but the gel slice and therefore the DNA itself is contained within dialysis tubing so the DNA migrates out of the gel and into the tubing to be recovered.
2. The gel slices were placed into separate ¾ inch dialysis tubing, which were clipped at one end to seal. The tubes were then submerged in 0.5xTBE and the other end clipped to seal, removing all air bubbles at the same time. The tubes were then placed across the flow of current in a small gel tank filled with 0.5xTBE and the power was run at 3V/cm for 3 hours. At the end of the run, the current was reversed for 30-45secs to detach any DNA from the walls of the dialysis tubing.
3. The tubes were then removed from the gel tank and placed, still sealed and containing the gel slice, into 900-1000ml of 0.5xTE (not T0.1E). The tubes were then dialysed in this at 4°C overnight, with gentle stirring on a magnetic stirrer.
4. The tubes were then placed into 0.5xTE containing 30% PEG8000 and dialysed for 1 hour to concentrate the solution to about 300µl. The solution was then removed from the tubes using a wide bore pipette to prevent damaging the DNA and placed into an eppendorf 1.5ml tube.
5. A small amount of each size range (1µl) was run on a 0.8% check gel for 45mins to check the recovery efficiency of the electroelution.

2.2.3 Ligation, transformation, plating and gridding

Materials and Equipment

- 30% Polyethylene glycol (PEG) (Sigma)
- 1mM $MgCl_2$
- T4 DNA ligase (5U/µl, Roche)
- T4 DNA ligase buffer (Roche)

It was important to optimise the ligation to a 1:10 molar ratio of insert: vector to reduce the number of non-recombinant clones and maximise cloning efficiency. A number of attempts were made, all with varying degrees of success and as there were a number of insert DNA preparations, the concentration of DNA recovered did not always remain the same. So, the ligation conditions changed very slightly every time a ligation was set up, depending on the concentration of the insert DNA to maintain the 1:10 molar ratio. The ligations were set up as follows:

pBACe3.6 Vector	2µl
DDW	to make up to 50µl
Insert DNA	1 to 20µl*
30% PEG8000	9µl
$MgCl_2$	1µl
Buffer	5µl
T4 DNA Ligase	1µl (of a 1/10 dilution)

*Amount dependant on concentration of insert DNA to maintain insert: vector ratio. Incubated at 16°C overnight. The ligations were then transformed using the same protocol already described in section 2.2.1.3 and plated out and incubated as also described in that section.

1. An A, B and C control was also set up to compare with to determine whether or not the ligation had worked. These were set up in the same way, but the A control had no DNA in it, the B control had no DNA or ligase and the C control had both ligase and the λ DNA fragment as its control DNA.
2. DNA was extracted from a number of colonies as described in section 2.2.1.3. Clones were analysed be digesting thesis minipreps with *Not*I and separating the digests by CHEF electrophoresis as previously described.

2.2.4 Replication and gridding

Materials and Equipment

- LB Broth (see appendix 2)
- Chloramphenicol (20mg/ml)
- 96 well, deep well boxes (Beckman)
- Polythene plate sealers
- Flat bottomed microtitre plates (Beckton and Dickinson)
- Pre-cut 78x119mm Nytran N nylon filters (Schleicher and Schuell Bioscience)
- 3MM filter paper (Whattman)
- 10% SDS (see appendix 2)

- 0.5M NaOH
- 1.5M NaCl
- 0.15M NaCl
- 0.5M Tris-HCl (pH7.4)
- 50mM Tris-HCl (pH7.4)
- 2xSSC (17.53g Nacl, 8.82g Hydrated Trisodium Citrate to 1ltr with DDW)/0.1% SDS

To allow for easy screening of the library by hybridisation the clones were gridded out on a nylon filter. All picking of colonies and media aliquoting was done in sterile conditions in a fume hood to minimise contamination.

1. LB broth (1ml) containing 20mg/ml chloramphenicol was added to 36x96 deep well boxes. Then, using toothpicks, single colonies were picked from the plates and placed into every well of all 36 boxes in a random order. This was now the set order for the clones.
2. The boxes were sealed with a cellophane plate sealer and each well punctured with a small hole to let air in. The boxes were grown overnight at 37°C in a shaking incubator at 300rpm.
3. Culture from each well (70µl) was added to 36x96 flat bottomed microtitre plates. and glycerol was added to a final concentration of 7.5%. These were then stored at -70°C until needed. Overall, 6 copies of the library were made for either archiving.

4. From the remaining culture, the library was sent to be gridded out onto labelled nylon filters robotically, using a specialised machine built in house and according to the specific protocol of that machine. The arrays were set out in the same form described in chapter 3, section 3.2.1.1. The filters with freshly printed and inked colonies were grown at 37°C overnight on agar plates.
5. Two large sheets of filter paper were placed on 2 large trays. One tray was saturated with 10% SDS and the second tray was saturated with 0.5M NaOH/1.5M NaCl. The bubbles were removed from the paper by rolling out with a glass roller. Excess solution was then drained. Filters were then placed colony side up on the first tray for 5mins, ensuring no solution is washed onto the upper surface of the filter.
6. The filters were then removed and any excess SDS was blotted onto tissue. The filters were then placed onto the second tray, ensuring no air bubbles for 10mins.
7. The filters were removed from the solution and then placed onto a dry piece of filter paper for 10-20mins, ensuring no over drying.
8. The filters were neutralised by submerging fully, colony side up, in 0.5M Tris-HCl pH7.4/1.5M NaCl. The solution was changed and replaced and then the solution containing the filters was placed on an orbital shaker for 5mins at low speed. This step was repeated once more.
9. The solution was replaced with 1 litre of 50mM Tris-HCl pH7.4/0.15M NaCl and placed on the shaker for 5mins.
10. The solution was drained and the filters rinsed in 1 litre of 2xSSC/0.1% SDS for 5mins, shaking.

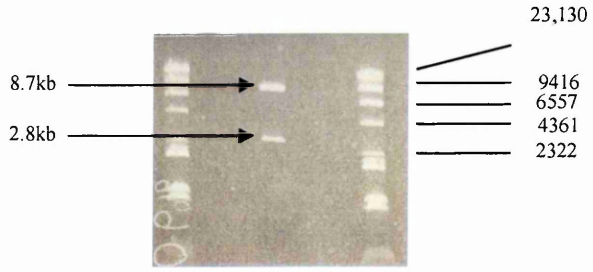
11. The solution was drained and the filters rinsed in 2xSSC for 5min, shaking.
12. The filters were then rinsed twice in 1 litre of 50mM Tris-HCl pH7.4 for 5mins, shaking.
13. The filters were then air-cooled colony side up on chromatography paper for 4-5 hours. The filters were then U/V cross-linked colony side down for 2mins to fix the DNA to the membrane.

2.3 Results

2.3.1 Stuffer fragment removal

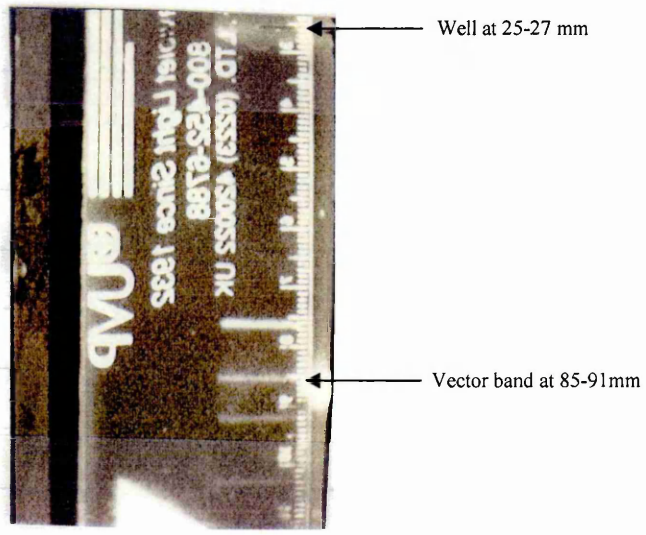
The stuffer fragment was removed to liberate the cloning site and make the vector amenable to ligation, as described in section 2.2.1. The digestion reaction to remove the stuffer fragment was checked to determine whether the reaction had worked. A small amount was run on a small agarose gel (Figure 7).

Figure 7. Agarose gel of *Bam*HI digest of pBACe3.6, showing the 8.7kb vector band and the 2.8kb "stuffer" fragment clearly defined and separated



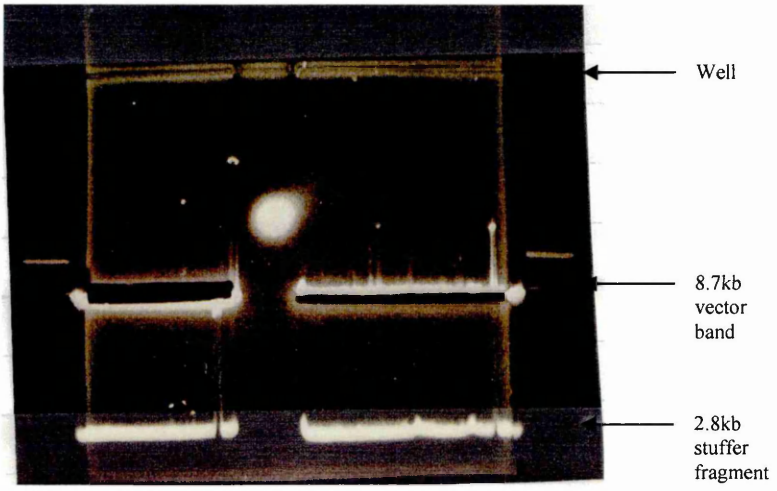
This shows that the digest had worked successfully with no miscellaneous banding or smearing. The digested vector was then run on a CHEF gel once the check had shown the digestion had worked (Figure 8).

Figure 8. One edge of a CHEF gel containing the marker and a small amount the vector band. The vector band in the gel slice glows with Ethidium Bromide staining and the distance from the well the band has migrated is measured. This is recorded and used to locate the band in the main central section of the gel, which contains the main body of the vector band. The same distance from the well is then measured and the band can be cut from the gel without staining the main section.



When the location of the bands had been completed, the linearised vector was cut from the main body of the gel from the corresponding positions that were identified on the marker sections. (Figure 9)

Figure 9. This is an image of the central section of the gel. The linearised vector fragment was excised and the remaining gel sections stained to check that the vector had been removed cleanly and from the correct location.



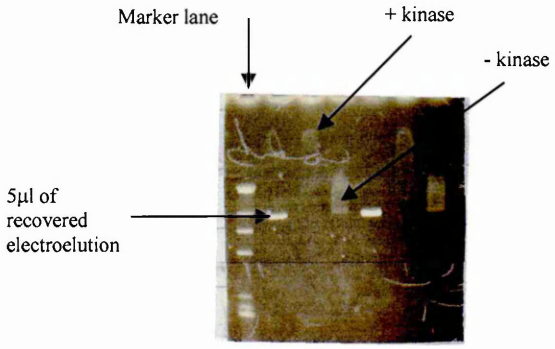
2.3.2 Cloning site check

If the removal of the stuffer fragment was successful, the cloning site should be of good quality and amenable to ligations, as previously described in section 2.2.1.

The kinase reaction would show whether or not the vector could be ligated. If the kinase treatment works, then the vector can then self-ligate again. Running the DNA on an agarose gel takes advantage of the differing migrations properties that this causes.

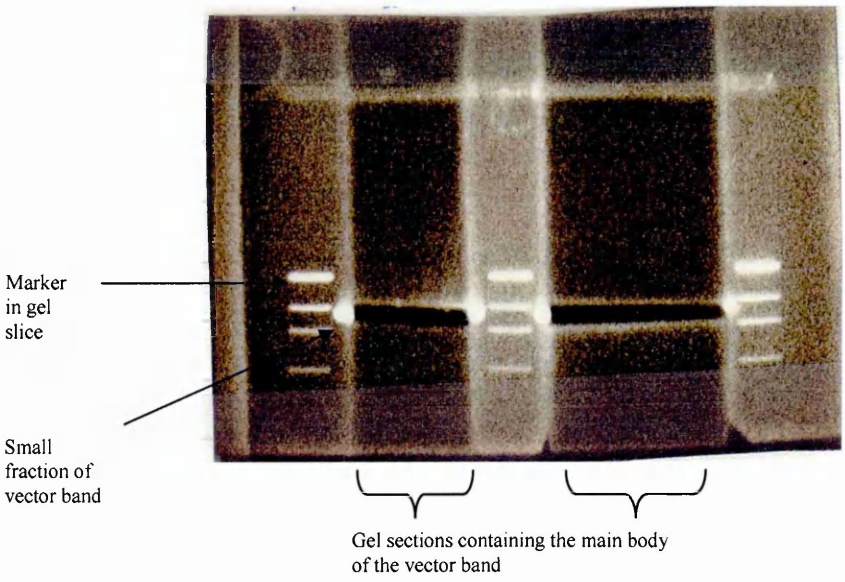
Figure 10 shows that this is exactly what had happened and proves that the cloning site is good and usable.

Figure 10. Here the image shows a check of the ligated DNA with and without kinase treatment. The ligations with kinase treatment did not migrate easily through the gel and have remained near the top of the gel. The ligations not treated with kinase have migrated further down the gel. The brighter bands are 5µl of the non-ligated vector to check quantities recovered by electroelution. Here, the band is bright and strong, suggesting the recovery had been efficient.



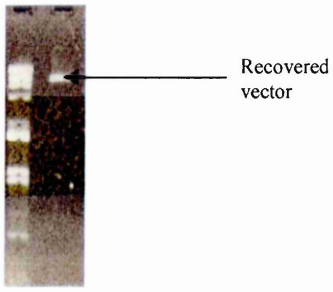
Once it had been determined that the vector could be ligated, a bulk ligation was set up and run on a CHEF gel to remove the non-dephosphorylated DNA (Figure 11)

Figure 11. CHEF gel containing vector band after bulk ligation to remove any non-dephosphorylated DNA that may still be present. Here, the gel was treated in the same way as before, with the outside marker lanes being removed with a small amount of vector, stained and the distance from well noted and then the band removed from the main body of the gel.



The DNA was recovered from the gel slices by electroelution and a final check was done to ensure a good recovery by running 1µl of elute on a check gel (Figure 12).

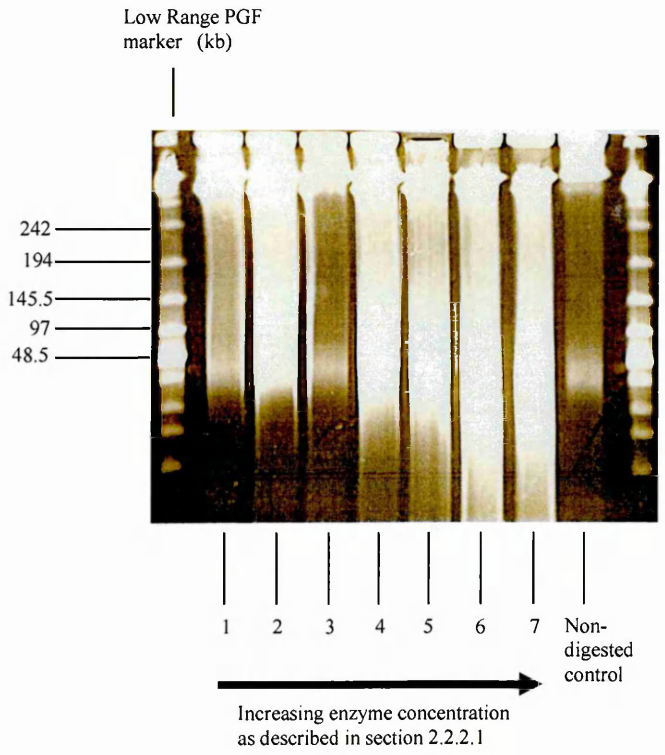
Figure 12. A good recovery of the vector was made. The relative brightness of the vector band compared to the top band of the marker gave an indication of the quantity of recovered DNA as well as the quality.



2.3.3 Size fractionation

The enzyme digestion gradient gave good results. The general trend was as expected, with the amount of digestion increasing with the amount of enzyme. But one or two plugs either did not digest or digestion was insufficient, possibly due to the inconsistent concentration or quality of DNA in the plugs. Figure 13 shows the digestion pattern.

Figure 13. Trial partial digestion of *Aspergillus* plugs with increasing *Sau*3A concentration.



From the evidence in Figure 13, the optimal digestion conditions were taken from 2 of the digests. Lane 4 (3U enzyme) had good strong DNA and a high yield in the size range

required. Lane 7(10U enzyme) also had very good results. The fact that there is a large difference in the concentration of the enzyme must mean that there was possibly a different yield of DNA in each of the separate plugs, regardless of the fact they were prepared at the same time. Lane 7 does also appear to have a much higher concentration of smaller fragments than lane 4. This may be due to the increased digestion received by this plug from the higher concentration of enzyme.

A full digestion was then done on a number of plugs using the conditions from these two digests. Half were digested with the conditions from lane 4 in the series and half were done with the conditions from lane 7 in the series (Figure 14).

Figure 14. Gel showing large scale partial digestion of *Aspergillus* plugs with desired insert fragment sizes removed

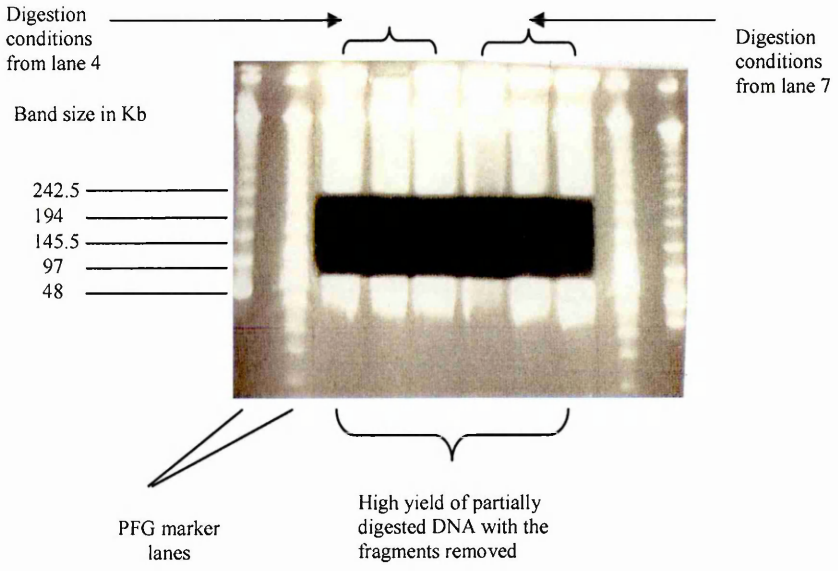


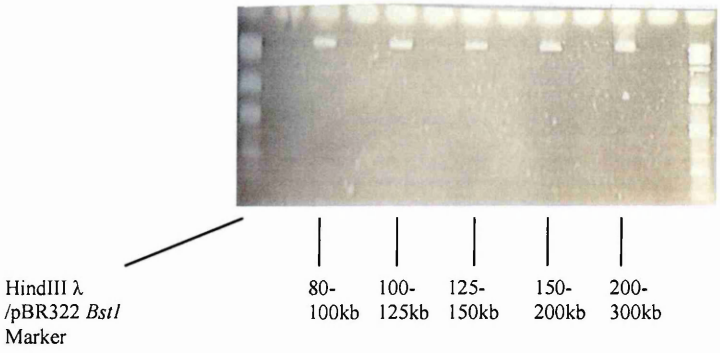
Figure 14 also shows that the pseudo double sizing step was successful in removing smaller fragments of DNA. It can be seen that the bottom of the digests have clean, straight cut off points and that the smaller fragment of the outside markers are also missing. This is a good sign as it means that there is a reduced possibility of DNA trapping, as the smaller fragments will have been removed before the larger fragments have migrated in the gel.

2.3.4 Electroelution

The electroelution process itself was straightforward. At the electroelution and recovery stage, the large DNA fragments were concentrated by dialysis against 30% PEG. This posed a few problems. Not only was the solution containing the DNA very difficult to get out of the dialysis tubing, it was invariably of differing volumes with each attempt. The length of time it took to reduce the volume to a given amount for each attempt also fluctuated wildly. The results ranged from a few minutes for a complete reduction in volume to a few hours just to reduce it a few microlitres. There was no way of knowing how quickly or how slowly the volume reduction was going to take, even when all conditions were reproduced. Nonetheless, a careful, watchful eye was a kept to ensure the bags did not dry out and the DNA was recovered after maximal concentration.

There was a good, consistent concentration in all of the plugs and more than adequate DNA in the 80-300kb size range. DNA was cut at the required size ranges, extracted and 5μl run on a small check gel to determine yield and quality of DNA (Figure 15).

Figure 15. 1µl aliquots of DNA recovered from each of the size ranges excised from the gel.



The yield from each size range was good and the integrity of each appeared excellent.

2.3.5 Ligation

The aim was to achieve an average insert size of between 125-150kb, but achieving this proved difficult. Eventually one ligation gave an average insert size of 74kb. This was still well below the size ranges of the fractions that were removed from the gel, but was usable for a BAC library and was not prohibitively small for sequencing. This ligation had a DNA volume of 10µl in a total ligation volume of 50µl to try and optimise the frequency of high molecular weight inserts over smaller inserts and non-recombinants. To try and reduce the amount of enzyme that would be there generating non-recombinants, the enzyme was diluted 1:10 with 10xBSA. The ligation reactions were set up to try and optimise maximum cloning efficiency for large insert sizes, using the conditions recommended in the Osoegawa paper as guidelines. There was 25ng of pBACe3.6 vector (approx) and up to 65ng of insert DNA that gave a recommended ideal

1:10 molar ratio of insert to vector. Overnight ligations were done instead of 4hr ligations. A series of ligations were carried out at 4 hours, but this seemed to give a very low titre. When ligations of at least 16 hours were done, the titre increased to levels that were much more usable.

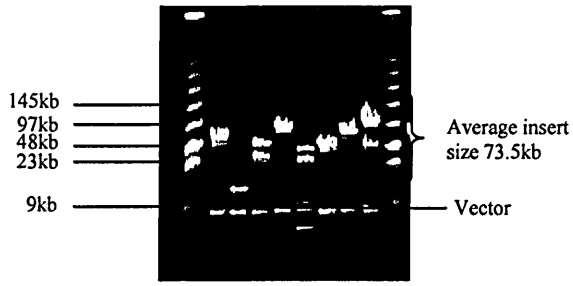
10µl of DNA was used in the ligation and 1µl of the ligation used to transform the DH10B cells. Three ligations from the successful size fractionation step were done, designated B28, B29 and B30. These represented the size fractions removed from the gel in 80-100kb, 125-150kb and 150-200kb respectively (see Table 10)

Table 10.

	Expected insert size from gel excision(kb)	Vol. DNA (µl)	Plated (µl)	Colonies	Ave. Insert Size (after NotI digestion)
B28	80-100	10	1	600	73.5 kb
B29	125-150	10	1	380	24kb
B30	150-200	10	1	400	33.9kb

Figure 16 indicates that the ligation had worked with the correct size inserts being present. The *Not*I digest gave an average insert size of 73.5kb. This was not an ideal size as the ligation was from the 80-100kb size range, but nonetheless, it was a usable, workable size range and was much better than the other size ranges and previous attempts at making the library. Sizes ranged from 12kb to >150kb.

Figure 16. The final check gel after digestion with *Not*I that confirmed that the ligation had worked sufficiently to continue to generate the library.



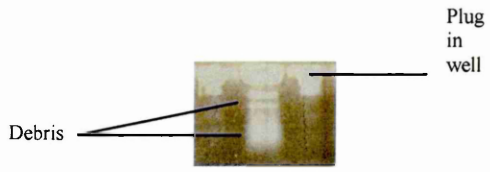
2.4 Discussion

Although Osoegawa used *Eco*RI to remove the 'stuffer' fragment and therefore, *Eco*RI was used as the 'cloning' enzyme, the use of a six base cutting enzyme, such as *Eco*RI, may bias the digestion and therefore bias the library. *Bam*HI was decided upon because of its multiple compatibility with other enzymes. It has a 6-base recognition sequence but has compatible cohesive ends with *Sau*3AI. *Sau*3AI has a four base cutting enzyme which was thought would give a more random digestion in an unknown genome.

The insert preparation process was the most problematic area of the library construction. This was mainly due to a poor DNA concentration in the DNA plugs supplied by the University of Manchester. The aim was to obtain as clean and as high a yield of DNA as possible from the stock in agarose plugs that was sent to us. Originally following the methodology in the Osoegawa paper as closely as possible, a number of attempts were made from the plug stocks we had available. Most were unsuccessful and it was very

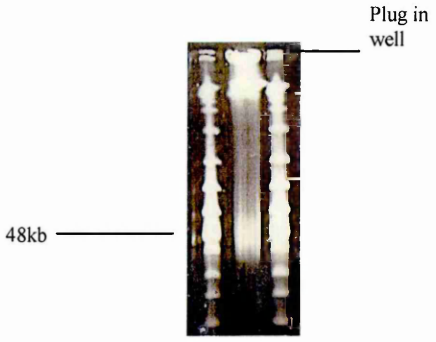
puzzling. There did not seem a logical explanation until another step (pre-electrophoresis) that was not used in the final library production method brought the problem to light. The DNA impregnated plugs were placed into a gel and run between 1 and 6V/cm, 1sec switching time, 14°C for any length of time between 1 and 6 hours to remove the very small DNA and RNA fragments that may be present in the plug, as well as any left over cellular debris that may have had a detrimental effect on the library production. Unfortunately, when the pre-electrophoresis stage was carried out, it seemed that it took a sizeable amount of good DNA with it, reducing the overall amount that could be used for digestion. In addition to this, it showed that the concentration of the DNA that was in the plugs was not that high in the first place. It was fortunate that some DNA did come out of the plugs under some conditions because otherwise it would have been quite difficult to spot that the DNA concentration *was* actually quite low. Moreover, in theory it was a good idea to remove the unwanted material, but every batch of plugs that were sent to us did not respond as expected at this step. On a number of occasions, there was evidence to suggest that the conditions had been optimised and that the debris was being removed with minimal DNA removal, (Figure 17a) and consequent examination on a CHEF gel would show that there was a lot of what seemed good quality DNA left behind. But this was not to be the case. As digestions were done after the pre-electrophoresis stage, it was soon evident that the quality of the DNA, as well as the amount being recovered, was far removed from the standards that were required to generate a library. (Figure 17b). A number of further attempts were made on several batches of plugs. All had the same result.

Figure 17a



An example of a pre-electrophoresed plug run at 1V/cm, 2 hours, 1 sec switching time and 14°C. It plainly shows banding of what could be either debris or RNA at high concentrations.

Figure 17b

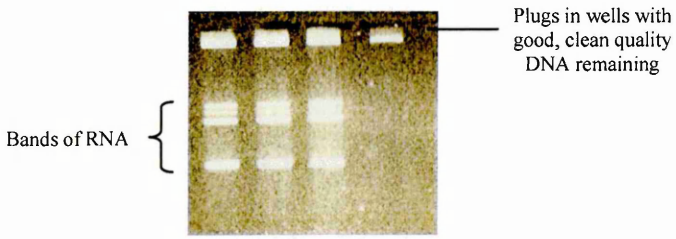


The same plug run at digestion conditions (6v/cm, 16 hours, 15 sec switching time and 14°C. This shows that there is still, even after pre-electrophoresis, a large concentration of debris. This was detrimental to the downstream techniques we were using to create the library as the smaller fragments of DNA would compete against the larger fragments, as well as the non-DNA/RNA debris would contaminate the ligations and prevent good sized inserts for the library.

The digestions conditions themselves were sometimes problematic. Various batches of DNA plugs were sent to us from Manchester. Digestion conditions remained the same with every batch, yet the results were very variable. Some plugs would go to complete

digestion with the minimum amount of enzyme; others would not digest regardless of enzyme concentration. Digestion times were varied, with mixed success. These findings were communicated and plugs with the desired quantity requested from Manchester University. The new plugs seemed to have more consistent digestion and gave good DNA yields. After consultation, the plugs from Manchester were of high DNA concentration and with a much better quality of DNA. Even with the pre-electrophoresis stage being used only as an indicator to the quality of the DNA, it showed that there was very little debris in the plugs (Figure 18).

Figure 18. Pre-electrophoresis stage showing good quality DNA being left behind in the plug, yet there is little debris and the RNA can clearly be seen to have travelled cleanly from the plugs.



The size fractionation was much more controlled after digestion, so there was consistency at this stage with all plugs. The rotation of the gel through 180° made sure that there were as few small fragments of DNA as possible. It became clear that the methodology was sound and when the quality of the starting material was improved, success was achieved after two attempts. The first attempt ascertained the conditions to digest at and to determine the quality of the DNA. The second attempt was to generate the library. It was

also less contaminated with debris and consequently there was less of the small size DNA that could compromise the ligation reactions. It can be concluded that the successful outcome of the experiment was probably due to the much higher quality of DNA provided, with less smaller size DNA being "trapped" with the larger size DNA and co-migrating with it during electrophoresis.

The "pseudo" double run was designed to help the removal of the smaller fragments and cellular debris and was relatively successful. There are other methods, most notably the "double run" method, described by Peterson *et al* (2000). This method is relatively efficient at purifying the DNA, but is much more time consuming and also uses more resources. In this method, DNA is run through a gel, side by side with markers. The markers are removed and visualised, noting where the desired fragment sizes are and then these are compared back to the main body of the gel. The fragment sizes required, using the marker gel slices as a guide, are then removed from the main body of the gel. This is then cut into three equal sizes corresponding to the three different size fractions desired and then placed into the wells of another gel. The fragments are then run again to remove any small, trapped fragments of DNA that had migrated with the larger fragments in the first gel. The principle is repeated, where the markers are stained and the fragment sizes noted and compared back to the main body of the gel and the correct, "double purified" fragments are cut out and the DNA extracted.

The method that was used for the *A. fumigatus* project was chosen due to its higher recovery of DNA, its efficiency of removing the smaller fragments of DNA, is more

conducive to being used with smaller concentrations of DNA and is also more practical as the DNA remains in one gel throughout the process. As we expected small amounts of DNA, a method like this was ideal, as the Peterson method would have a high level loss of DNA due to the extra purification steps. The method we used only ran DNA off the end of the gel, removing the unwanted smaller fragments completely, yet leaving the main fractions of DNA untouched in the gel, with little loss overall (Figure 14). Although both methods are designed to achieve the same results, we thought that with the possible constraints we might have with this project (low DNA concentration, low DNA integrity etc), the method described in the Osoegawa paper seemed to be the logical choice.

Electroelution was used to recover the large DNA fragments from the excised gel slices after the size fractionation for two main reasons. It has been shown by Strong *et al* (1997) that the high molecular weight DNA can be recovered with a higher degree of integrity compared to that of enzyme based agarose digestion for extraction of the DNA. The study showed that the integrity of the DNA before and after extraction by comparing the size of the DNA was better when electroeluted and not enzyme digested from the agarose. Enzyme digestion requires a number of steps that may be detrimental to the state of the DNA, such as heating, pipetting and agitation. Small fragments of DNA may be broken off and these can the ligation. Electroelution uses the same principle to get the DNA out as it does to get the DNA into the gel. Direct manipulation of the DNA is kept to a minimum, as well as using wide bore pipettes to prevent shearing of the DNA. This reduces the risk of breaking the DNA into smaller fragments. Also, electroelution allows the use of high melting point agarose, rather than low melting point agarose that enzyme

digestion works best on. This increases resolution of separation and helps when handling the gel, as high melting point gels are much stiffer and less prone to shearing than low melting point gels.

The running buffer used was the same as the Osoegawa paper methodology, with 0.5xTBE as the main buffer. It has been noted that borate ions may inhibit ligations further downstream (Ioannou *et al* 1994; Strong *et al* 1997), so they were removed by a dialysis step against TE after elution. This allows the running of the gels to be in 0.5xTBE, as TBE buffer allows the best resolution of the size fragments during size fractionation (Osoegawa *et al* 1998).

There also seemed to be differences in the times that PEG concentration took to reduce volumes within the tubing that could sometimes be inconvenient. Some PEG dialysis would take no more than a few minutes, while some would take hours with little or no discernible difference in internal volume. A possible explanation for this inconvenience is the differences in starting volumes in the dialysis tubing. If there were a lot of liquid to start, then it would possibly take longer for the PEG to remove the water. Also, as the water component was removed by the PEG, it could cause localised dilution of the PEG, reducing the osmotic gradient between the tubing contents and the PEG, which will in turn increase the time to remove more water. If there were low volumes of starting liquid in the dialysis tubing, then this effect was reduced, as there would be less liquid to remove at the start and also, as a consequence, less localised dilution of the PEG, so

maintaining the osmotic difference for longer and removing the water component much more quickly.

There is a combination of factors that make for a good library. Large average insert size, low background of non-recombinant clones and high titre are essential to making a good library. Table 10 in section 2.3.5 showed that the highest efficiency of the ligations was from the ligation with the smallest insert size. This is expected to a degree, as the larger the DNA becomes, the harder it would be to transform into the cell. But there were sufficient numbers to generate a library. What is harder to explain is the insert size fluctuation. Although B28 had an insert average size that was reasonably close to that of the original excised size, it was still some way below what was expected. What was stranger was the large reduction of average insert size in the other two ligations. Not only were they much reduced over that of the B28 ligation, they were just a fraction of the size of what they should have been. The reason for this the phenomenon of DNA trapping.

On a number of occasions after size fractionation had been optimised to a more reliable methodology, excising the insert from the vector after ligation would instantly show that the ligation, although successful, had not worked in obtaining the largest possible average insert size. Figures 19a and 19b shows that although there were inserts that had been excised, they were small and very sparse. This kind of thing happened on numerous occasions when it was thought that there would be a possibility of success.

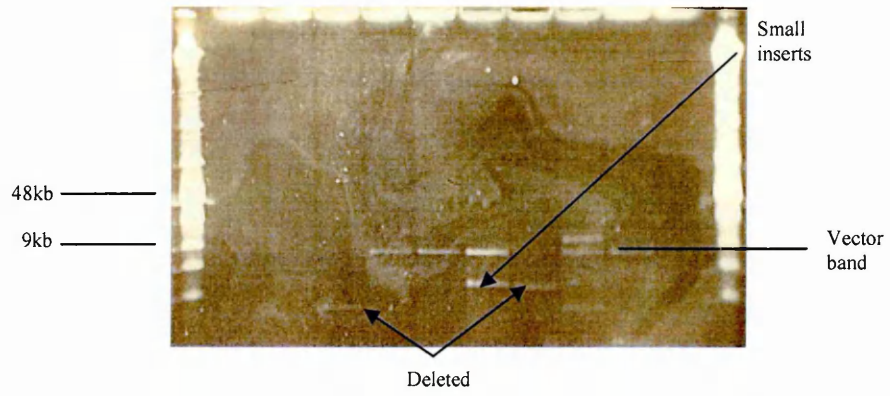


Figure 19a An example of one of the failed ligations. Digested with *Not*I enzyme to excise the insert.

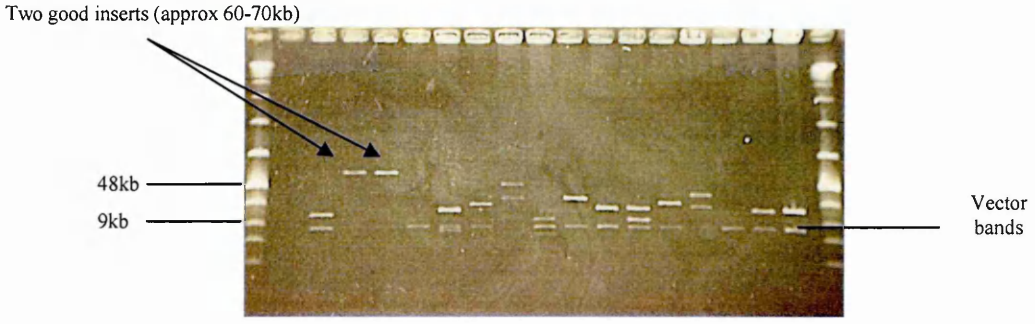


Figure 19b Example of a ligation that nearly worked, as there were two inserts that were over the required size, but as it was only two from the library, it was deemed unusable.

Although numerous attempts were made over the course of approximately a year, each attempt gave a greater degree of success until a single methodology was settled upon that gave a good usable insert size to generate a high quality, stable library. This would hopefully pave the way for relatively easy sequencing, making interpretation of the sequence much easier. More of this is discussed in the following chapters.

3. Physical Mapping and Fingerprinting

3.1 Introduction

Physical mapping of a genome provides an invaluable resource in a number of ways. It gives the location of physical clones and markers across a genome. It also makes possible the selection of clones to be sequenced, ensuring minimal sequencing redundancy and providing the maximum coverage possible of the genome. It localises clones with respect to other clones that may have already been sequenced. Also, when fingerprinting clones, sizing of the clones enables verification of the accuracy of shotgun sequencing of each clone that is identified.

There have been a number of mapping projects that relate to the *A. fumigatus* project, namely the *Aspergillus niger* project (http://www.aspergillus-genomics.org), the *S. pombe* project (http://www.sanger.ac.uk/Projects/S_pombe), using cosmids and the model fungus *Aspergillus nidulans* utilising BACs. The latter also used plasmids and fosmids and although they are relatively reliable, the number of clones required for genome coverage is high. Ideally, the clones used would be as large as possible to create a minimal tiling path (the least number of overlapping clones to cover the entire genome). Plasmids and fosmids have small insert sizes making the logistics problematic and interpretation difficult.

Of most relevance to this study is the closely related organism, *Aspergillus nidulans* sequencing project of

(www-genome.wi.mit.edu/annotation/fungi/aspergillus/background.html). *A. nidulans* and *A. fumigatus* have similar genome sizes (approx 30Mb) and it is thought that they both have 8 chromosomes. *A. nidulans* is used as a model for cell biology, gene regulation and fungal genetics. It is additionally important for carbon and nitrogen regulation studies. The Whitehead Institute for Genome Research (WICGR) is in collaboration with Monsanto to produce a 10x genome sequence coverage for *A. nidulans* i.e. a genome library that represents the genome of the organism up to 10 times over. They are aiming to achieve this by using a shotgun sequencing strategy and incorporating both BAC and Fosmid end sequences into the assembly, to provide linking information. This information will then be integrated into existing physical and genetic maps. So far, Monsanto have contributed a 3x coverage consisting of 16,144 contigs over 29,123,109bp. The estimated genome size is approximately 31Mb and is thought to contain up to 12,000 genes. 432 genes have already been mapped to a locus. 254 of these genes are cloned and sequenced.

But work at WICGR has continued apace and exceeded the original aim alone, already producing whole genome shotgun sequence from 4kb and 10kb plasmids, 40kb Fosmids and 110kb BACs. These were assembled with Monsanto's reads and the result is a 13x assembly that was made public in March 2003, 3x more than was originally aimed for. The aim was to make the annotation public by spring 2003, but this is still ongoing and not yet fully available. What information that is available is primarily used for BLAST searching and downloading by the fungal genetics community.

For the shotgun strategy, the genomic DNA was shattered into small fragments of 4, 10 and 40kb. The fragments were then ligated into either 4 or 10kb plasmid vectors or 40kb

Fosmids. An external collaborator supplied the 110kb BAC library. The ends of the fragments were then sequenced to give paired sequencing reads. Software was used to pair the reads and generate contigs and these contigs were then linked to supercontigs using the read pair information already generated in other contigs.

The total genome length is 30,068,514bp and there are 248 contigs that are longer than 2kb and there are 89 supercontigs covering the genome. The average contig length is 121kb and the average supercontig length is 338kb. A clickable overview of the physical map can be found at http://aspergillus-genomics.org/physical.fls/overview.html.

Other projects, such as the *C. elegans* (http://www.sanger.ac.uk/Projects/C_elegans) and the *S. cerevisiae* (http://www.sanger.ac.uk/Projects/S_cerevisiae/) sequencing projects have also used clone-based maps and have been completed using this information. The *C. elegans* project was one of the first "large scale" genomes to be sequenced (Waterson and Sulston, 1995). The project pioneered various different approaches to physical mapping. The physical map constructed consisted mainly of overlapping cosmids and YACs. At the time of the construction, these two methods were thought to be the best techniques available. YACs were used because of the large inserts, the ease of propagation in yeast, the fact that, long range continuity was more achievable and DNA that was unclonable in cosmids could be cloned. Cosmids were used as they gave a much higher resolution locally and were technically more amenable. A number of techniques were used to construct the map. Restriction enzyme fingerprinting was used to generate cosmid contigs; YAC-cosmid hybridisation to grid the arrays; sequence tag site (STS) assays were used to detect YAC-YAC overlaps; and hybridisation to *C. elegans* DNA for

long range ordering. There was no single technique used that would provide a full linkage of clones and as there was a level of redundancy between each technique, each technique complemented the next in a logical manner.

The *C. elegans* map consists of 17,500 cosmids and 3500 YACs. It was predicted that there were over 13,500 genes within the *C. elegans* genome. This was calculated by using EST (expressed sequence tag) data. The number of predicted genes in a given sequenced region was divided by the fraction of cDNA library for which exact matches found in that region. Even though there was only a quarter of the genes of the genome in the cDNA library, this made no difference to the calculation as the figure calculated rested solely on the assumption that the expression of the genes in the sequenced region is typical of the genome as a whole.

The *C. elegans* project was a project that had data published from its work over a number of years and as a result, a number of independent studies were initiated using the information as it was published during the project. The University of Leeds project involved collecting expression data by using transgenic reporter constructs of the predicted genes (Hope 1994; Lynch, Briggs and Hope 1995); Kohara of the National Genetics Institute at Mishima, Japan used the map to continue to gain information and collect data for his set of sequence tagged cDNAs and is using the information to determine the expression patterns by *in situ* hybridisation and Baillie at the Simon Fraser University and Rose at the University of British Columbia used the map to generate transgenic strains incorporating sequenced cosmids. This will allow the rescue of lethal and visible mutants to determine a precise correlation of the genetic map with the sequence (Howell and Rose, 1990; Schein *et al*, 1993). Each one of these laboratories and

projects makes use of the map extensively. But the main use of the map and sequence have been for the study of specific *C. elegans* genes and for comparative genomics, which may involve up to 150 laboratories around the world.

Initially, it was thought that a whole genome shotgun for the *A. fumigatus* project would be logistically far too difficult because of the scale and cost. So the pilot project was set up to assess the viability of a complete genome sequencing project based on a methodology very similar to that that was being employed by the Human Genome Project and that of the Mouse Genome Project, a BAC based physical map. It was estimated that the pilot project would sequence no more than 1MB by choosing a BAC contig from this map centred around the *niaD* gene. The region of the genome surrounding *niaD* was chosen for this study as it is a well studied gene in *A. nidulans* within a gene dense region, and is thus a good choice for synteny studies.

3.2 Materials and Methods

3.2.1 Identification of the *niaD* gene

Materials and Equipment

- Thick polythene sheet
- Small sandwich box
- Shaking incubator
- Perkin Elmer Cetus PCR machine
- QuikHybe (Stratagene)
- Guys buffer (see appendix 2)

- 5mM d(ATG) (Amersham Biosciences, 100mM)
- AmpliTaq DNA polymerase (Applied Biosystems, 5U/μl)
- α-^{32}P-dCTP (Amersham Biosciences, 3000Ci/mmol)
- $(CA)_n$ =(1mg poly dA-dC poly dG-dT in 1ml T0.1E, Pharmacia Biotech)
- 20xSSC (see appendix 2)
- 2xSSC
- 0.2xSSC
- Total Human Placental DNA (10.8mg/ml, Sigma)
- 0.5x SSC/1% sarkosyl (BDH Laboratories)

The *niaD* gene in the nitrate reduction gene cluster was used as a starting point for the mapping. The library had to be screened for the gene by radiation hybridisation with a *niaD* probe, which was generated using DNA from the previously sequenced *niaD* gene, supplied in the pGEM-T Easy vector (Promega), by Michael Anderson of Manchester University. ^{32}P-radiolabelled probes were produced by PCR using gene specific primers (see below) to perform a hybridisation experiment against the library. Any spots that lit up designate both the plate number from the original library and also the position on that plate where the clone was located.

3.2.1.1 Array probing

1. The filters were robotically gridded according to the manufacturers guidelines as shown in Figure 20

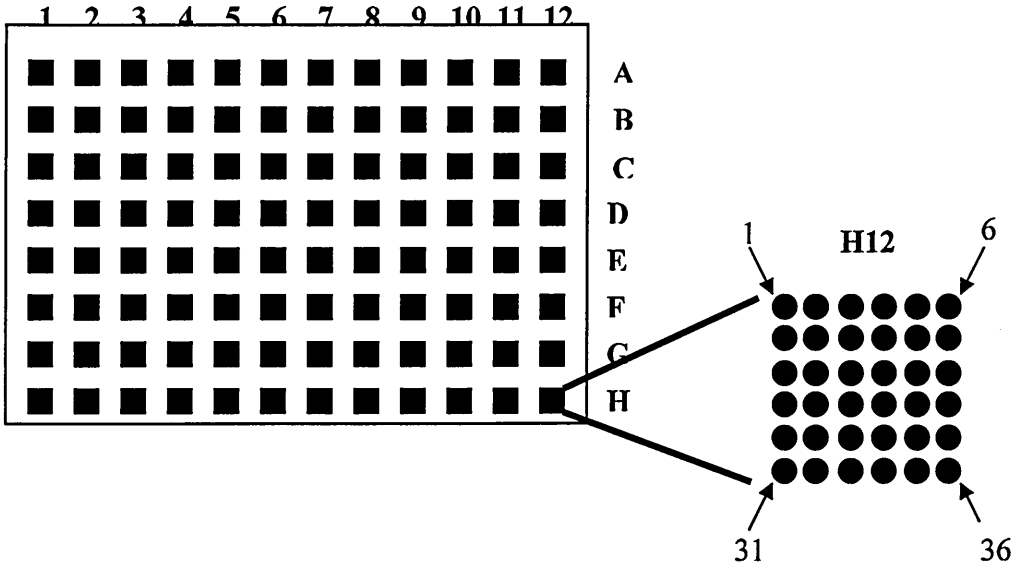


Figure 20. Diagram of gridded filter. From top left to bottom right, each position shows position of clone on the original library plate. Of the 36 dots found in each grid position, this shows the plate number.

2. The filter was pre-hybridised in a small sandwich box, immersed in QuikHyb with a small piece of polythene placed over the top to slow evaporation, for 3 hours at 65°C with shaking at 40 r.p.m. on a flatbed shaker.
3. To make the radiolabelled probe, 2µl *niaD* template DNA was pipetted into a sterile 0.5ml eppendorf tube and 0.2µl of each primer. The primers were sequences supplied by Anderson and represent the ends of the fragment that was

supplied. (3':GCCCCAATAACGCTAGTGTG;

5': CGTTCATCGCGCCACCACAT) (Tm: 60°C). A drop of mineral oil was added to each mix. A pre-mix of 1µl 10x Guys Buffer, 0.4µl d(ATG), 5.7µl sterile H_2O, 0.1µl Taq DNA polymerase and 0.4µl α-^{32}P-dCTP was made. 7.6µl of this pre-mix was added to the DNA/primer mix.

4. PCR was performed as follows (radioisotope shielded):

94°C	5min	1cycle
93°C	30 seconds	
55°C	30 seconds	35 cycles
72°C	30 seconds	
72°C	5 min	1 cycle

5. Following the PCR reaction, a competition mix of 235µl DDW, 5µl $(CA)_n$, 125µl total human placental DNA and 125µl 20x SSC were added to a screw capped eppendorf tube. The PCR was pooled into the competition mix and the cap screwed on. It was then placed in to a boiling water bath for 5min then quenched on ice for 2min.
6. The plastic was removed from the pre-hybridising filter and the filter removed. The probe/competition mix was added to the QuikHyb (plus a little more QuikHyb if some evaporation had occurred) and mixed. The filter was returned to the box, covered with the plastic and placed in a shaking incubator at 65°C overnight.

7. 2 litres of 0.5x SSC/1% sarkosyl were pre-heated to 65°C in an incubator. In a large sandwich box, 1 litre of cold 2x SSC added and the filter transferred to the solution. The filters were "swirled" in the solution and the solution discarded. The process was repeated and the solution discarded. Then 1 litre of pre-warmed 0.5x SSC/1% sarkosyl was added to the filter and the box was placed in a shaking incubator at 65°C for 30min. The solution was discarded and the process repeated. The filter was then rinsed twice using 1 litre of 0.2x SSC. The filters were then placed face down on some pre-flattened cling film and the whole package placed into an autoradiograph cassette face up, a film placed on top with "Glo-gos" (Stratagene) in the cassette to mark the correct position of the film when developing. The cassette was left at -70°C overnight (or longer with fresh film if needed) and then developed.
8. When developed, the film was then placed back over the top of the filters in the cassette and correctly positioned using the "glo-gos" as a guide. Any spots that were "illuminated" on the film could then be scored to the correct position on the filter, giving the correct plate number from the library and the correct position on the plate.

3.2.1.2 Colony PCR

Materials and Equipment

- TYE Agar plates (see appendix 2) containing 20μg/ml chloramphenicol and 5% sucrose
- T0.1E (see appendix 2)

- 40% sucrose/creosol red (see appendix 2)
- 10xNEB PCR buffer (New England Biolabs)
- 100xBovine Serum Albumin (BSA) (10mg/ml, New England Biolabs)
- β-mercaptoethanol (Sigma)
- Taq Polymerase (AmpliTaq 5U/ml, Applied Biosystems)
- dNTPs (Amersham Biosciences, 100mM)

This was used as a confirmation that the "hits" that were found on the gridded array were true positive hits and not anomalies that were picked up by the process.

1. A small inoculum of each "hit" was taken from the original glycerol stock plates from the library and was streaked out onto TYE Agar plates containing 5% sucrose and 20μg/ml chloramphenicol and was grown overnight.
2. A single colony was then picked from each plate and transferred to a microtitre plate containing 150μl T0.1E.
3. A premix of 5.425μl sucrose/creosol red, 1.5μl PCR buffer, 0.495μl of a 1:10 dilution of BSA, 0.21μl of a 1:20 dilution of β-mercaptoethanol, 0.12μl Taq polymerase and 1.5μl dNTP's.
4. The colony inoculum (5μl) was used as template, 0.75μl of the premix was added and PCR was carried out using the same conditions according to the above conditions set out for the array probing.

3.2.2 End Sequencing

Materials and Equipment

- 48 well, deep well boxes (Beckman)
- Plate sealers
- 2xTY (see appendix 2)
- 1mM Tris pH8.5

3.2.2.1 Template DNA preparation

1. Two deep well boxes were filled with 2.5mls of 2xTY media containing 20μg/ml of chloramphenicol. These were inoculated from the original library, in the correct order of the plates.
2. The deep well boxes were then sealed with plastic, self-adhesive plate sealers and small holes punched into the tops.
3. The boxes were grown for 24 hours at 37°C in a shaking incubator at 220rpm.
4. The boxes were spun at 3800rpm at room temperature for 10min.
5. The supernatant was then poured off and the box was dried upside down on a paper towel.
6. To extract the DNA, the REAL96 kit from QIAGEN was used in accordance to the manufacturers specifications (http://www.qiagen.com/literature/Handbooks/PDF/Plasmid_DNA_isolation/INT/Miniprep/PLS_REAL_Prep_96/1019510_QPREALHB_p18_22.pdf)
7. The DNA was re-suspended in 22μl of Tris.

3.2.2.2 Sequencing the DNA

Materials and Equipment

- MJ Research PTC-225 Peltier Thermal Cycler
- BigDye© Terminator sequencing reagent (ABI Prism)
- T7 or SP6 30pmol/µl (120,000pmol/ml, Sigma)
- Precipitation mix (see appendix 2)

1. A sequencing reaction was set up as follows in 96 well PCR plates:

10µl DNA
1µl primer
12µl BigDye©

2. The reaction conditions were set up as follows:

95°C	5min	
95°C	30secs	X40 cycles
50°C	20secs	
60°C	4min	
10°C hold		

3. After the reaction had finished, 50µl of precipitation mix was added to each well.
4. The plates were then spun at 4000rpm at 4°C for 40min.
5. Supernatant was poured off, dabbed dry upside down on paper towel and 200µl of 70% ethanol was added.
6. The plates were spun for 5min at 4000rpm at 4°C. Supernatant was poured off; the plates were dabbed dry and the process repeated.

7. The plates were then spun briefly at 500rpm upside down to remove residual supernatant. The plates were air-dried and then they were loaded onto an ABI Prism 3700 sequencing machine.

3.2.3 Fingerprinting

An enzyme had to be chosen for fingerprinting that would restrict approximately 20 times in each BAC to provide sufficient data for map construction. The only dataset that was available to use for enzyme determination was the BAC end sequences. This data was analysed using a program called EmbossRestrict , which can be found at http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/restrict.html. Restrict uses the REBASE database of restriction enzymes to predict cut sites in DNA sequence. The program will display a number of enzymes showing the number of cuts for that sequence and the starting and finishing point of recognition. From the BAC end sequence that was inputted into the program, the enzyme that was chosen was *Pst*I. Although this enzyme was not ideal as the number of cuts it achieved was nearer the lower end of the acceptable range, there was no other reasonable alternative. It was calculated that *Pst*I would cut a BAC clone to be fingerprinted between 20 and 30 times.

3.2.3.1 Micro-prepping of BACs.

Materials and Equipment

- 96 well, deep well boxes (Beckman)
- 2xTY (see appendix 2)
- 10% w/v glycerol
- Solutions I, II and III (see appendix 2)
- Isopropanol
- 4.4M lithium chloride
- 70% ethanol
- T0.1E (see appendix 2)

1. Growth media (500µl of 2xTY containing 80µl of 25mg/ml chloramphenicol) was dispensed into 1ml deep well Beckman 96 deep well plates.
2. Using a "hedgehog" 96 pin inoculating tool, the medium was inoculated from the BAC library stock glycerol storage plate.
3. The plates were sealed with a microtitre plate cap and incubated at 37°C on a rotary shaker for 18 hours at 300rpm.
4. 250µl of each overnight culture was transferred to a new microtitre plate and centrifuged at 2500rpm at room temperature for 2min.
5. The supernatant was then decanted.
6. The remainder of the culture was stored at -70°C after adding glycerol.
7. Solutions (I, II and III) were made up

8. To each well, 25µl of Solution I was added and the plates were tapped gently to resuspend the pellets.
9. Then 25µl of Solution II was added and this was mixed by again tapping the plates. They were left at room temperature for 5min.
10. Finally, 25µl of Solution III was added and mixed by gently tapping the plates. This too was left at room temperature for 5min. The plates were then covered with plate sealers and vortexed vigorously for 10 seconds.
11. The plates were then centrifuged using a microtitre plate rotor for 10min at 3400rpm and 4°C.
12. 75µl of the supernatant was transferred to a new microtitre plate that contained 100µl of isopropanol in each well.
13. The plates were re-sealed and then placed at -20°C for 30min to 1 hour.
14. The plates were then centrifuged again at 3400rpm, 4°C for 10min. The supernatant was decanted and the plates were drained upside down with care not to dislodge the pellets.
15. Water (25µl) was added and the pellets were resuspended by tapping the plates again. Then to this, 25µl of lithium chloride was added and mixed gently by tapping the plates. This was left for 1 hour at 4°C.
16. The plates were then again centrifuged at 3400rpm at 4°C for 10min.
17. The supernatant from each well was transferred to another plate containing 100µl in each well.
18. These were then put into a -20°C freezer and left either for 1 hour or overnight.
19. The plates were centrifuged at 3400rpm at 4°C for 10min.

20. The supernatant was decanted; the pellet washed with 200µl of 70% ethanol and the previous step was repeated. The pellet was then air dried. The pellet should dry to transparency.
21. Each pellet was resuspended in 10µl of T1.0E. To help the resuspension, the TE was pipetted up and down gently at the base of the well. The resuspended DNA could then be either used straight away for fingerprinting or could be stored for up to 2 weeks at -20°C.

3.2.3.2 Fingerprint digestion

Materials and Equipment

- *Pst*I restriction enzyme (New England Biolabs, 20U/µl)
- Buffer 3 (New England Biolabs)
- 6X Buffer II dye (see appendix 2)

Digestion mixes were made up according to the number of plates that were to be loaded on to gels.

1. The enzyme of use, *Pst*I, was used with its supplied buffer. Premixes of the appropriate volume were first constituted on ice (Table 11)

Table 11

	x1 well	1 plate	2 plates
***Pst*I enzyme**	0.5µl	55µl	110µl
Buffer 3	0.9µl	99µl	198µl
Water	2.6µl	286µl	572µl

2. Then 4µl of this mix was added to each of the wells of the resuspended DNA using a combitip dispenser set to 1 with 0.2ml tips.
3. The plates were covered with plate sealers and agitated gently with a vortex to mix and were centrifuged briefly to 1000g to bring the residue to the bottom of the well.
4. The plates were then incubated at 37°C for 2 hours.
5. The plates were again briefly centrifuged to 1000g and the reaction terminated by adding 2µl of 6X Buffer II dye using a Hamilton pipette.
6. The plate was again centrifuged briefly to 1000g. The plates could then be either stored for up to two weeks at 4°C with plate sealers covering them or used straight away.

3.2.3.3 Gel electrophoresis for fingerprinting

Materials and Equipment

- Each gel was made using SeaKemLE agarose (FMC BioProducts).
- 1xTAE (see appendix 2)
- Promega ladder (see appendix 2)
- Vistra Green (Amersham Life Sciences)

Each gel required 5 litres of 1xTAE. It was ensured that each buffer that was made up to make the gels was also the buffer that went with that gel in the electrophoresis tank. This prevents any contamination and also helped ensure straight running of the gels and to the correct distance.

1. For each gel, 4.5g agarose and 450ml 1xTAE was heated in a microwave to melt the agarose for approximately 5min and allowed to cool for 5min in a cold room at 4°C.
2. The agarose was then poured into an gel bed with the ends sealed with plastic end sealers. A divider was put at approximately half way to separate the gel into two. A 121-well comb was placed in both halves of the gel at the same ends and the gel was allowed to set for at least 3-4 hours. The remaining buffer was also left to equilibrate at 4°C.
3. Once the gel had set, the sealers at each end were removed and the gel was placed into the gel tank and 3-4 litres of the buffer was poured in. The divider and the combs were then removed.
4. All reagents were stored on ice as the Promega ladder is particularly sensitive to temperature changes.
5. To the first well, 0.8µl of marker was added and then to every fifth well (see appendix 2).

6. To the wells in between the markers, 1µl of sample was added, starting from A1 on the plate and working across the row. This was also done for the second lot of wells with another plate.
7. Once all 121 wells were loaded the gel was run at 90V for 15 hours at 4°C.
8. Once the gel had run, the gel, in its gel mould, was removed from the tank. The gel and mould was placed into a suitable plastic tray, placed onto a shaker then stained with Vistra Green.

    5ml Tris (pH 7.4)

    500µl 0.1M EDTA

    Made up to 500ml with ddWater

    50µl of Vistra Green

9. This was poured over the gel and it was checked that the gel was floating in the stain to ensure even staining of the gel. The gel was covered with a tray to prevent light exposure and the whole thing was stained for 45min at a very slow shake.
10. The gel was drained and then rinsed with 500ml of water, drained and this was repeated to remove any excess stain. The gel was then ready to be scanned into the computer.
11. The scanning was carried out using a Phosphoimaging Scanner and bespoke software.

3.2.3.4 Data Capture

Data from the gel image was processed from the scanner and stored in a computer database. The database is a network accessible database that is part of the editing program called Image (http://www.sanger.ac.uk/Software/Image/). Image is a program of algorithms for processing gel images of restriction fingerprints. The raw information was then used to continue with the analysis.

Firstly, the captured image was visualised on screen within Image (Figure 21)

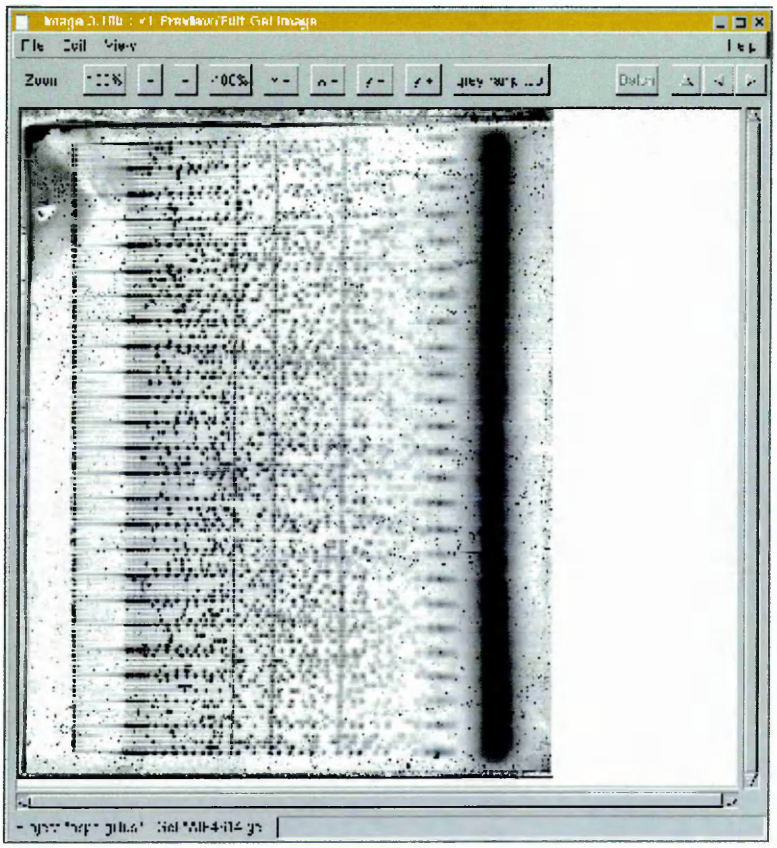


Figure 21. This is an example of how a gel image looks at the first step of the image analysis before any manipulation.

Image takes the normalized banding patterns from each clone (lane) on a gel image and the program performs several procedures in turn. Firstly, the lane tracking has to be performed. The program has a facility to recognise the lanes itself and can superimpose a grid over each lane on the gel image. This is so the program can recognise what is a band or clone. It does this for all 121 lanes on a gel (the 96 wells from the plate and the 25 marker lanes) (Figure 22)

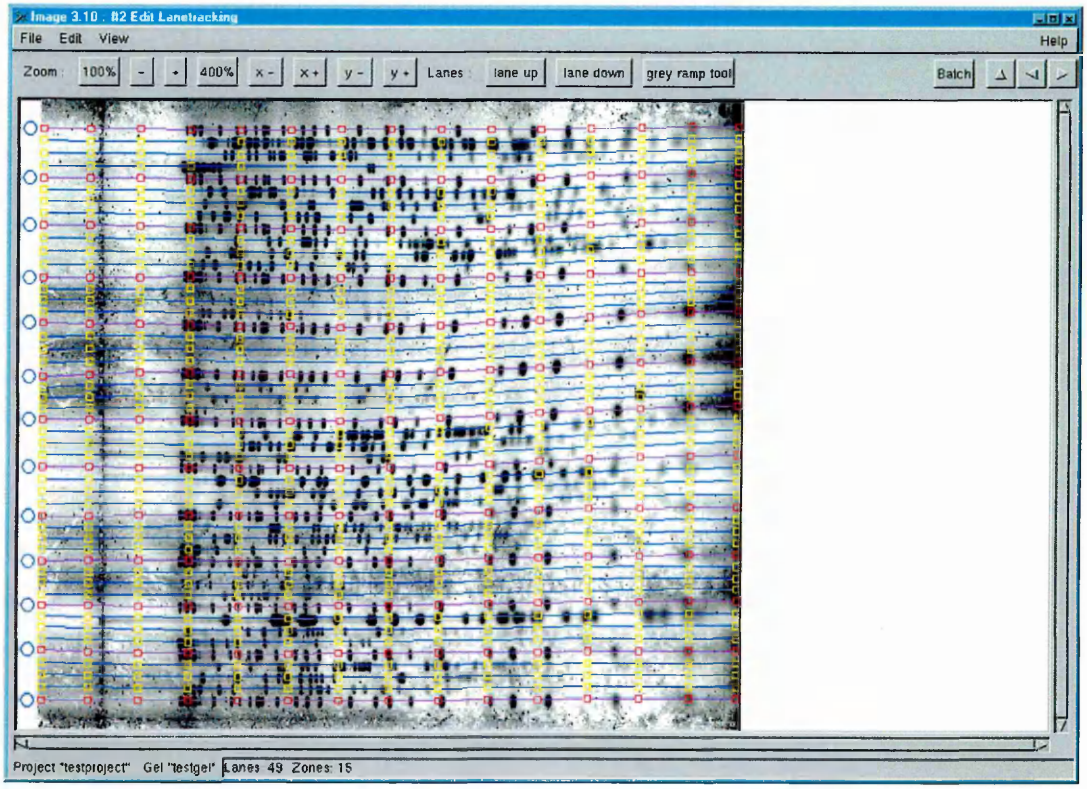


Figure 22. Here the test image has had the lanes tracked by the computer. The tracks in red show where the marker lanes are. The very small, indistinguishable bands have been removed from the image.

Once the lane tracking had been completed, the next step was to call the bands. The program looked along the tracks that has just been laid over every lane of the gel and

looked for bands by making comparisons with the background and foreground contrasts. This step required frequent human intervention, as there were always artefacts on gel images that could not be completely eradicated. If something that was not DNA was stained and illuminated when being scanned, the scanner picked up the artefact and processed it with all the other band information. The gel image would transfer these artefacts and they would have to be eliminated by eye, as the program would recognise them as bands (Figure 23)

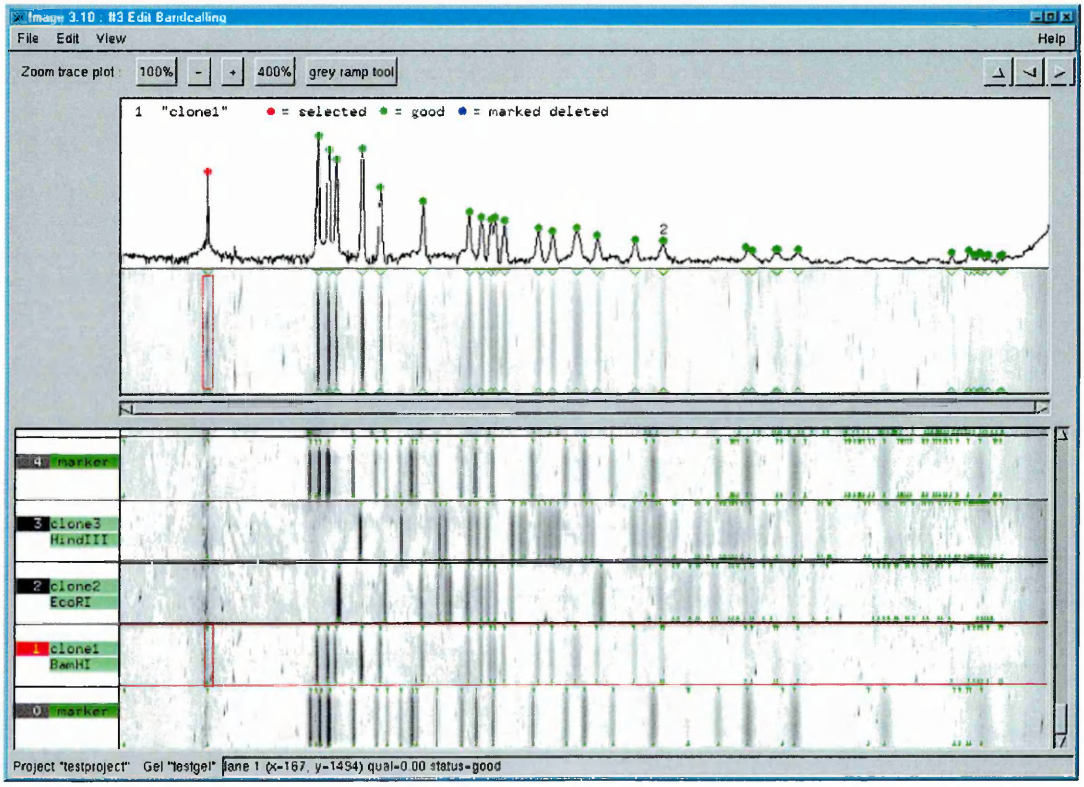


Figure 23. A screenshot of a test gel representing the band calling procedure.

The next step was to lock the markers for the program to have a reference to measure the band sizes. In order for the program and me to compare banding patterns from different

gels, the band positions had to be normalized to one "master "gel. DNA marker fragments of a known size were loaded onto the gel at every 5th lane, with the samples in between the markers. This pattern of bands was then matched to the "master" pattern. The program was then used to get the best match between two sets of bands between the master pattern and the banding pattern of the gel under analysis. (Figure 24)

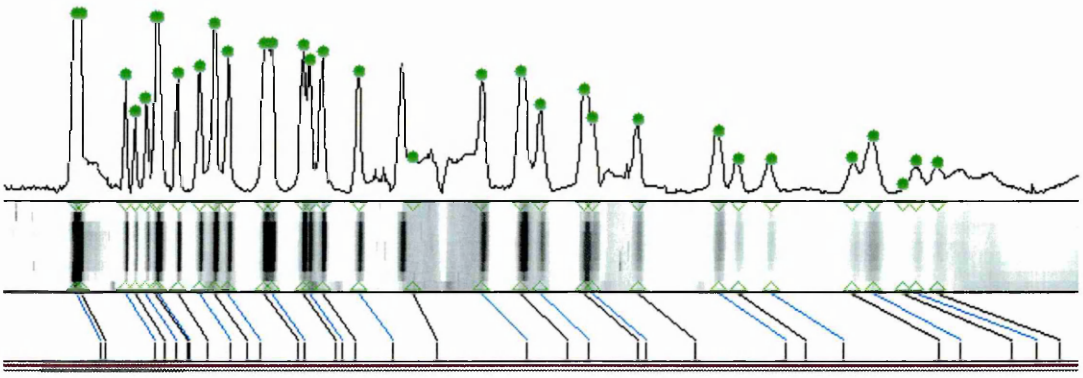


Figure 24. Here the standard lane is "locked" with all the bands in the correct pattern and number. This is compared against a "standard" pattern within the program to ensure that the markers on all of the gels are standardized to the "standard" pattern. Human intervention is inevitably needed at this stage to ensure that there are no bands missing, or there are no extra bands that the program may have called. The upper part of the figure shows the central trace and the bands in the marker lanes. The triangles in the lower part of the plot corresponding to the bands in the standard lane.

Once the marker lane patterns were locked onto the standard lane, then band positions of the samples were then normalized against the locked marker. Each gel would then appear to have been run on the "master" gel with all the distortions due to temperature

differences, voltage discrepancies and so on metered out. The program was used to calculate the correct sizes for the bands by comparing the distances between the locked bands of the normalized marker and those of the sample banding pattern.

Once the bands had been sized in Image, the data were then processed into a database called FPC, which is an automated comparison tool that clusters the fragments from the clones together under certain stringency conditions. FPC uses more algorithms to automatically join clones into contigs by using a scoring system based on the probability of similarity between clones. For each contig, FPC created a "consensus" band (CB) map that had a lot of similarities to the restriction map, but had none of the errors that a restriction map can introduce. The CB map was used as a template to assign coordinates to the clones based on map alignment and generated a detailed visual picture of the clonal overlaps. I used the tools within FPC to adjust coordinates to fit and also remove any badly fingerprinted clones from the system. As new clones were fingerprinted, FPC also had the function of being able to integrate the new data into the system, ready to be used when needed. Contigs could then be split, merged and deleted as the analysis continued (See results). Clonal markers could also be displayed in the program, along with the appropriate contig that the clone was in, making matching the clones much easier. All of the data from the fingerprints was entered into the FPC database and the construction of the initial assembly started using the autonomous facility of the program. This was fine tuned to make the assembly as consistent and complete as possible before undertaking the manual work that was required to complete the task.

The sequence from individual clones as they were sent for sequencing after initial identification of overlaps with other clones was completed could be viewed in a program called Gap4. This allowed the identification of overlaps in clones and was especially useful when the initial clones to start contig building were identified.

3.3 Results

3.3.1 Identification of *niaD* gene

Probes were generated and hybridised to the filters, as described in 3.2.1.1. After overnight exposure the film was developed and the image was analysed. 5 spots were illuminated on the film. The "hits" were designated 5C11, 8D5, 4H9 33H12 and 34H12 (Figure 25). One hit (33H12 and 34H12) was slightly ambiguous as it was not clear that the hit was covering either one or two spots and whether it really was two spots very close together, or just an anomaly in the processing. The information was clarified by performing a colony PCR to confirm the hits were the correct clones. This was also to ensure that the probing had actually worked correctly and that the hits were not just artefacts on the grid.

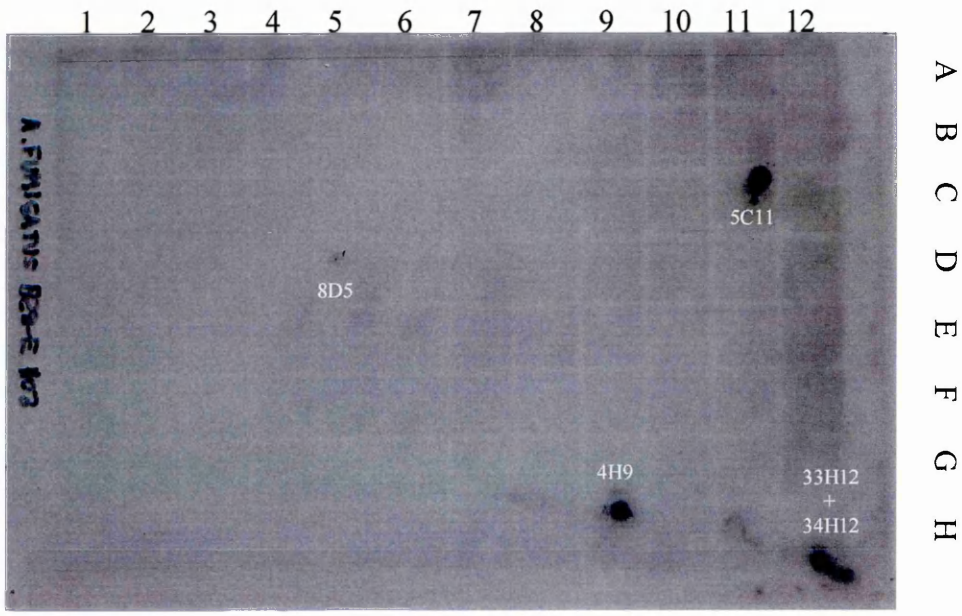


Figure 25. Although the illuminated spots can be clearly seen, they are a little diffuse. This becomes less of a problem when the original filter is placed underneath so "scoring" can be performed. The filter spots can then be seen through the film and the spots can be matched to those on the filter.

3.3.2 Colony PCR

Colony PCR was carried out according to the method set out in 3.2.1.2.

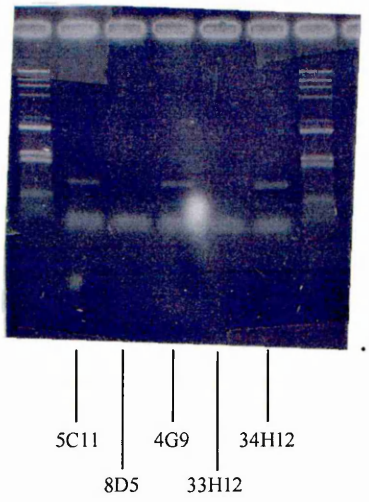


Figure 26. The bands produced from the colony PCR correspond exactly to the spots on the filter. The strong spots on the filter gave by far the best products from the PCR. Of the 5 possible hits, 3 were positive.

If the colony PCR gave bands of good clarity and the correct size, then it could be assumed that the array probing actually worked and that the bands could be confirmed as the *niaD* gene. As can be seen from figure 26, three out of the five hits tested showed positive and strong bands. When building contigs, 5C11, 4G9 and 34H12 were strong candidates to "walk" from. The clone 8D5 only gave a positive on the array probing and not on the colony PCR, but although the hit was not a strong one it was significantly visible, so on this evidence alone, it was sent for sequencing anyway, along with the other three hits. Hit 33H12 would need further confirmation and is discussed in a later section.

3.3.3 End sequencing

The radiation hybridisation identified three clones that were candidates for a starting position and these clones had the end sequences compared against the rest in the database and the best candidate was picked on the grounds of greatest overlap and closest sequence homology. Using the BAC end sequences, the strategy was to "walk" from the starting BAC in either direction to construct contigs that way. Selected BACs were either skimmed (minimal sequencing to generate sequence for use in clone location) or "shotgunned" (further random shattering of the BAC inserts for ease of sequencing and put into pUC vector to create a "mini-library" of that clone) to 8-fold coverage for finishing (knowing the lengths of the selected BACs, by fingerprinting, was helpful in deciding how many plates to shotgun sequence). The fingerprints would then confirm that the clone to be sequenced was correct from the size information and the banding homology. Of the 6912 reads in the database that were produced from the end sequencing ((36 plates x 96 wells per plate) x 2 for read pairs), there were 5685 useful reads. Of these 5685 reads, it could be seen that 4468 reads were actually pairs. This left 1217 reads unaccounted for. These reads were seen as either single reads (there were approx. 300 single end reads) that were still of some use as they were designated to one end of a particular clone, or that they were repeats that were of no use at all.
Which meant that from the 3456 read pairs that were predicted, 2234 read pairs were of use. This is 64.6% of the total complement from the library.

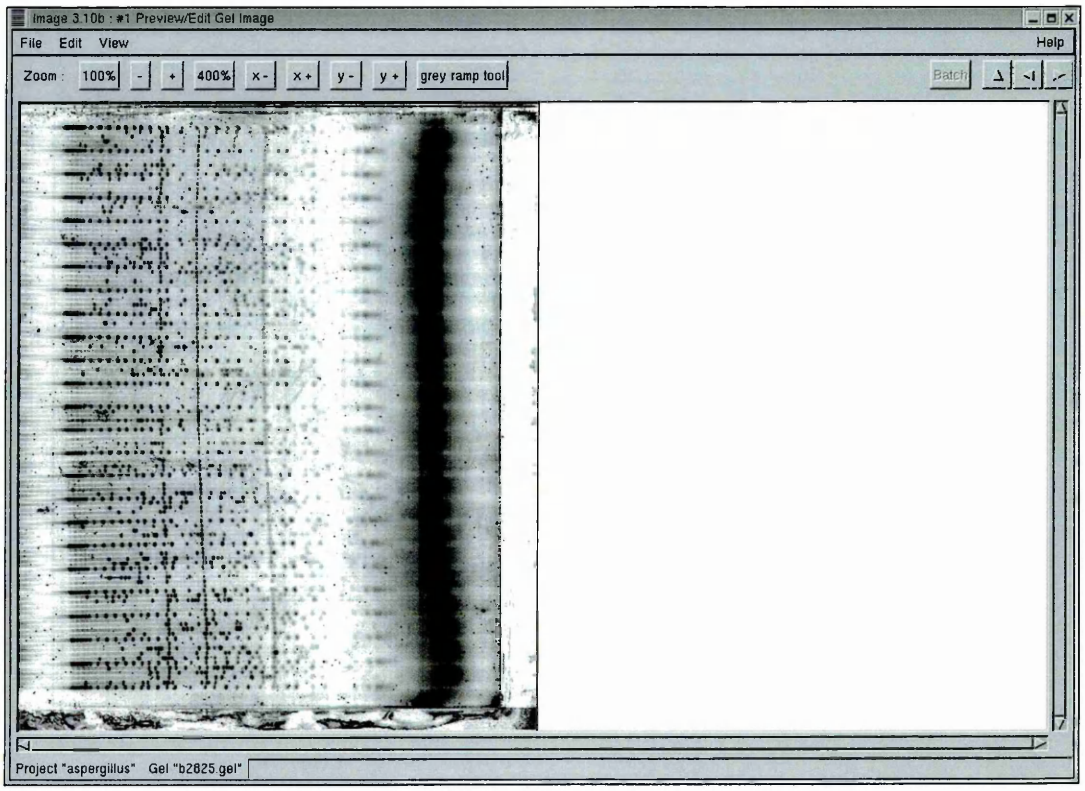
3.3.4 Fingerprints

The fingerprint data were first to be edited. Fragments on the gels that were less than 600bp were discarded as they gave false indications of information to the database and the database would predict very low estimates of clonal overlaps. Another problem was the incidence of "doublet" bands, where two bands of similar or exactly the same size travel though the gel at the same rate and are in approximately the same position when the gel run is over. There was a need for reliability in determining the correct overlaps between clones, so registering them as only one band in the fingerprint effectively eliminated the "multiple" bands. The reliability was increased accordingly.

There also had to be a tolerance threshold to accept two bands from different clones as being the same. If the stringency is too high then bands that are very close on the gel might not be seen as the same band size. Too low and bands on the gel that have high distances between them will be seen as the same band when they are clearly not. The stringency was a fairly fluid parameter due to the high number of variables that can happen with the gel runs. This had to be taken into account when collating this data in the database and the stringency was adjusted at all times, within its own set parameters, to allow discrepancies in gel run distances, as well as manually putting clones together when the program has made obvious mistakes. These parameters also were applicable to the datasets that were in the FPC software and data manipulation described in section 3.3.5 (Figure 27).

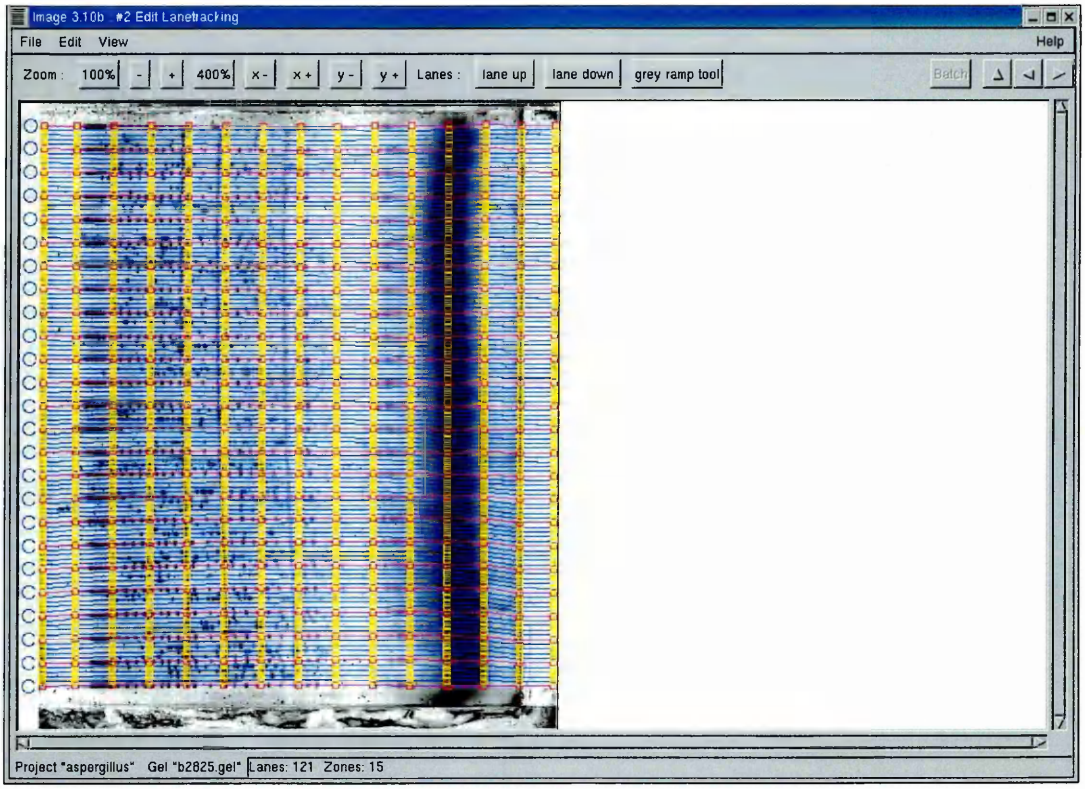
Figure 27. Image of one of the gels generated at different stages of the process. This gel is designated B2825 (B28 is the library designation and 25 is the plate number)

a) Image capture



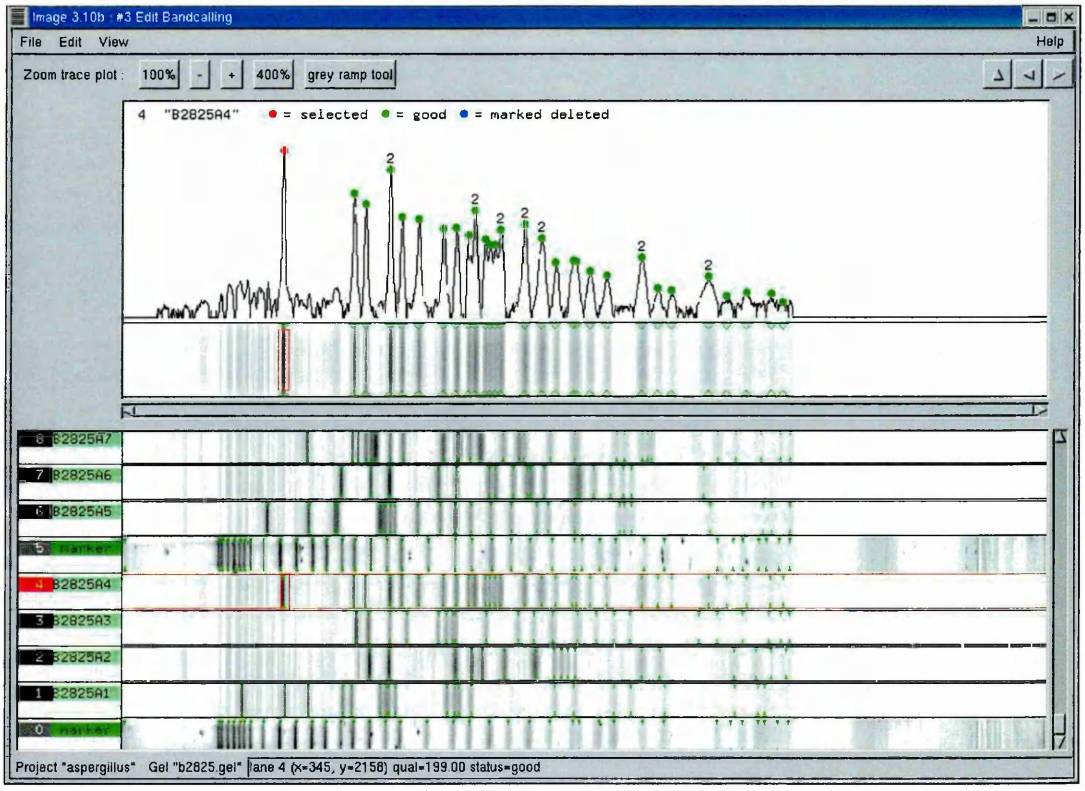
With this image, indistinguishable fragments towards the bottom of the gel are visible. In the analysis, these fragments would be discounted. The gel has run relatively straight, but some lanes are still not completely straight and this needs to be compensated for.

b) Lane tracking



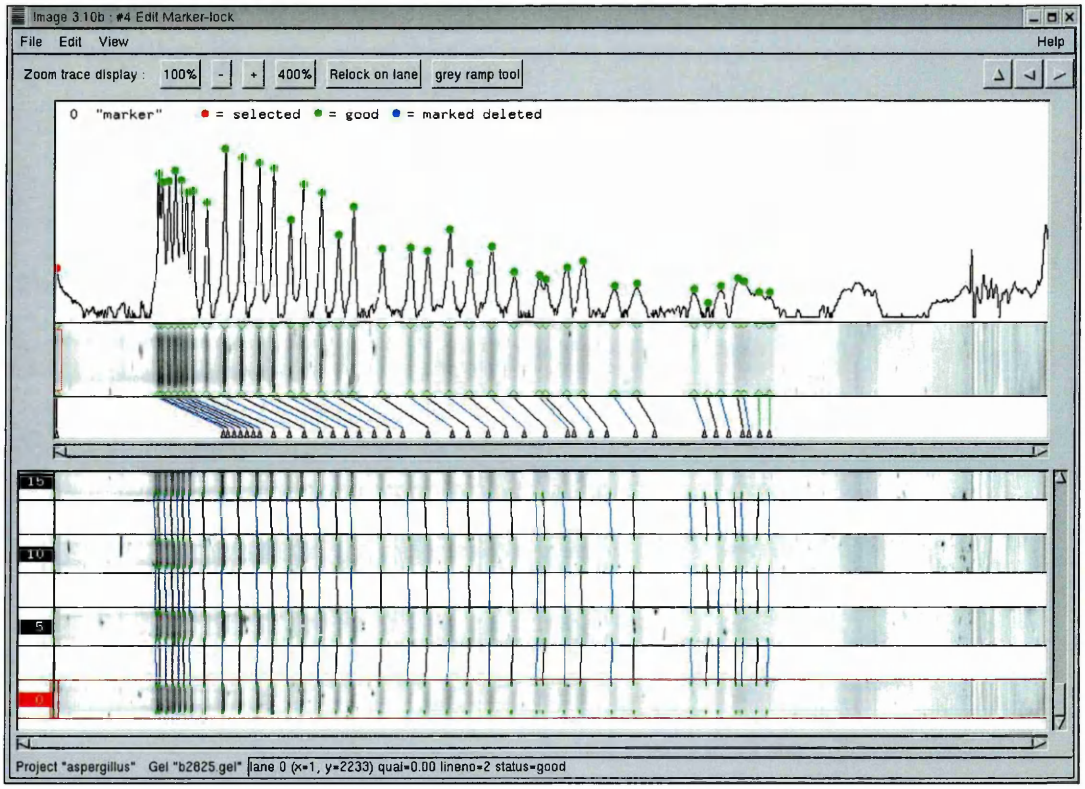
Here, the lane tracking has been completed. Each lane has been locked onto by the imaging software and has picked out what it thinks are the correct lanes. Some manual intervention is required the majority of the time at this stage as the software does not always get it 100% right every time. Some track may be off the middle of the lane or may well have kinks in them due to artefacts that are on the gel showing up more brightly and the software detecting it as a band. The tracks can be moved by the operator to give a more accurate representation.

c) Band calling



This image shows the band calling step. The software detects each band by comparing the contrast of background and foreground. Unfortunately, this can lead to some artefacts being picked up, adding extra "bands" or can also lead to some bands missing if the contrast is too faint. Again, operator intervention is required to make sure all the correct bands are called and no artefacts are picked up.

d) Marker locking



This image shows the last step of the editing process before the image is stored for the information to be used at a later date when comparing clones for contig building. Here, the marker standard is being locked so the software knows the correct distances to call the clone bands to. There should be a total of 37 bands for the marker. For each of the bands, it is up to the operator to make sure that the peaks correspond to the bands on the image and make sure they are accurately picking up the correct band. For the first 30 bands this needs to be very accurate, as this is where the majority of the clone bands migrate between on the gel. The accuracy needed for the last seven bands is not so important as the software rarely uses the information from fragments that small.

As with the end sequencing, all 3456 clones were fingerprinted from the AfB28 library. Assuming an insert size of 74kb and an estimated genome size of around 30-35Mb, then this would give genome coverage between 7.5 and 8 times. The pilot project proposed to only do 1Mb, so the overall coverage was more than adequate.

In total, the B28 library had a total of 2635 fingerprints put into the database. But this was combined along the way with a further library, AfB46, which was needed to fill gaps and enhance the data. This was not the original intention, but as the second library had been constructed, it was thought that it might be valuable to integrate its data. Therefore, figures for the B28 library alone are difficult to obtain. The combined data, made up in the majority of the B28 library gave the following:

104 contigs in total (4180 clones)
1507 Singles (clones not in contigs)
Total genome length 30.032Mb
Longest Contig 1.786Mb

30 BACs were picked in total in the project and 16 of these were "shotgunned" (i.e. broken down again into smaller fragments and sequenced again to attain a higher level of detail) to at least 8-fold coverage and 14 were "skimmed" (i.e. Sequenced to a level so detail could be seen, but not completely. Used only as an indicator as to whether it is worth sending the clone for full "shotgun" sequencing. Usually done on the number of plates, not overall coverage of the clone). (see Supplementary Figure 1).

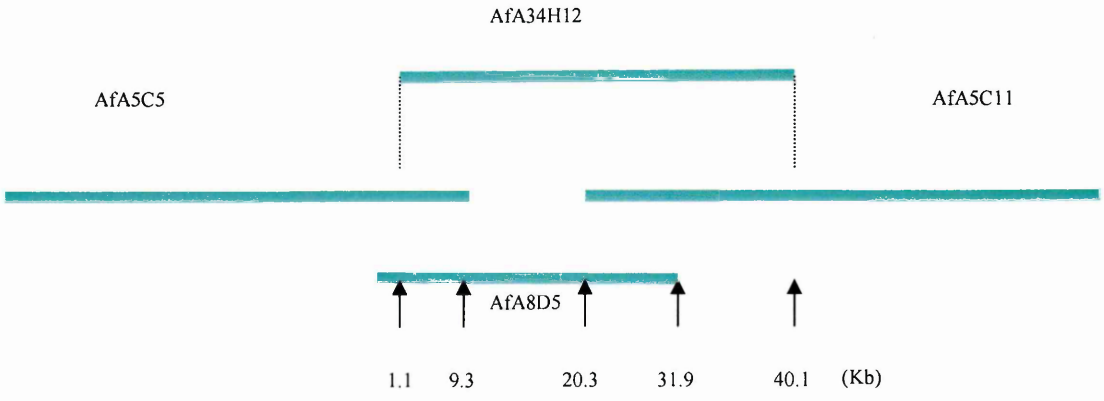
A total of 166x96 well plates were sequenced and the total length of the 30 shotgunned BACs was 1880500bp. 31872 (166x96) attempted reads gave 25500 passed reads (at 80% pass rate) and a total of 12750000bp (at 500bp/read). This works out at an average of 6.8-fold coverage of 1880500bp.

The finished status was 16 finished BACs that cover a region of 921536bp. The aim was to cover 1Mb.

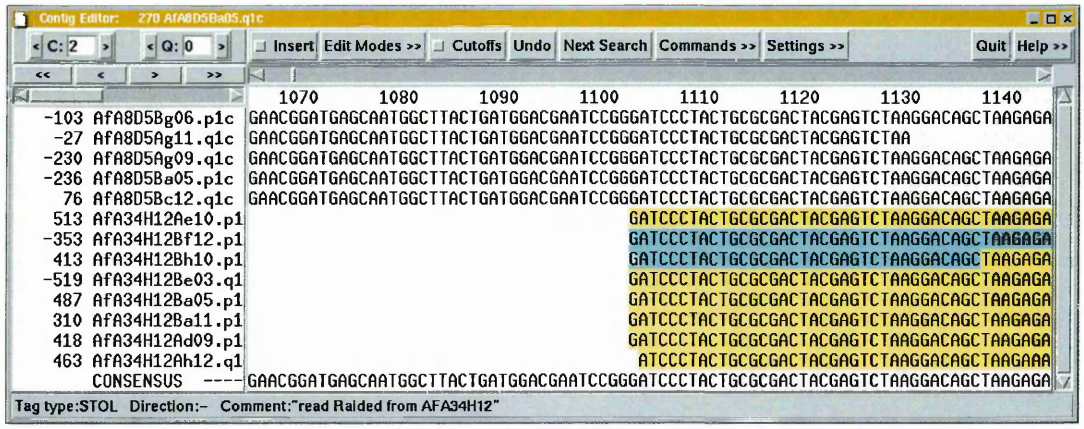
3.3.5 Data capture

When the sequence was viewed in a sequence editor (3.2.3.4), it could be seen that all four of the hits had significant overlaps with each other, showing that 8D5 also had overlaps (Figure 28). Supplementary Figure 2 in the next chapter also shows where in relation to the clones indicated below the *niaD* gene is.

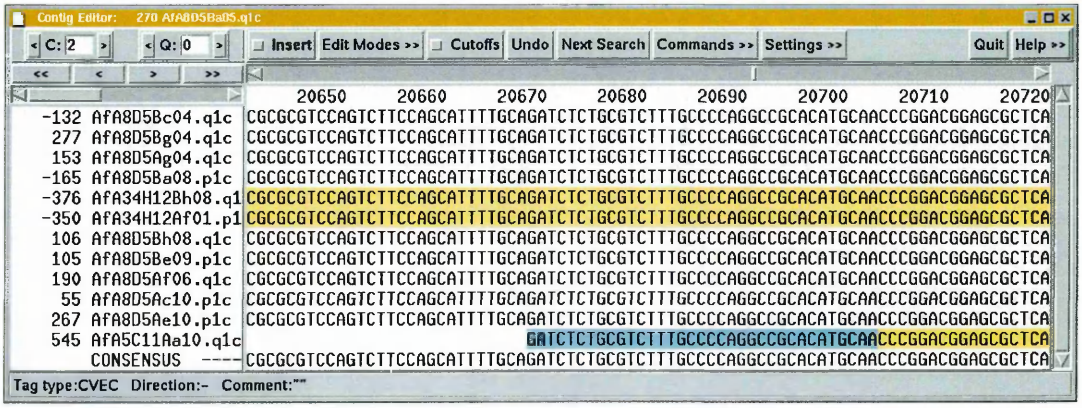
a)



b)


Contig Editor: 270 AfA8D5Ba05.q1c
Insert
Edit Modes >>
Cutoffs
Undo
Next Search
Commands >>
Settings >>
Quit
Help >>
1070 1080 1090 1100 1110 1120 1130 1140
-103 AfA8D5Bg06.p1c GAACGGATGAGCAATGGCTTACTGATGGACGAATCCGGGATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
-27 AfA8D5Ag11.q1c GAACGGATGAGCAATGGCTTACTGATGGACGAATCCGGGATCCCTACTGCGCGACTACGAGTCTAA
-230 AfA8D5Ag09.q1c GAACGGATGAGCAATGGCTTACTGATGGACGAATCCGGGATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
-236 AfA8D5Ba05.p1c GAACGGATGAGCAATGGCTTACTGATGGACGAATCCGGGATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
76 AfA8D5Bc12.q1c GAACGGATGAGCAATGGCTTACTGATGGACGAATCCGGGATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
513 AfA34H12Ae10.p1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
-353 AfA34H12Bf12.p1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
413 AfA34H12Bh10.p1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
-519 AfA34H12Be03.q1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
487 AfA34H12Ba05.p1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
310 AfA34H12Ba11.p1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
418 AfA34H12Ad09.p1 GATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
463 AfA34H12Ah12.q1 ATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAAA
CONSENSUS ---- GAACGGATGAGCAATGGCTTACTGATGGACGAATCCGGGATCCCTACTGCGCGACTACGAGTCTAAGGACAGCTAAGAGA
Tag type:STOL Direction:- Comment:"read Raided from AFA34H12"

c)


Contig Editor: 270 AfA8D5Ba05.q1c
Insert
Edit Modes >>
Cutoffs
Undo
Next Search
Commands >>
Settings >>
Quit
Help >>
20650 20660 20670 20680 20690 20700 20710 20720
-132 AfA8D5Bc04.q1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
277 AfA8D5Bg04.q1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
153 AfA8D5Ag04.q1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
-165 AfA8D5Ba08.p1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
-376 AfA34H12Bh08.q1 CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
-350 AfA34H12Af01.p1 CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
106 AfA8D5Bh08.q1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
105 AfA8D5Be09.p1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
190 AfA8D5Af06.q1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
55 AfA8D5Ac10.p1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
267 AfA8D5Ae10.p1c CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
545 AfA5C11Aa10.q1c GATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
CONSENSUS ---- CGCGCGTCCAGTCTTCCAGCATTTTGCAGATCTCTGCGTCTTTGCCCCAGGCCGCACATGCAACCCGGACGGAGCGCTCA
Tag type:CVEC Direction:- Comment:""

d)

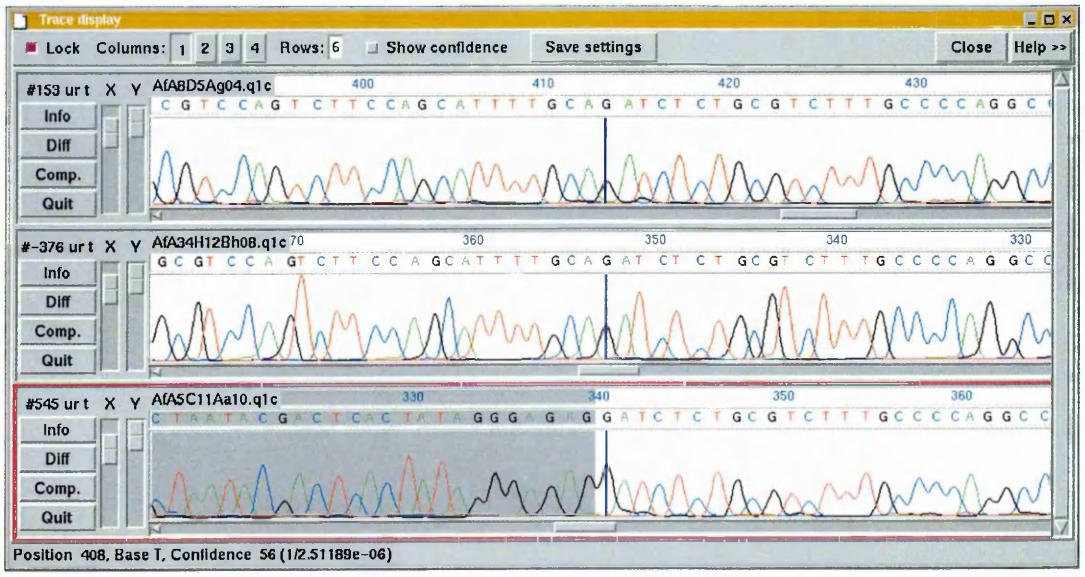

Trace display
Lock
Columns: 1 2 3 4
Rows: 6
Show confidence
Save settings
Close
Help >>
#153 ur t X Y
AfA8D5Ag04.q1c
400 410 420 430
Info
Diff
Comp.
Quit
#-376 ur t X Y
AfA34H12Bh08.q1c
370 360 350 340 330
Info
Diff
Comp.
Quit
#545 ur t X Y
AfA5C11Aa10.q1c
330 340 350 360
Info
Diff
Comp.
Quit
Position 408, Base T, Confidence 56 (1/2.51189e-06)

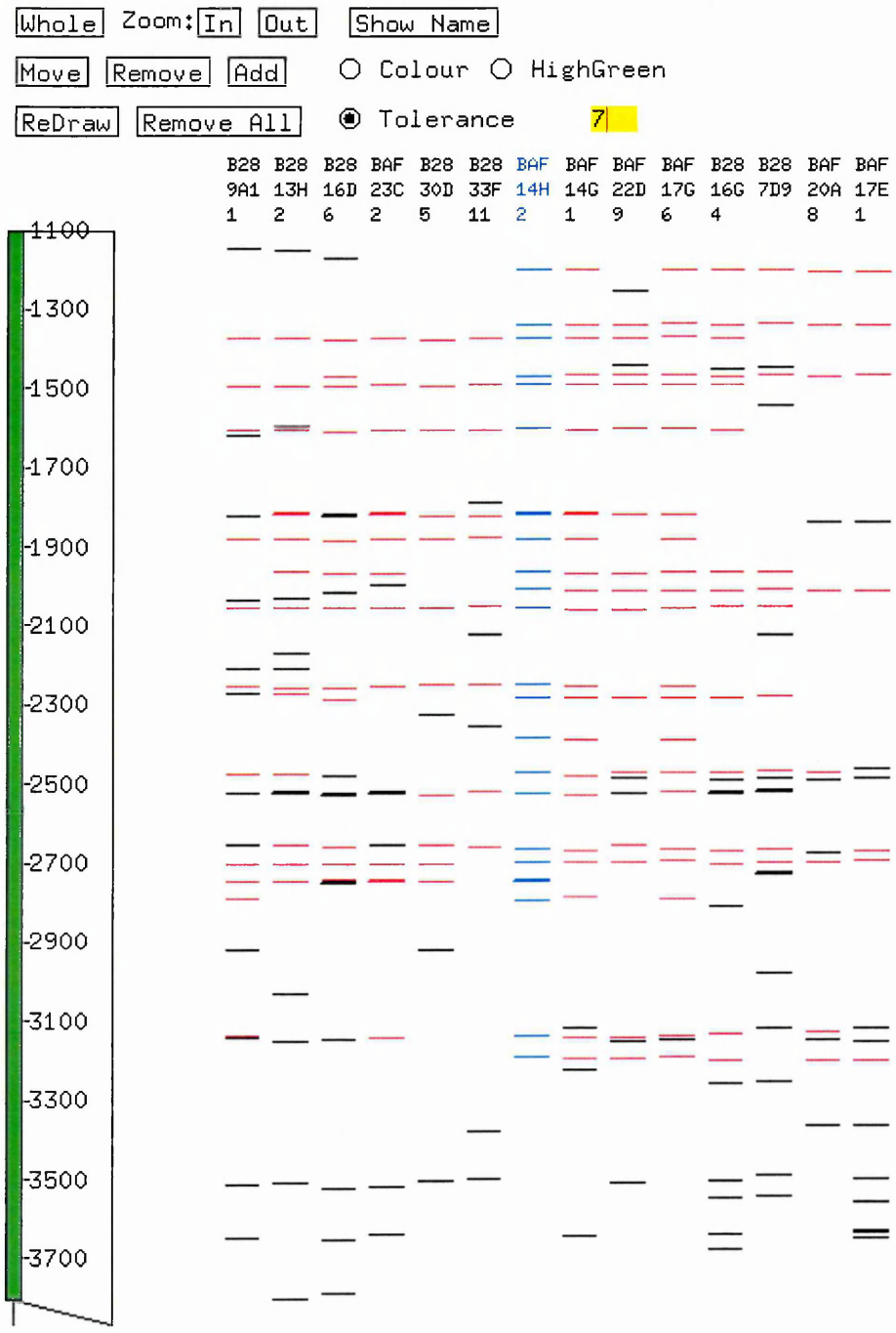
Figure 28. This shows the sequence overlap between the clones 8D5, 34H12 and 5C11 in a viewing program called Gap4 (Bonfield *et al* 1995). a) This is a graphical representation of the overlaps and their sizes between the clones that were sequenced from the original four. Note that one of the clones, 4G9, is not there. In its place is 5C5, which came from BLASTing the sequence against the end sequences in the database that had already been completed for the library. 4G9 turned out to be a sequencing failure in that region, but 5C5 was a match. The numbers under the arrows represent the number of kilobases from the left end of the clone 8D5 that the overlaps occur. b) This shows the sequence alignment between 8D5 and 34H12. The empty gap represents the cloning vector and where the sequence of the insert starts. It can be seen when in comparison with a) where the overlap occurs. The figures at the top of the window show the number of kilobases from the left end of the clone, as in a). c) This shows the sequence alignment at the other end of the clone where 5C11 starts the overlap. Again, with comparison to a), it shows nicely the distance from the left end of the clone and the good sequence alignment confirmation. d) These are known as "traces" which are the raw data that is produced by the sequencing machines. The peaks and colours represent the four different bases. Here, it can be clearly seen that there is good sequence matching with all three clones.

As can be seen from the images above, 4G9 clone was not seen in the final sequence alignment, but another clone, 5C5, was found to match the closest. 5C5 was found by comparing its end sequence from a BLAST search of an internal database which had all of the end sequences available on it and pasting in sequence from three of the clones to see if there were any matches in the database. 4G9 did not have any matches, but 5C5 did. This was the clone that was used to continue the work.

FPC was used to build the contigs. There were a number of clones that were initially not included in the original build, but after careful manipulation of the data and stringencies, some of these clones were incorporated into the build. Figure 29 shows and example of

an intermediary build with a number of clones. Once this build with this group was done, it was joined with other contigs that were built at overlaps that were identified by either the software or myself.

a)


Whole
Zoom: In
Out
Show Name
Move
Remove
Add
Colour
HighGreen
ReDraw
Remove All
Tolerance
7
B28 9A1 1
B28 13H 2
B28 16D 6
BAF 23C 2
B28 30D 5
B28 33F 11
BAF 14H 2
BAF 14G 1
BAF 22D 9
BAF 17G 6
B28 16G 4
B28 7D9
BAF 20A 8
BAF 17E 1
1100
1300
1500
1700
1900
2100
2300
2500
2700
2900
3100
3300
3500
3700

b)

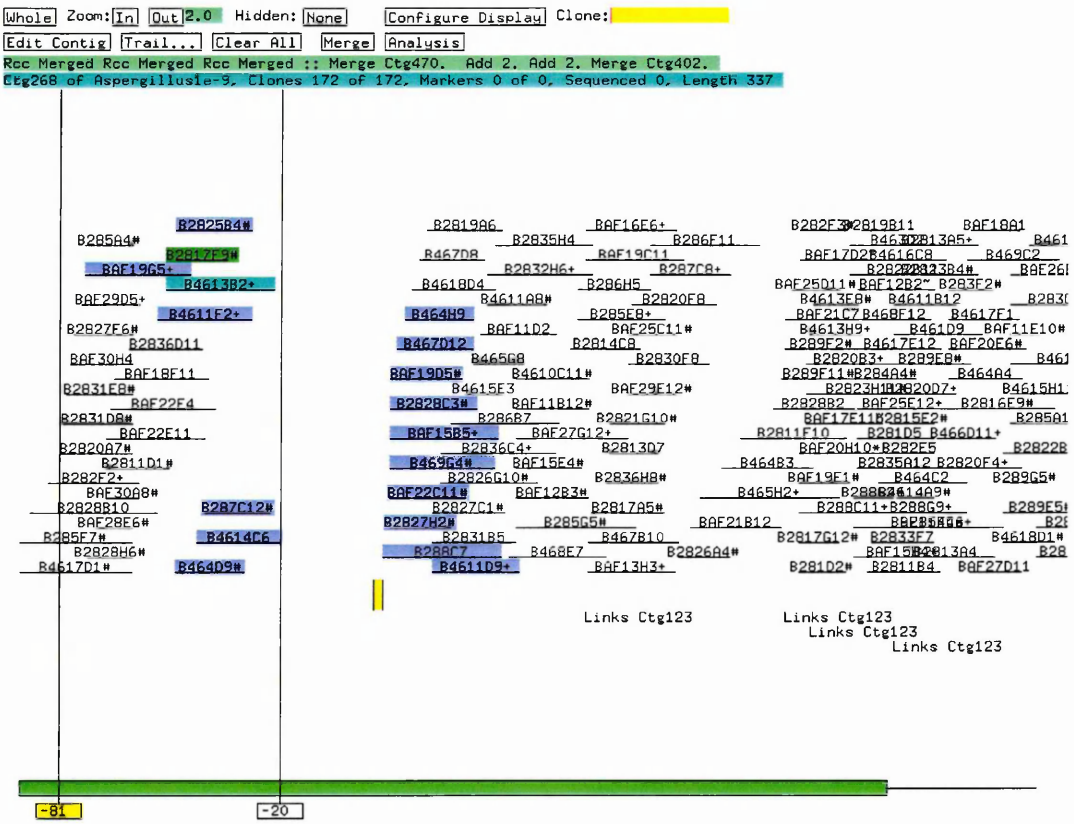


Figure 29. a) This screen shot shows a number of fingerprints side-by-side so visual comparison can be made.

b) The fingerprint in blue is the clone that is being compared to the fingerprints the computer estimates is closest in match. Screen shot from FPC showing possible merge of a contig. Clones in purple are overlapped.

The data from the fingerprint gels was visualised in the first part of the FPC software (Figure. 29 (a)) by lining up the banding of each clone side by side to compare the sizes of each band in that clone. The software then determined the overlaps in the clones by the number of bands that were in common with each clone. The higher the number of bands

in that were the same size, the higher the likelihood that the clones overlapped. It was inevitable that although a number of clones would have overlaps the software would not pick up the connection of the band number between the clones because the bands were not sized *exactly* as they should be. Hence stringencies had to be altered to allow for any running error on the gels that would stop the software from recognising the banding sizes. Once the clones had overlaps, the contigs could then be built in the second part of the software. This linked the clones by overlapping them by using information from both the end sequences in the database and the clones that overlapped from the FPC software. By looking at both sets of information, the clones could contiguated as seen in Figure 29 (b). The software and I combined the data generated so that "best-fit" overlaps could be made. Each type of information was used against the other to ensure that the clone that was thought to overlap was actually the correct clone.

3.4 Discussion

The identification of the *niaD* gene using the probes and the fragment supplied turned out to be a relatively easy process. The only complication that arose was after the clones identified were sent for sequencing. One clone, 4G9, did not show in the database as having overlaps with the other clones, even though it had been detected on the array and also by colony PCR. This may have been due to 4G9 being a very small clone and although detectable on a hybridisation experiment, the sequencing did not produce enough data to give positive matches. It is also possible that 4G9 was still in the *niaD* region, but not directly linked to the other clones in the gene i.e. either to the left or to the

right of the other clones in the gene, but with no overlaps. 5C5 on the other hand showed up in the database as having significant overlaps with the other clones. The identification of this clone failed in the array probing and this was possibly down to a failed PCR, hybridisation or that the clone only had a small amount of sequence that was compatible with the probes to bind to but had high levels of overlap at the opposite end. The reason could be identified more clearly at a later time when there is more information in the databases and the clones could be placed more accurately.

The colony PCR also gave good confirmation on the hits on the array. The technique did clear up the confusion of the ambiguous hit of 33H12 and 34H12. When both were tested, it was easy to see that 33H12 was not a positive and proved that there was only one hit in that area and the confusion was brought about by highly dispersed radioactivity from the 34H12 hit in that area. The labelled probe had obviously not adhered completely to 34H12 and had spread too far into the surrounding area, which exposed the film more and made it look like there was the possibility of two hits instead of one. The would not have been too much of a problem once the clone had been sent for sequencing as it would have been immediately apparent that the clone did not have any overlaps with the other hits, but it would have been a waste of resources and time to have let the process go this far for one clone. Luckily, the colony PCR technique is usually reliable and identified that the hit was in fact false and prevented any further work from being done on this clone at this stage. But as proven above, the reliability of the technique is still not 100%, as it identified and confirmed the 4G9 hit which consequently did not have significant overlaps with the other clones. But for the purpose of identifying clones from an array, it was a very useful tool.

As discussed in section 3.3.3, the number of read pairs produced from the end sequencing that were of use was 64%. Although this figure is not optimal (other BAC end sequencing projects aim for 80-85% pass rate), it was still more than useful as the read pairs there were of confirmed good quality. This confirmation is done on a scoring system produced by the software indicating the quality of the read pairs that have come through the system. Also, if the sequencing reactions have just not worked, then there will be no data to score. As the end sequencing was relatively new on a project like this, there were initially some mistakes in database nomenclature that may account for some lost reads, some miscalculations from reads and so on. This itself bought down the overall pass rate. But what was important was the quality of the sequence from the reads and this was of a very high quality, ensuring that the lower pass rate would not be too much of a problem. As it was a 1MB contig that was to be sequenced and not the whole genome at this point, then a 64% pass rate was more than enough to ensure coverage for a 1MB contig construction. Ideally, all of these would have passed and each read would be a clean, mistake-free sequence covering every clone in the library. But this is not so and there are inevitably some failures.

As with the end sequencing, all 3456 clones were fingerprinted from the AfB28 library. Assuming an insert size of 74kb and an estimated genome size of around 30-35Mb, then this would give genome coverage between 7.5 and 8 times. Although the library had smaller inserts than envisaged, it was known that the library was of high quality, with few non-recombinants and relatively uniform insert sizes. This was confirmed by the excision of the insert by *Not*I digestion, showing the good quality of the insert size and also by the

end sequencing, which as already explained gave good results for DNA quality, even if the actual pass rate was less that ideal. The insert size reduction could be due to a number of factors such as a library bias of smaller fragments that were generated when removing the fragments from the gel. The clones were digested and loaded on to gels with 121 lanes. 25 lanes were taken up with "marker" DNA to help determine standards and calculate band sizes. Each plate of the library was digested at once and the whole plate was loaded on to the gels. This was time consuming and laborious, but this was slightly relieved when a revised method of tandem gels was introduced (described in section 3.2.3.3), which halved the time it took to fingerprint the whole library and also did not degrade the integrity of the results regarding the accuracy of the clone tracking or the fragment sizes. The original method only used one gel per plate. There was no division of the gel into two distinct sections and there was only comb. The main problem of this method was the time it took to do each plate. When the one gel/two plate method came in, it became much quicker to fingerprint the whole library. There was an investigation to determine if this would have a detrimental effect on the running characteristics by some colleagues who concluded, after a little manipulation of a few of the minor preparative practices in making the gels that there should be no effect on the results produced by a tandem gel. Before the preparation protocols were changed and tandem gels were tried, the DNA tended to "bend" in the gel, with the outside lanes "drifting" out towards the edge of the gel and usually losing the smaller fragments from the bottom of the marker lanes. As already discussed, the smaller fragments are not that useful, but the tandem gels lost bands from the marker that were of larger size and these bands were required for band calling a sizing. The modifications to the preparation were very small, but effective.

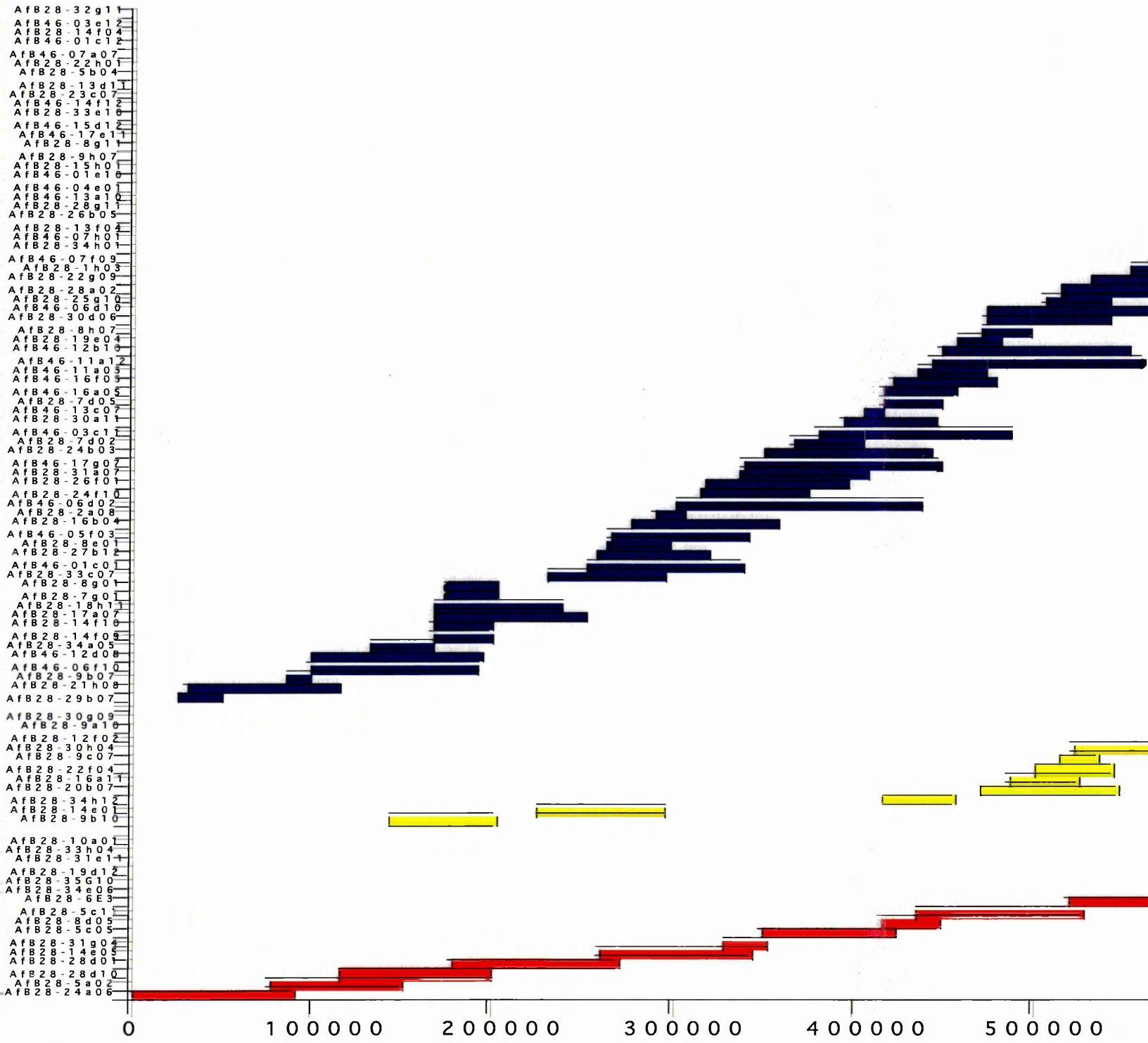
Procedures such as leaving the gel to equilibrate at 4°C in a cold room for at least 2 hours before loading made a difference, probably due to the temperature being equal throughout the gel as opposed to some areas of the gel being warmer or colder than others and affecting the gel properties. Also, making sure the buffers were also equilibrated for a minimum amount of time at the same temperature all had small, but significant effects. Once these small problems were corrected, the time benefits were obvious over a single gel. This meant that we could start on the next processes much more quickly.

The data capture was more or less fully automated, with the software doing most of the work capturing images and interpreting and analysing them before human intervention. The main intervention was at the stage of contig building using FPC. Here, it was the number of single clones that were not integrated into the contigs from the first round of automated building by the software that was a slight problem. Stringencies within the software could be changed manually, which relaxed the guidelines and allowed more of the single clones to be added. This may seem like a haphazard method when the strict guidelines already posed by the software has not included the singles, but when the data from the fingerprints was visually analysed and the data were combined with both the end sequencing data (which was more useful for the walking) and the BLAST searches, which compared the sequence data with other clones in the database, an overall picture of the contig could be built up which integrated the single clones into the contig. The confirmation that the clones were the correct ones was the data from the end sequencing, which compared the ends of each of the clones and indicated that my intervention had actually picked out the correct clones.

The physical map is and was a very useful tool in regards to deciding which clones were to be used for sequencing. A number of techniques were used to reduce the eventual number of clones that needed to be sequenced to cover the entire region to a minimum of 16 clones. Overall, it was a painstaking process, but in hindsight, probably the most efficient and accurate way of confirming the clones. Supplementary figure 1, at the end of this chapter, shows the breakdown of how it was achieved. The original end sequenced clones marked out the clones and put each in relation to the next clone correctly by overlap of similar sequence. The choice of a starting point was done using the hybridisation of a known gene, *niaD*, and then using this to find the corresponding clone in the library. Each clone was "walked" out using the end sequencing data and further sequencing from each end of the original *niaD* containing clone and the map was slowly generated. Once a certain "tiling path" had been determined, each clone was sequenced much more accurately and the clones put together by their overlaps. This then reduced the number of clones covering the contig to 16. Supplementary figure 1 shows the "map" of the contig with the levels of coverage that were needed to obtain the final 16 clones. The map can and will still be used in the future as a good starting point for the whole genome sequencing project, as well as being useful for researchers who may want to look at particular genes or regions within the contig more closely. Genes can be identified and then located to a particular clone from the library using the map. The clone can then be studied much more closely.

The sequencing of the clones was done in accordance to the standard sequencing and subcloning protocols carried out at the Sanger Institute and the methods and outlines of

Supplementary Figure 1



The 922kb contig as a schematic. The nucleotide positi
used for sequencing is in red. Yellow BACs were skim
end sequenced BACs used for the map.

this can be found at http://www.sanger.ac.uk/Teams/Team53/methods/methods.shtml and also at http://www.sanger.ac.uk/Teams/Team55. Once the clones had been sequenced, the sequence data was fed into a database that was accessible by a number of programs already described and will also be described in the next chapter.

4. Annotation and the findings of the sequence

4.1 Introduction

The final stage in the project was to annotate the DNA sequence that has been elucidated from the methodology described in Chapters 2 and 3. It is necessary to try and detect mobile elements, pseudogenes, repetitive elements and viral fragments as well as determining the most important regions, the areas of interest. At the nucleotide level, "landmarks" have to be identified such as detection of tRNAs, rRNAs, other non-translated RNAs, repetitive elements and duplications. Additionally, the genetic markers confirmed by experiments have to be found. This gives a "map" of reference points from which to work. Genomic landmarks then have to be identified, with PCR based genetic markers and restriction fragment length polymorphisms being good indicators. The identification of open reading frames is a vitally important step to determine putative genes. Bioinformatic tools can provide context, near species matches and some information on possible function. Some of the programs and databases used in the analysis are outlined below.

Possibly the most difficult part of the annotation is the process level annotation. How does the protein of interest fit into the overall biology of the organism? What is its function? This was a very difficult process to begin with, as there was only comparison with experimental evidence to back up submissions and queries. But this has changed recently with the advent of Gene Ontology (GO). This "database" allows the annotator to pinpoint much more accurately the whereabouts and its function within the cell by allowing searches against other proteins and domains that have already been characterised. This is further described in 4.1.3.

This chapter will give an overview of some of the techniques used to decode the information we gained from the sequencing of the 1Mb of *A. fumigatus* DNA, as well as give an insight into what has been found within that sequence.

4.1.1 Artemis

There are a number of sequence viewers either commercially available or written specifically by various organisations, such as AceDB (http://www.acebd.org), but these more often than not have limitations. Artemis (http://www.sanger.ac.uk/software/artemis) on the other hand is freely available yet has many more capabilities than most visualisation tools. The Sanger Institute has used Artemis for many organisms to date and was custom built for the annotation and display of both prokaryotic and lower eukaryotic genomes, including the complete genome annotation of the malaria parasite *Plasmodium falciparum*, *Neisseria meningitides,* and *Campylobacter jejuni*.

Artemis can be used solely as a sequence viewer. A number of formats of sequence and annotation can be entered into the program including those taken directly from EMBL (European Molecular Biology Laboratory) (http://www1.embl-heidelberg.de/) and GenBank (Benson *et al* 2000) (http://www.ncbi.nlm.nih.gov/) format files. It displays stop codons in all frames, making visualisation of ORFs and possible genes much easier. Display of G+C content, an important identifier of many genes, can be visualised as graphs and charts. But Artemis is mainly an annotation tool and was used to identify multiple or single features for further work with BLAST and FASTA. Artemis is currently the main annotation and visualisation tool used within the Pathogen Sequencing Group of the Sanger Institute and will be continually updated as time and requirements move on. Artemis allowed the manual prediction of genes and identification of unusual features much more easily. It allowed the interactive prediction of gene models whilst allowing the annotator to view the evidence at the same time, which in turn allows the highlighting of genes and exons that have been overlooked by the gene prediction programs. Very large sections of sequence can be annotated and analysed at the same time, which allows the identification of unusual features from rearrangements or acquisitions.

4.1.2 Gene prediction and other tools

Used in conjunction with Artemis were other tools, many web-based, that allowed not just the ability to view the sequence, but to decipher the main features of the sequence displayed. GlimmerM (Salzberg *et al* 1999), Phat (Cawley, Wirth and Speed 2001) and

GeneFinder P (P.Green, unpublished) helped predict genes from the evidence such as G+C content or standard open reading frame views. The results of the outputs of these programs could be adjusted to ensure that the predictions were correct, for example, splice sites. Percentage GC plots were also used in exon prediction, as this was the only tool to use when any of the gene finding programs failed to find a number of them, which could happen on a number of occasions due to the complexity of some sequence regions and the inability of the programs to decipher the code.

To identify genes and to make gene predictions, a number of web based were used. Programs such as :

- TMHMM (http://www.cbs.dtu.dk/services/TMHMM/). Transmembrane proteins are seen as particularly important in that they are the gateways for cell secretions and could be used as drug targets. This program was used to predict transmembrane helices in proteins and gave links to literature on the nature of those particular proteins and their functions. It also helped the annotator designate function much more easily.
- SMART (Simple Modular Architecture Research Tool) (http://smart.embl-heidelberg.de/help/smart_about.shtml) (Schultz *et al* 1998) allowed the identification and annotation of genetically mobile domains and the analysis of domain architectures, especially in signalling and extracellular proteins.

- InterPro (http://www.ebi.ac.uk/interpro/) (Mulder *et* al 2003) is another external program and was used to identify protein families, protein domains and protein functional sites.
- Pfam (**P**rotein **Fam**ilies) (http://www.sanger.ac.uk/Software/Pfam/index.shtml) is a large collection of multiple sequence alignments and hidden Markov models that covers many of the common protein families and domains. This was used to look at multiple alignments and to view protein domain structures. It also allowed the annotator to examine the distribution of a family within species narrowing down the annotators searches.
- ExPASy (http://ca.expasy.org/enzyme/) identifies enzymes and allocates an Enzyme Commission (EC) number. The EC number can be fed into ExPASy, which pinpoints the enzyme and its function whilst the annotation is ongoing.

4.1.3 Gene Ontology (GO)

Gene Ontology is a relatively new process, but is a very useful tool and a good final point to the annotator's main job. To make time consuming function searches easier and less confusing, as well as having a standard from which all work can then be set, the Gene Ontology (http://www.geneontology.org/) consortium was initiated in 1998. It is a collaboration of three different model organism databases;
FlyBase (http://flybase.bio.indiana.edu/), the *Saccharomyces* Genome Database (http://www.yeastgenome.org/) and the Mouse genome Database (http://www.informatics.jax.org/). GO has now incorporated a large number of

databases into its fold, including some of the worlds biggest plant, animal and microbial genome databases.

The collaborators working on GO are developing three structured controlled vocabularies. These represent biological process, molecular function and the cellular component. The biological process component comprises of its function within the cell. A general rule of thumb would say that the process must have more than one distinct step too. The types of functions would go from processes such as cell growth and maintenance to more complicated and focused processes such as pyrimidine metabolism. Molecular function on the other hand describes the function of a particular protein at molecular level, such as catalytic or binding activities. It does not specify when or where the process is taking place in the cell or in what context. It is purely a functional designation. Cellular component is stating that it is part of the cell and is also part of a larger object in the cell, whether that is anatomical (e.g. endoplasmic reticulum) or part of a gene product group (e.g. ribosome). These three categories help scientists by describing the protein of interest in a species independent manner. The vocabularies are structured so that it is easy to look at a certain protein in different ways. For example, you can search the databases for all of the gene products that are associated with signal transduction within the mouse genome or you can target in directly on something much more specific. Depending on what is known on a gene product, annotators can then assign function on a gene product at different levels but with a much greater degree of accuracy. GO is not a series of databases or catalogues nor is it a central unification of all databases in use. It is therefore also not a gold standard and should be seen and used only as a means to quickly and

accurately assign a function to a gene product that has already had a certain amount of analysis performed on it.

GO annotation is usually one of the last processes to be completed before the entire sequence is submitted for viewing. The process can be quite complicated as a number of the terminologies can overlap or a product might have more than one molecular function in one or more places within the cell. But it is still more accurate and a lot less labour intensive doing this than by curating manually and assigning function without knowing the background to that gene product.

4.1.4 Other aspects of annotation

By the very nature of annotation and genome sequencing, annotation is not a finite process. As time goes on, even on supposed completed genomes, the annotations will need to be updated to include any new information that is gathered from various studies upon that organism. Experimental verification will confirm or deny annotation of particular predicted genes and this will have to be identified and corrected. Independent viewers may find mistakes or anomalies in the original annotation that needs to be discussed, agreed upon and confirmed. Even though a publication may go out and the annotation data go live on the Internet, it is still an ever-evolving situation that has to be refined and updated to reflect this situation. Annotation is done in very different ways by different laboratories and scientists, and it can be done either by a single individual or by an entire group or consortium. It can even be done by a whole community interested in a

particular organism. This approach has both good points and bad points. One of the good points is that it spreads the work around which can be especially important on a large genome, such as the Human Genome Project and it can also bring a lot of expertise together. On the other hand, with so many people involved, it can also bring confusion and misunderstanding and errors can creep into the annotation. The annotation process for the *A. fumigatus* pilot project was a single effort of one institution, but a collaborative effort of three or four annotators. The proposed Whole Genome project that will come from this will be a collaborative effort by a number of institutions across the globe.

4.2 Methods

Annotation is a continuous, difficult and time consuming process and the screen capture images shown here are of the actual stages of the process. But they are not necessarily in the order that the process they were performed due to the very "fluid" nature of the process. Therefore, to explain the methods, I will be using the images in a rough "chronological" order, to show approximately the required steps that are needed for annotation. The images are from the actual *A. fumigatus* contig.

4.2.1 Method overview

1. Genes were identified by annotators within Artemis using the output of various gene finding programs as previously mentioned such as GlimmerM and Genefinder. These programs were "trained" to search for genes using a gene set from *A. nidulans*.

2. Functions were assigned to the genes from the results of BLAST and FASTA searches of public databases. InterPro (Apweiler *et al* 2000), TMHMMv2.0 (Krogh *et al* 2001), t-RNA scans (Lowe and Eddy, 1997) and SignalPv2.0 (Nielsen *et al* 1999) searches were used to assign functions to the predicted genes. Where possible, GO terms were assigned manually to the predicted genes.
3. Initially, the possible GO terms were chosen by searching for sequence similarity in a database of protein sequences and their previously assigned GO terms. The database drew from the databases previously mentioned in 4.1.3.
4. Once sequence alignments had been agreed, GO terms were assigned from either the candidate list or directly from the ontology. If a previously characterised gene was identified, the GO terms were assigned as above. But the inferences were no longer based on sequence similarity; therefore alternative experimental evidence codes were used to relay this.
5. BLAST comparisons and the Artemis Comparison Tool (ACT) (Rutherford, unpublished) were used to view the comparison of the quinate utilisation gene cluster regions of *A. fumigatus*, *N. crassa*, *Podospora anserina* and *A. nidulans*.

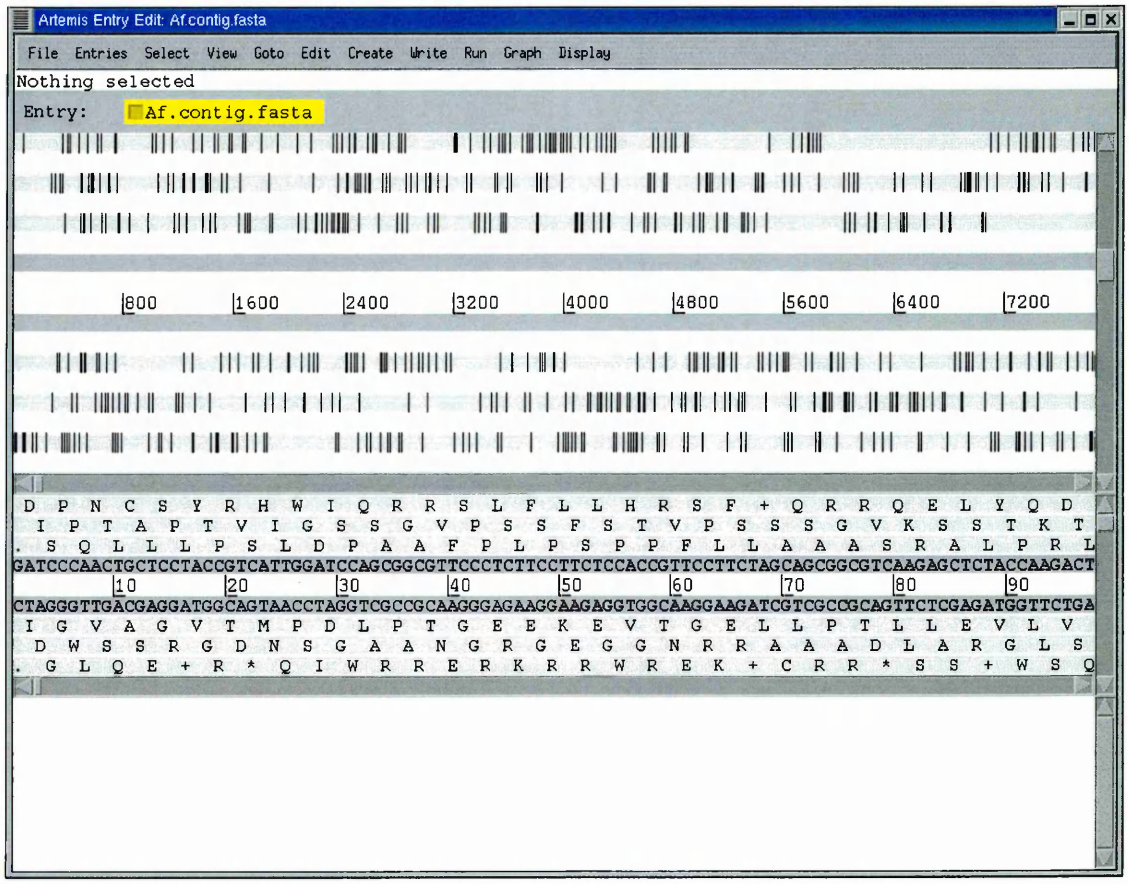
4.2.2 Method used

The sequence was put through a gene finding program via Artemis, which predicted genes from the information that was on screen using the stop/start codons, codon usage within the genome, open reading from size, gene splice sites and many other algorithmical properties. At the moment, the Sanger Institute uses the two previously

mentioned programs, GeneFinder and GlimmerM. Both of these gene finding programs initially had to be "trained" to look for genes in *A. fumigatus* by using a training set of genes from possibly a related genome or a different organism that carries a lot of similarities. These training sets helped the program to predict where genes were likely to be throughout the genome. Once the raw sequence was fed into Artemis and the gene finding programs had predicted where genes were likely to be, the annotator or I had to view and edit the predictions meticulously to ensure the predictions were correct. Splicing had to be correctly altered; ORFs had to be joined if the prediction programs had missed splicing sites and stop end codons needed to be checked to ensure the gene was correctly aligned.

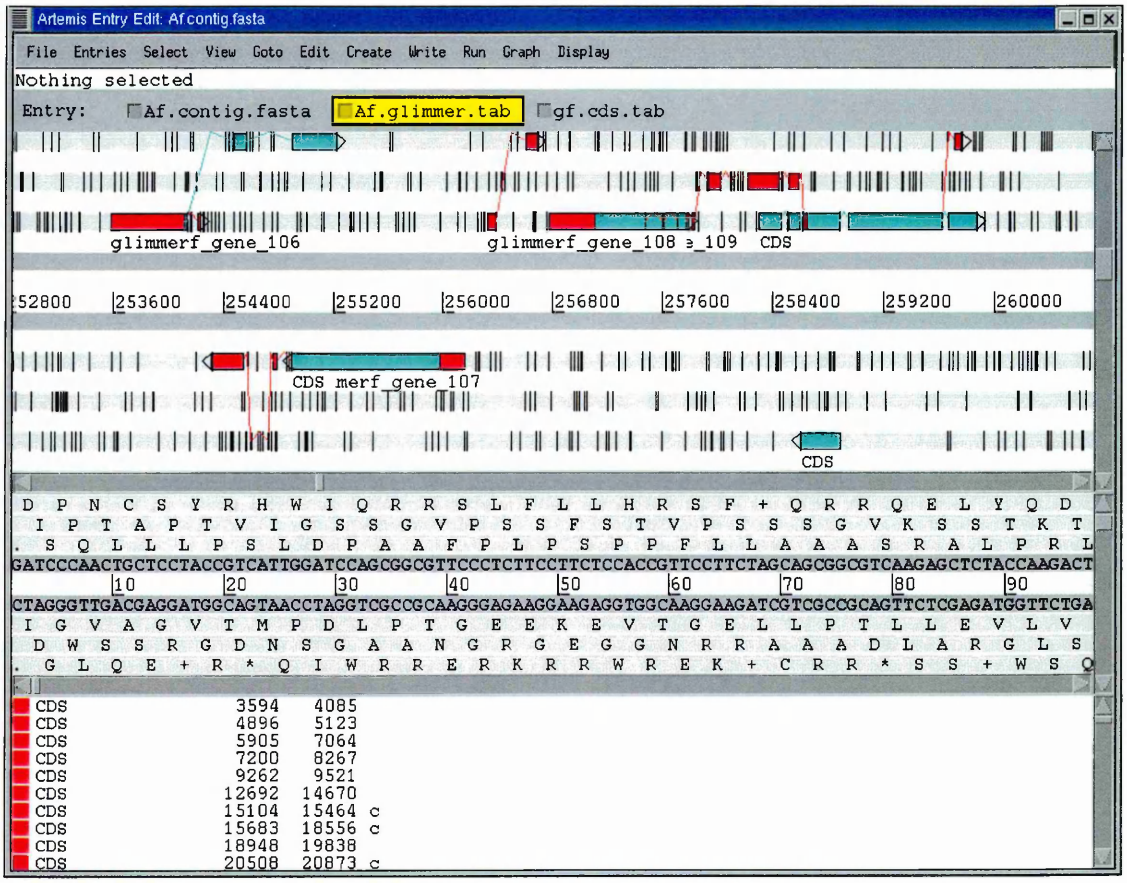
1. The sequence was fed into Artemis in the form of files that were compatible, such as those mentioned in 4.1.1. It displayed this sequence on screen with features in the genome already picked out, such as stop and start codons, open reading frames, amino acid sequences and so on.
2. Artemis was used to read features and sequence from a file and display the features on a six-frame translation of the sequence. Two views of the sequence are shown, both of which can be zoomed in to the base level, or zoomed out to display the entire sequence. Figure 30 shows a blank window with just the sequence file read into it, the start and stop codons and amino acid sequences displayed.

Figure 30. This image shows the three frame read in each direction (top to the right, bottom to the left) and the stop codons are shown as small, vertical black lines. The three frame amino acid sequence is directly below this.



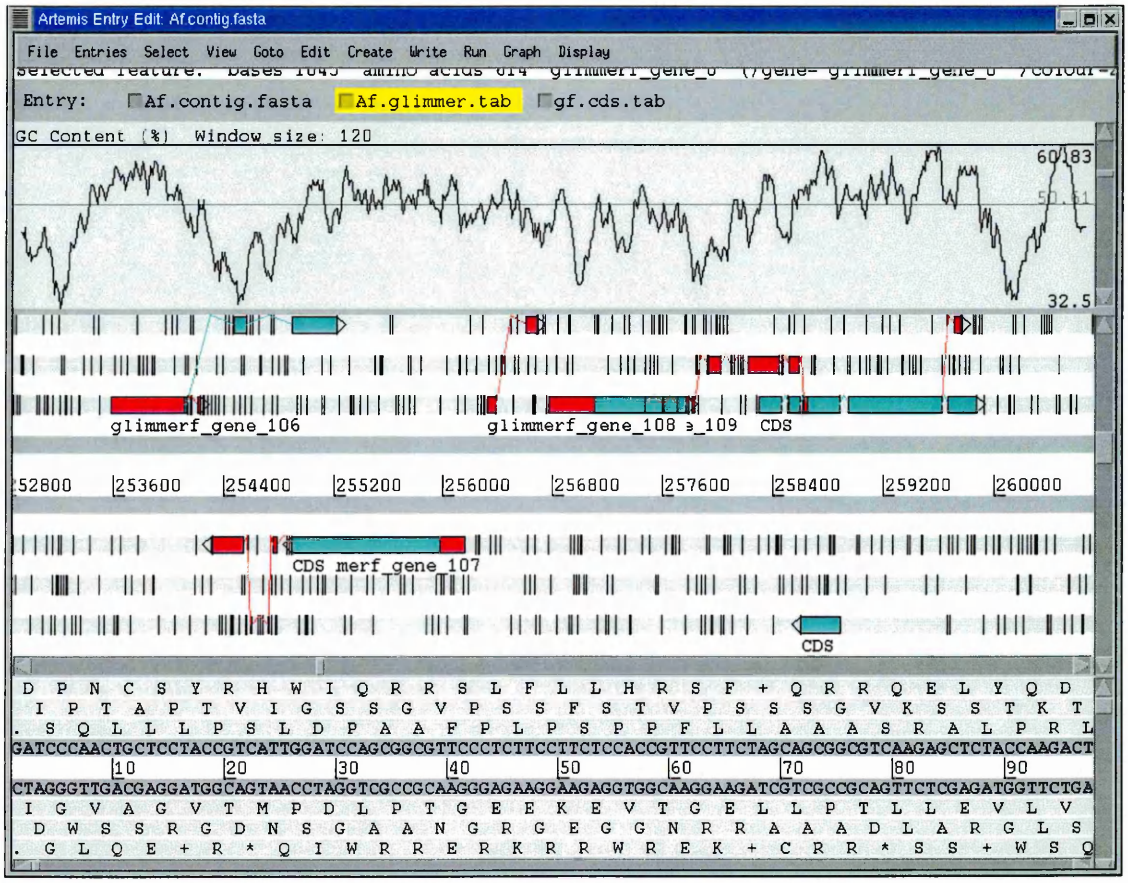
3. Next, the gene finding programs were read into the sequence and the predicted genes and ORFs were displayed within the three frames, with the exons spread over the three frame plot, showing where the predicted splice sites were. Manual curation of this process was time consuming as each prediction had to have the correct splice sites and not all had been correctly predicted by the programs. Moving of the splice sites was a simple click and drag process of the ORF to where the correct splice site was thought to be (Figure 31).

Figure 31. The gene finding programs, GlimmerF and GeneFinder are read into the sequence and the results displayed. The red ORFs are the Glimmer results and the GeneFinder results are in blue. Note the differences in the predictions due to the differences in the programming algorithms. Both are needed though as to make accurate predictions on the correct ORFs.



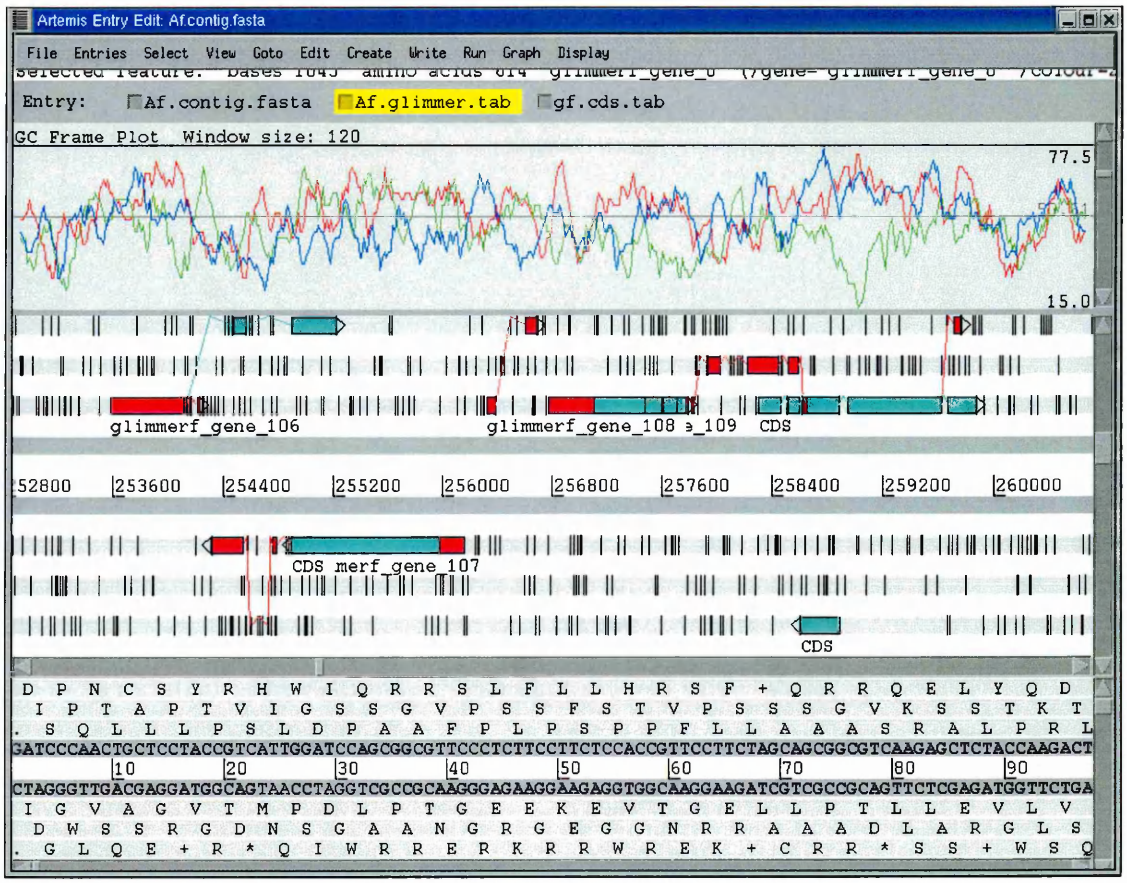
4. Artemis was also used to plot the results of calculations on the sequence, or on any of the CDS (coding) features. This helped confirm ORFs and predicted genes by looking at %G+C plots, when usually there is a higher percentage of G+C in a gene or ORF. Figure 32 shows the same sequence with a simple %G+C plot for the DNA sequence, and a plot of amino-acid properties for one of the CDS features.

Figure 32. %G+C plots showing the increase in the G+C percentage where there are predicted ORFs. The line is the cut off by which the percentage is measured.



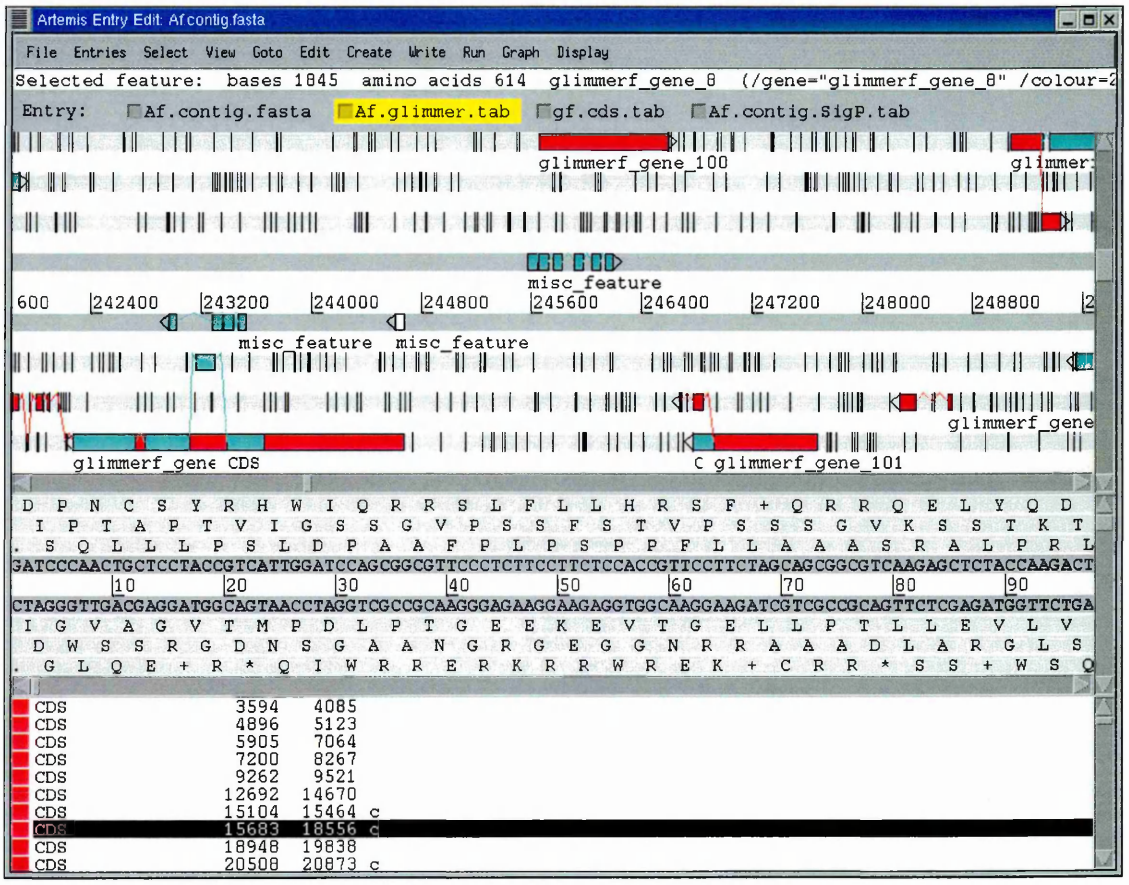
5. Other tools, such as a GC frame plot also helped confirm the predictions and help with corrections of splice sites. Figure 33 shows the three colour graphical display of each %G+C frame plot.

Figure 33. The frame plot shows the GC percentage for each part of the gene, indicating whether the splice sites are correct and confirming that each part of the gene predicted by the gene finding programs is correct.



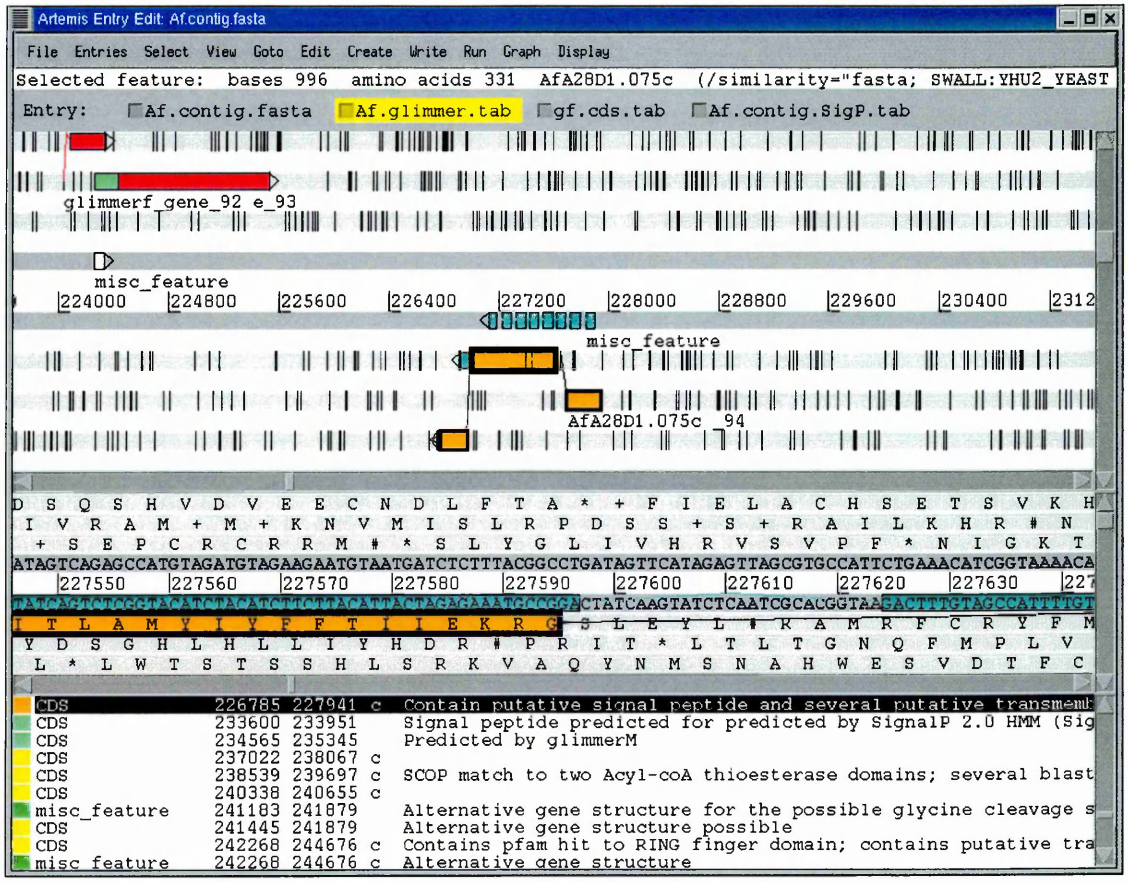
6. Next, the annotation programs and files were read into Artemis. These started the actual gene identification process and also helped identify specific features that would help confirm what the gene and its function was. Figure 34 shows a SignalP and a TMHMM file read into the sequence.

Figure 34. SignalP and TMHMM have been read into the sequence and the results displayed. Note the new features that have appeared in the two middle grey lines at certain points. The small white features are SignalP results, usually indicating a signal peptide has been identified in that gene and the blue features are TMHMM results, usually indicating transmembrane helices at transporter sites within cell membrane. As there are usually more than one, the TMHMM shows the results for all of the helices identified within that region.



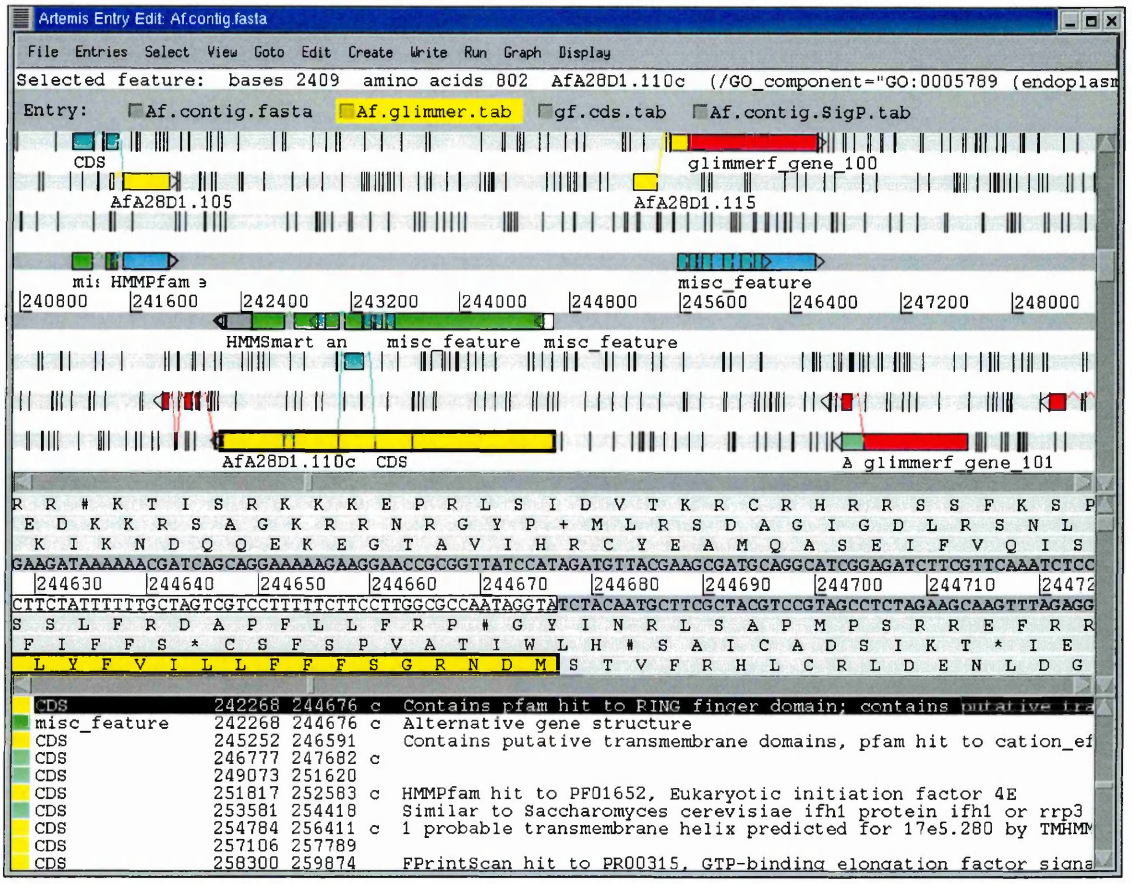
7. Further programs and files were then read into the sequence to identify other features. InterPro, Pfam and SMART scans all helped identify the genes and their function. Figure 35 shows a further step with a number of these files read in and the results displayed.

Figure 35. More features have been edited in this window and the features themselves are indicated in text at the bottom of the screen. Each ORF or feature can be clicked on and the information that has been entered by the annotator about that feature can be read or further edited as more information about it is found.



8. Further files were read in and then web based databases and programs were used to finally come up with a gene model that was close to finishing. As analysis on the genome is a continuous process, the annotation cannot ever be said to be completely finished at this stage, but it is certainly in a form that can be submitted to various databases and to the web. Figure 36 shows annotation close to completion. This would have also gone through the GO function assignment process too, but the results of this are hidden in text, so do not show on screen.

Figure 36. This is close to the final gene model with annotation and program file results fed in and displayed. Each one of the features highlighted is clickable and all annotated information on that feature will be displayed in a window that can also be edited. The main features of each ORF are also displayed in the bottom window for quick and easy viewing of what is in that particular region.



9. Each one of these checks verified what the predicted genes were likely to be. The genes usually come in a number of categories. There are hypothetical genes which are as the name suggests, possibly genes but have no direct, indisputable evidence to say otherwise. There are putative genes where the searches show that it has function analogous to that of other genes, but it is not definite. There are pseudogenes that look like genes but have little or no known function.

10. Each gene was individually looked at by the annotator to ensure that the predicted gene had been identified as accurately as possible and that its sequence had been dealt correctly. The final annotation file was then written.
11. All of this information was fed into an "edit" box as the process proceeded and the information was saved at every point. Once all the genes had been edited and annotated, the whole section of the particular genome under scrutiny was saved as a large file that could be fed back into Artemis at any point, either to be viewed or to be edited at a later date.

4.3 Results

4.3.1 Annotation overview

From a sequence of 921,539bp, 54% was coding sequence with 50.6% as the GC content. There were a total 341 predicted putative protein coding genes and 8t-RNA genes (see Supplementary Figure 2 (**a**) & (**b**) and table 14 at end of this chapter). Gene density was calculated at an average of 0.37/kb of sequence. Each gene had 3 exons per gene on average. The mean gene length, excluding the introns, was 1462bp. This is comparable to 1424 for *S. cerevisiae* (Goffeau *et al* 1996), 1673 for *N. crassa* (Galagan *et al* 2003) and 1626 for *D. discoideum* (Glockner *et al* 2002). It was predicted that 80% of the genes would have introns, which is similar to *Arabidopsis thaliana* at 79% and *N. crassa* at 72%. It is much higher than that found in *S. cerevisiae*, which only has 5% of genes with introns and is still considerably higher than *Plasmodium falciparum* that has 54%,

Schizosaccharomyces pombe has 43% and still higher than *D. discoideum,* which has 68%. There were also 8 predicted t-RNA genes, 5 of which contain a single intron and were arranged in an array. There were no r-RNA genes or arrays found in the small region that was sequenced or any transposable elements, even though there was one possible putative transposase gene. There were no telomeric or centromeric regions in the 900kb sequence. Just 16% of the predicted proteins have been previously been described across the *Aspergilli* genus. From BLAST/FASTA searches, 44% have significant hits to other organisms in the database. Of the predicted proteins, 38% could not be assigned a function, but a significant proportion of these (25% approx) had conserved homologues in one or more organisms. These have been marked in the annotation as "hypothetical protein, conserved". There were 1 or more transmembrane domains in approximately 21% of the total of the predicted genes and there were also 20.5% predicted to have signal peptide or signal anchors. It was seen that the top five FASTA hits to functionally annotated and conserved hypothetical proteins have a number of homologues present only in bacteria and plants. Included in these are proteins involved in secondary metabolism. The proteins predicted in *A. fumigatus* were also compared to the proteome of *N. crassa* and 70% (238 from 341) were identified as putative orthalogues. When the whole proteome dataset of *A. fumigatus* becomes available, a complete comparison with *N. crassa* is going to be of great interest. Table 12 shows a breakdown of the analysis.

Table 12

Feature	*A. fumigatus*	*N. crassa.*	*S. cerevisiae*	*S. pombe*	*D.discoidium*
(G+C) content (%)	50.6	50	38.3	36	22.2
No. of predicted genes	341				
Mean gene length* (bp)	1461	1673	1424	1426	1626
Gene density (bp per gene)	2702	1953	2088	2528	2600
% coding	54	40.2	70.5	57.5	56.3
% coding (including introns)	59.8	46.5	ND	ND	ND
Genes with introns (%)	80	72	5	43	68
Exons					
Number	1089				
No. per gene	3	2.6	ND	ND	ND
(G+C) content (%)	53.8	55.9	28	39.6	29
Mean length (bp)	487	538	ND	ND	711
Introns					
(G+C) content (%)	46.5	ND	ND	ND	13
Mean length (bp)	78	134	82	81	177
Intergenic regions					
(G+C) content (%)	46.5	ND	ND	ND	14
Mean length (bp)	997	ND	515	952	796
No. tRNA genes	8				

4.3.2 Functional classification of predicted gene products

From 153 predicted proteins, 187 Pfam domains were found. From these Pfam domains, the most abundant of them were Zinc-fingers and 9 proteins have been identified with either fungal Zn(2)-Cys(6) binuclear cluster domains or C2H2 type Zn-finger domain (Table 13). There were 74 putative proteins identified that were assigned Enzyme Classification (E.C.) numbers on the contig. This was approximately 22% from the total predicted proteins.

Pfam entry	hits	InterPro entry	Gene ids
PF00172 : Fungal Zn(2)-Cys(6) binuclear cluster domain	5	IPR001138	AfA24A6.035c; AfA5A2.050c; AfA34E8.070c; AfA10A1.045; AfA10A1.100c;
PF00096 : Zinc finger, C2H2 type	4	IPR000822	afa5c11.21c; afa35g10.15; AfA33H4.005; AfA19D12..065;
PF01488 : Shikimate / quinate 5-dehydrogenase	3	IPR002907	AfA5A2.035; AfA5A2.055; afa35g10.13;
PF00501 : AMP-binding enzyme	3	IPR000873	AfA24A6.045; AfA28D10.115c; AfA6E3.015;
PF00106 : short chain dehydrogenase	3	IPR002198	AfA19D12.080; AfA10A1.040c; AfA10A1.050;
PF00083 : Sugar (and other) transporter	3	IPR005828	AfA5A2.030c; AfA28D10.090c; afa35g10.20c;
PF00069 : Protein kinase domain	3	IPR000719	AfA28D1.020; afa5c11.10; afa35g10.01;
PF00023 : Ankyrin repeat	3	IPR002110	AfA14E5.12c; AfA5C5.015c; AfA10A1.125;
PF00005 : ABC transporter	3	IPR003439	AfA5C5.85c; afa5c11.08; AfA10A1.030;
PF02826 : D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain	2	IPR002162	afa35g10.04c; AfA10A1.060;
PF02716 : Isoflavone reductase	2	IPR003866	Afa31G4.020; AfA31G4.040c;
PF01545 : Cation efflux family	2	IPR002524	AfA28D1.115; AfA10A1.085;
PF01487 : Type I 3-dehydroquinase	2	IPR001381	AfA5A2.055; afa35g10.13;
PF01423 : Sm protein	2	IPR001163	AfA28D10.010c; AfA10A1.010;
PF01360 : Monooxygenase	2	IPR000733	AfA14E5.32; afa35g10.06c;
PF01202 : Shikimate kinase	2	IPR000623	AfA6E3.120; afa35g10.13;
PF00503 : G-protein alpha subunit	2	IPR001019	afa5c11.10c; AfA6E3.035c;
PF00400 : WD domain, G-beta repeat	2	IPR001680	AfA14E5.29c; afa35g10.09c;
PF00389 : D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain	2	IPR006139	afa35g10.04c; AfA10A1.060;
PF00155 : Aminotransferase class I and II	2	IPR004839	AfA28D1.001; AfA33H4.060c;

Table 13. Pfam and InterPro entries from the initial annotation process. Note the high abundance of the Zinc finger Pfam hits.

From the total gene products, 177 were assigned GO terms (52%) manually. The annotation highlighted a tendency to have a slightly higher proportion of genes annotated to function as either enzymes (39%) or transporters (13%) compared to other organisms

such as *S. cerevisiae* and *Arabidopsis thaliana* (36% and 30% enzymes respectively and 8% transporters in both organisms). Compared to *S. cerevisiae,* the proportion of secondary metabolism proteins was higher in *A. fumigatus* (18% compared to 13%), but was significantly lower than *A.thaliana* and the worm *C. elegans* (both approx. 23%). However, it did not reflect the genome as a whole for *A. fumigatus*, as it suggested that the contig that had been analysed may just have contained more than one functional gene cluster, which would have biased the percentage of proteins that had been annotated as enzymes or transporters.

4.3.3 Analysis and comparison of *A. nidulans* linkage group VIII and the *qut/qa* gene cluster synteny comparison

Analysis of the degree of synteny of the orthologous *A. fumigatus* genes and those mapped to the linkage group VIII in *A. nidulans* (http://www.gla.ac.uk/Acad/IBLS/molgen/aspergillus/index.html) was made and the results are illustrated below.

Figure 37 shows there was an inversion of two regions (*facC* to *veA*; *trpC* to *hisC*) with respect to the gene order could be observed in *A. fumigatus*. It was also seen that the *galG* and *fwA* loci between the markers *facC* and the *qut* gene cluster was missing when compared to *A. nidulans*. *qut* has a function in the utilisation of quinate. As the loci *brlA*, *pyrD*, *aldA* and *argC* all map to the linkage group VIII of *A. nidulans* genetic map, it was not unreasonable to expect that the region or cluster was to be found elsewhere on the genome.

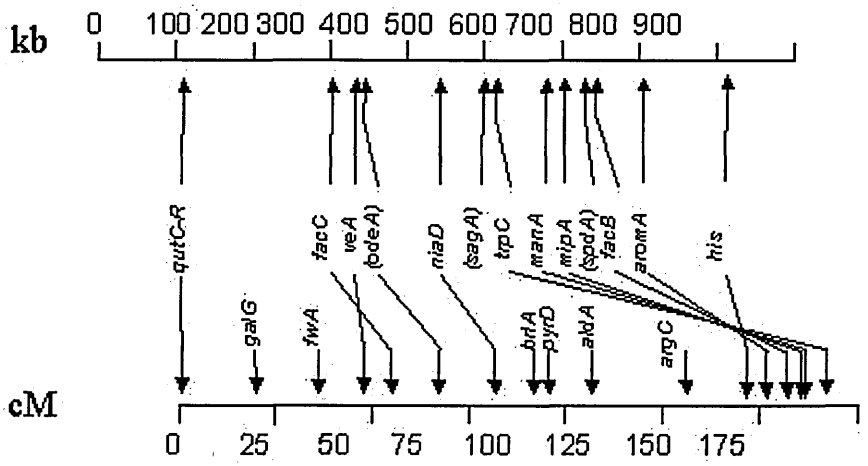


Figure 37. Synteny shown in linkage group VIII between *A. fumigatus* and *A. nidulans*. Note the regions of inversion at *facC, veA, trypC* to *hisC.*

The quinate utilisation gene cluster was compared between *A. fumigatus* and a number of other filamentous fungi, such as *A. nidulans, N. crassa* and *Podospora anserina*. This is known as the '*qut*' genes in Aspergillus spp. and *P.anserine* and known as '*qa*' in *N. crassa*. This cluster was looked at as the sequence to the *qut* and *qa* genes are already in public databases. The clusters are involved involved in the quinate utilisation pathway and the genes code for at least 5 structural genes and 2 regulatory genes. Artemis Comparison Tool (ACT) was used to perform the comparisons. In Figure 38, *A. fumigutus* and *A. nidulans* showed that the gene order in the *qut* genes was completely conserved, but the orthologous *qa* genes in *N. crassa* and *P. anserina* show little or no order of conservation.

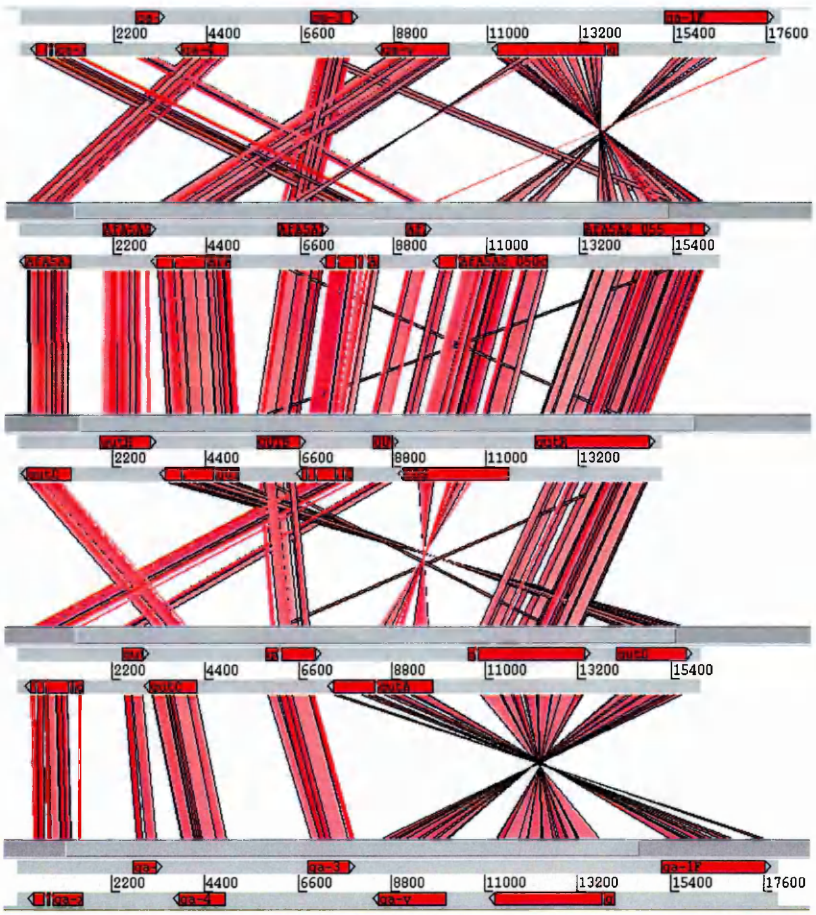


Figure 38. Artemis Comparison Tool (ACT) showing degrees of synteny of the *qut* cluster and its adjacent regions. The red stripes are regions of synteny. The blank spaces are regions where no synteny occurs. From the region, it is seen that *A. nidulans* and *A. fumigatus* have high levels of synteny along nearly all of the cluster, where other organisms show little or none across the region.

N. crassa and *P. anserina* showed synteny with complete conservation in orientation and order of the *qutG, B, C* and *B* genes, but in the two species the *qutA, R* and *D* genes were inverted. The *qutH* gene, which has an unknown function in the quinate utilisation pathway, codes for a gene with putative oxidoreductase activity. This gene is not found in *N. crassa* and *P. anserina* in the orthologous cluster. There was also a *qutD*-like

permease gene found 57kb away from the *qutD* gene found in the cluster in *A. fumigatus*. The *qutD*-like permease is predicted to have nine transmembrane domains compared to the eleven in the *qutD* permease gene or twelve in the *A. nidulans*. The *qutD*-like permease and the actual *qutD* permease gene have only 29% similarity compared to the 76% similarity between the *A. fumigatus* and *A. nidulans qutD* proteins. The three proteins are part of the sugar transporter family (MFS), which in turn is part of a major facilitator superfamily which is involved in the transport of organic alcohols, organic acids and carbohydrates. It is still unclear what, or if, the *qutD*-like permease has a role in the quinate utilisation gene cluster in *A. fumigatus*.

4.3.4 Aflatoxin-related genes

A. flavus, A. parasiticus and *A. nidulans* have all had extensive studies with respect to the organisation and expression of the aflatoxin B1 gene cluster (Woloshuk and Prieto 1998). These were used for the basis of our comparison. In *A. flavus* and *A. parasiticus*, the aflatoxin producing genes are located in a 75-90kb cluster and the genes for the aflatoxin synthesis intermediate, sterigmatocystin, are also shown to be in a cluster (Prieto *et al* 1996, Yu and Leonard 1995, Keller and Hohn 1997). The aflatoxin B1 biosynthesis gene cluster was not identified on the contig that was sequenced for *A. fumigatus*, but two genes with significant similarities to genes in the aflatoxin gene cluster were identified. The annotation assigned a 'vesicolorin b synthase-like' label to one which showed high similarity to the *vbS* gene of *A. nidulans* and the other as '*aflT*-like major facilitator superfamily' which shows significant similarity to *aflT* (aflotoxin efflux pump) gene of

A. nidulans. It is thought after extensive analysis that the aflatoxin biosynthesis cluster is located elsewhere in the genome of *A. fumigatus*, therefore the genes are not part of the aflatoxin B1 cluster and it is not known whether they have a role in aflatoxin B1 biosynthesis.

Another interesting gene found in the contig is an aldo-keto reductase gene. It has a 42% identity to the human aflatoxin B1-aldehyde reductase (AFAR) and is thought to code for a protein very similar to the AFAR. Both proteins are part of the aldo-keto reductase 7 family (AKR7), which in itself is part of the AKR superfamily (Jez *et al* 1997). There are a number of enzymatic systems that have evolved to deal with ketones and aldehydes, as these can present significant problems to cellular function. Members of the AKR1 such as aldose reductase and members of the AKR7 family are often found in mammalian systems, particularly humans and rats. A number of important amino acid residues, such as the highly conserved mammalian aflatoxin B1-aldehyde reductases and aldose reductases, can also be found in the *A. fumigatus* protein. The residues make up a catalytic triad and also bind to the co-factor NADPH (Wilson 1993, Ireland 1998). There is an orthalogue that can be found in *N. crassa*. There is 62% identity and the initial thought is that they may have a role in specific charged aliphatic and aromatic aldehyde detoxification, but there would need to be experimental verification that the AKR7 family has a similar substrate range.

4.3.5 Possible drug targets

The contig contains a gene called *aroM,* which codes for a 171kDa pentafunctional polypeptide. It has five functional domains, which is the same as other *aroM* orthologues.

Bacteria, plants and microbial eukaryotes all have a system for synthesis of aromatic compounds and an intermediary compound in this system is shikimate. The five functional domains in AroM orthologues are known to catalyse consecutive steps in the shikimate biosynthetic pathway (Bentley, 1990). The protein is required in Salmonellae for virulence (Gunel-Ozcan *et al* 1997) and is potentially a target for anti-fungal and anti-bacterial drugs (Kishore and Shah, 1988). The crystal structure of one of the five domains, dehydroquinate synthase, has been determined in *A. nidulans* (Carpenter *et al*, 1998). The AroM sequence in *A. fumigatus* is 80% identical to the *A. nidulans* orthologue and therefore is a candidate for more modelling to look for potential inhibitors.

There are also a number of other enzymes that have similarity to homologous plant and bacterial proteins. Examples of these are a short chain dehydrogenase/reductase, a dioxygenase (which is involved in the cleavage of aromatic rings), a putative 4-coumerate-CoA ligase and a putative 3-oxoacyl-[acyl carrier protein] reductase, which is also a member of the short chain dehydrogenase/reductase family. It may be that these proteins are components of biochemical pathways not found in humans. If this is the case, then as long as they are part of the fungal mechanism for growth in human tissue, then they are good, viable drug targets.

4.3.6 Drug/toxin transporters

A. fumigatus produces a number of secondary metabolites that are toxic to its own cells. Therefore the fungus has a number of transporters to remove the toxins from its cells. In the contig, there were a total of twenty putative transporters identified. Fortuitously, some of these transporters can pump anti-fungal drugs from the cells. Therefore, efflux is

probably the major mechanism of azole drug resistance in *Aspergilli* (Moore *et al* 2000). Three genes in the contig, AfA35C10.20c, AfA5C5.050 and AfA24A6.065 code for proteins related to a group found in the MFS class, which is involved in the efflux of drugs. Also, interestingly, another gene, AfA5C5.085c, is already previously characterised as *mdr4*. This is known to be highly over expressed in *A. fumigatus* mutants that confer high level itraconazole resistance. Its product is a member of the ABC transporter superfamily (Nascimento *et al* 2003). Afa10A1.030, whose product is also part of the ABC transporter superfamily, but it is known to have a role in itraconazole susceptibility (Mosquera *et al* 2002). The *A. nidulans* orthologue, *atrG*, has also been characterised and is known to have a similar role in the resistance of azoles (Andrade, 2002).

4.3.7 Transcription factors

The fungal specific zinc-finger domain, Zn(2)-Cys(6) binuclear cluster domain has already been shown to be the most abundant Pfam domain seen in the contig. This is an N-terminal region DNA-binding protein found in transcription regulators. There have been 17 proteins predicted to function as transcription factors. This equates to 8% of the functionally annotated proteins. Of these, the activator QutA and repressor QutR, of the quinate utilisation cluster; AfA14E5.23c, the Jumonji family transcription factor (Balciunas and Ronne 2000) and the acetate regulatory FacB orthologue are of most interest.

4.3.8 Putative host interaction molecules

Host interaction genes are likely to be of high interest to investigators and a number of interesting candidate genes have been identified in the contig. A gene identified as AfA10A1.015 encodes for a 580 amino-acid-long protein predicted by Pfam to contain 2 fascilin domains. It is thought that this might represent ancient a cell adhesion domain and usually proteins have 2 or 4 copies of this domain and are GPI-anchored to the membrane. The *A. fumigatus* protein has a predicted signal peptide, but there is no prediction of a GPI-anchor.

There is a very obvious *N. crassa* orthologue, AfA33H4.165, which encodes for a 1236 amino acid protein. The C-terminal half of this protein has 26% identity over 644 amino acids to the TGF-β receptor associated protein. But it may not be expressed on the cell surface as it is not predicted to have a signal peptide.

Also, AfA5A2.010 has been found to have 35% over 318 amino acids similarity to human integrin beta 1 protein 2 and has been annotated as a chord containing protein homologue. It is still not known whether *A. fumigatus* expresses an integrin beta 1 binding domain-containing protein to mediate interactions with the host, but this gene also has a very strong *N. crassa* orthologue and has no prediction to be secreted at the cell surface.

4.4 Discussion

This pilot project was designed to give a possible whole genome sequencing project a head start and to determine if sequencing the entire genome would be possible. Before

the project had started, there had only been one protein structure that has been deciphered in the *A. fumigatus* genome. This is the Mn-dependant superoxide dismutase (Fluckiger *et al* 2002). There are only 44 *A. nidulans* proteins known. To have only 16% of the 341 annotated genes identified in other Aspergilli so far shows the possible scale of the task ahead. Another indicator of the size of this undertaking is that there is one third of the proteins still not found to have an orthologue in *N. crassa.*

On detailed analysis, the sequence so far shows a gene density in *A. fumigatus* that is comparable to the fission yeast, *S. pombe.* It also shows that the majority of the genes are spliced. Significantly and interestingly, the percentage of spliced genes found so far is much higher than that of *S. pombe* and *N. crassa.* But it does compare favourably with *A.thaliana.* There have been other genes and proteins that have been reported for *A. fumigatus,* such as the retrotransposons Afut1 and Afut2 (Class I) (Glayzer *et al* 1995, Neuveglise 1996, Paris 2001) but they have not been found in the region that has so far been sequenced. In fact, this region has not shown any signs of any degenerative transposable elements. But other filamentous fungi have had both Class I and Class II transposable elements found and identified, so it is expected that these will be identified in other regions of the genome that have not yet been sequenced or analysed. They can be seen as important as they are natural causes of genetic instability in fungi and therefore they could be used for identification and manipulation of the fungus.

Drug targets are an obviously important area for further study and research and there have been a number of genes that have been identified that will facilitate the search for possible drug targets and allergens. The pilot project has identified a number of genes that are involved in diverse functions and as a number of genes are involved in secondary

metabolite metabolism, it is thought that *A. fumigatus* will have a large catalogue of transcriptional control points and export mechanisms. Within the contig, several transcriptional control proteins and transporters have already been identified and translate into about one fifth of the functionally annotated proteins. These could possibly give an indication of what may be found in the rest of the genome, but so far it is not known whether this is a localised property of the contig sequenced so far or whether it is an indicator into what will be found in the whole *A. fumigatus* proteome. But it is possibly a very good indicator to the *A. fumigatus* proteome content and therefore with the trend for a large number of transcriptional control mechanisms so far identified, will show why *A. fumigatus* is such a highly opportunistic pathogen of man.

There is a very high likelihood that *A. fumigatus* has evolved to produce a number of transporter pumps to remove the secondary metabolites that it produces that are toxic to its own cells. Although the contig represents only a small fraction of the genome, it does indicate that there could be a high number of transporters within the rest of the genome, which will become of great interest with the increase into drug resistance in the fungus (Del Sorbo *et al* 1997, Warris *et al* 2002). Regions of synteny will indicate and identify a number of significant factors, such as evolutionary significant genes, potential regulatory elements and ORF identification and annotation. As more sequences from different genomes become available, there has been a sharp increase in the amount of interest in these syntenic regions, including filamentous fungi (Seoighe *et al* 2000, Hamer *et al* 2001, Pederson 2002). This study has shown that *A. fumigatus* and *A. nidulans* show high levels of synteny in the quinate utilisation gene cluster. The study also showed that in other Ascomycetous fungi, evolution has kept the genes located in the quinate utilisation

cluster as a cluster, but there have been a number of genetic rearrangements within the cluster. As other genomes and annotation become available for Ascomycetous fungi, it would be interesting to study the levels of synteny of other metabolically related gene clusters.

The project also showed that there is a high similarity in the putative aldo-keto reductase gene in *A. fumigatus* and the mammalian aflatoxin B1-aldehyde reductase and this could prove highly significant. This protein may provide a role in the detoxification of aromatic and aliphatic aldehydes including aflatoxin B1 catabolism. In humans, the aflatoxin B1 protein (afb1) is metabolised by aflatoxin B1 aldehyde reductase and has been shown that it may have some function in the protection of the liver against acute and long term carcinogenic effects of aflatoxin B1 (Neal *et al* 1998; Kelly *et al* 1997). It is known that G-T transversions in the p53 tumour suppressor gene at codon 249 (Eaton and Gallagher 1994) is caused by aflatoxin B1 and therefore *A. fumigatus* having an aflatoxin B1 detoxifying gene will make this a very good candidate for further study into aflatoxin metabolism.

Identification of the fungus encoded molecules that influence the outcome of infection during invasive aspergillosis has become an interesting focus of attention (Muhlschlegal *et al* 1998; Latge and Calderone 2002). A significant number of specific interactions between *A. fumigatus* conidia and host molecules have already been described (Sturtevant and Latge 1992; Mendes-Giannini *et al* 2000; Clemons *et al*, 2000; Allen, Voelker and Mason, 2001; Robinson *et al*, 2001; Monod *et al*, 2002; Tronchin *et al* 2002). This study has identified a number of genes that may be involved with host molecule interaction, such as the integrins. There are more than 75 IgE –binding allergen molecules produced

by *A. fumigatus* and to understand the patho-physiology of allergic bronchopulmonary aspergillosis, further identification of these molecules needs to done (Kurup *et al* 2000; Banerjee *et al* 2002; Kodzius *et al*, 2003).

Further comparison of invasive *A. fumigatus* isolates and non-invasive isolates or non-pathogenic *A. nidulans* strains will be able to be carried out when the whole genome sequence of both organisms becomes available and the various comparisons will hopefully be a pointer to why *A. fumigatus* has become so much more of varied pathogen and is capable of much more varied and extremes of infection, from allergies to invasive diseases. Hopefully, the further comparisons will also allow *A. fumigatus* specific pathogenic pathways and increase the identification of new and novel drug targets.

5. Summary

With the ever-increasing prevalence of *Aspergillus fumigatus* infection and invasion threatening to become a major problem in the coming years, the sequencing of the genome is taking on a much greater importance in the medical community. This pilot project was prompted and designed in a way as to determine whether the sequencing of the genome was a feasible objective by carrying out a small-scale sequencing project to sequence just 1Mb of the *A. fumigatus* genome. What I hope that this thesis has proven is that not only was it feasible, but also that it went beyond expectations in proving what could be done in the constraints of time and budget but also, more importantly, what can be found within the genome.

Many people are at risk from this opportunistic pathogen. Healthy and immunocompromised people are exposed to the airborne spores of this fungus 24 hours a day and can be infected, resulting in symptoms ranging from skin lesions and eye infections to lethal lung and central nervous system debilitation. It's the most common mould causing infection across the globe and is almost impossible to treat sufficiently. There have been some genetic studies on this or organism, but most are small-scale, specialised studies into specific areas of interest. This pilot project was devised to determine whether this understanding could be increased significantly and allow this genetic information to be available to all who could use this information in the fight against infection. This information could and will be used in virulence, pathogenicity and metabolic studies in order to help understand its workings and design anti fungal drugs to combat infection. The information could be used in comparative and post genomic studies in order to better understand its biology and metabolism and determine if anything useful could be taken from what the sequence holds.

The initial attempts at the construction of a genomic library good enough for sequencing, following a rough guideline from the Osoegawa paper, were dotted with problems and it took over a year to get a good, viable library. There were a number of reasons for this but after some serious thought and some consultation with collaborators, it was eventually resolved. There needed to be a high yield of clean, good quality DNA and some experimentation to remove the smaller fragments of DNA from the process that could detrimentally affect downstream procedures by DNA "trapping" were carried out, with differing levels of success. When the consultations with collaborators provided greater

quality DNA, it was found that the experiments to remove the smaller fragments and cellular debris were not required, which became a turning point as it was not long after a viable library was made. Techniques such as "pre-electrophoresis" and electroelution helped us further in refining the techniques, removing smaller fragments of DNA and helping keep the DNA in tact to optimise the process to give the best viable library we could make with the resources available to us.

For us to use a library, the insert sizes had to be consistently large and above an arbitrarily picked size of about 60kb. Many attempts gave deleted inserts, inserts of inferior quality and size. The overall insert size that was eventually generated was not ideal, being just over 70kb, but it was more than good enough to cover the genome and be a usable, working library for sequencing.

The next step was to find a starting place for the sequencing and mapping process using radiation hybridisation. The *niaD* gene was used as it was one of the few well characterised gene in *A. fumigatus*, as well as *Aspergillus nidulans.* Once the "starting" point had been identified as a clone on the array, it was a simple case of "walking" out from this clone by a combination of end sequence matching and fingerprint information. Good overlaps between clones were found, few were mismatched, most clones were intact; it could be said that the *A. fumigatus* genome was almost the "model" genome to assemble.

The only slight problem encountered was the less than ideal read pair pass rate figure of 64% from the 3456 clones the end sequencing produced when normally a pass rate of around 80-85% is the target aimed for. This is a 7.5-8.0-fold coverage of the genome. The fact that the quality of the clones that *had* passed meant that the programs used to "fit" the pieces back together (FPC, BLAST etc) had no problem achieving this and the lower than expected pass rate could be overlooked to a certain extent as there was still more than enough coverage to complete the goal. The reasons for the reduced pass rate were numerous, such as sequencing reactions not working, database mismatching and wrong nomenclatures.

The project ended with an overall 16 "super-contigs" covering just under 1Mb of the genome, centred around the *niaD* gene and its associated cluster genes. This information was fed into a central, purpose built database that could be accessed by a number of other programs for use further downstream. The map was also useful for the whole genome shotgun that has continued on since this project has come to an end, being used primarily as a good anchor and starting point for that project.

Before the pilot project started, only the Mn-dependant superoxide dismutase protein had been characterised and deciphered from the *A. fumigatus* genome. In comparison, only 44 proteins were identified in *A. nidulans* at that time. What emphasised the importance of the pilot project was finding that only 16% of the 341 genes found in the 1Mb were identified in other Aspergilli. The fact that this figure is so low, even in such a small proportion of the estimated 30Mb genome size showed the scale of the whole genome

project. The pilot project sowed the seeds for this to continue, as it showed that the practicalities could be overcome.

The analysis of the 1Mb region shows a gene density comparable to that of the *S. pombe*. There is also a high degree of gene splicing, much higher than *S. pombe*. Some genes, such as the retrotransposons Afut1 and Afut2 that have been reported are yet to be identified in the 1Mb region, but these may be found elsewhere in the genome with the whole genome project underway.

Possible drug targets are obviously important to identify (if possible) and a number of genes were found in the 1Mb that could help with this search. The *aroM* gene was identified, which is involved in the synthesis of aromatic compounds and is part of the shikimate pathway. Identified in Salmonellae for virulence, it is potentially a target for anti-fungal and bacterial drugs. Further searches on the whole genome would hopefully highlight more of these genes. A number of genes coding for enzymes homologous to plant proteins were also found, such as the putative 4-coumerate-CoA ligase that may be a component in the pathway involved in fungal growth, which make them highly conducive to drug target studies. A number of secondary metabolism genes have been identified, which indicates that there could be a high number of transcriptional control point and export mechanisms. About one fifth of the functionally annotated proteins from the 1Mb are already identified as such. If the whole proteome is found to follow this trend, it could indicate why *A. fumigatus* is such a highly opportunistic pathogen.

Transporter genes seemed to be relatively prevalent in the 1Mb, indicating more to be identified in the whole genome studies. These could give good indicators to the mechanisms involved in the organism's drug resistance, further increasing interest in studies to exploit these findings further downstream as the data from the whole genome sequence is unravelled.

Synteny studies will also become important in searching for evolutionary significant genes as studies on other filamentous fungi highlight areas for comparison. For instance, it has already been shown that *A. fumigatus* and *A. nidulans* show significant synteny in the quinate utilisation cluster. Interestingly, the studies have also shown that evolution in ascomycetous fungi has kept the location of these genes in a cluster across a number of other fungi, although there have been some genetic rearrangements. Comparison studies from these findings help further in the understanding at a molecular level for not just *A. fumigatus*, but a number of other fungi too.

Further genes, such as the putative aldo-keto reductase genes and the 75+ IgE-binging allergen molecules will all give rise to further study due to the importance of them in aflatoxin metabolism and the patho-physiology of allergic bronchopulmonary aspergillosis, two highly significant areas of interest. The former is thought to be one of the most important factors in fungal virulence and the latter focuses on a very important, specific illness that can arise from the infection. Comparing these two here shows the diversity of the experimentation and studies that are likely to come from the whole genome sequencing, from the interaction of a single molecule with the host to identifying

and combating the processes involved with one of the more prevalent and dangerous forms of infection.

Further work on this organism will continue unabated now that the pilot project has been a significant success in pointing the way for the whole genome sequence. It has also shown that there is still plenty of information and data that can and will be taken from the rest of the genome. Comparison studies will hopefully highlight why *A. fumigatus* is such a highly virulent organism compared to its close relatives and why it infects in such varied ways.

The whole genome sequencing project for *A. fumigatus* started towards the end of the pilot project as preliminary results from this were first appearing. It was clear from the outset that the whole genome project would be feasible and that work should start immediately. This has now been going for almost two years now, with results published on the Sanger Institute web pages every day (www.sanger.ac.uk). The annotation process has not been completed as yet, so figures and indications of what has been found in the genome are few and far between. But there is preliminary annotated data in GeneDB (www.genedb.org).

Other groups within the fungal study community have done further complimentary work. One such study is the "optical" map of the genome by the company OpGen (www.opgen.com). Using the information from 19 contigs of the whole genome sequence, they have completed a whole genome map for the international consortium in

order to complete the genome sequence. The data from the OpGen optical map has allowed the contigs to be ordered and orientated on what is now believed to be a confirmed eight chromosomes. This has allowed the size, location and number of gaps remaining to be finished on the project. The map corresponded to over 300 times coverage of the genome and has showed that the optical map is extremely powerful and versatile. Not only that, it has helped the international consortium immensely in being able to identify areas of the genome that are not complete and therefore speeding up the process to completion and ready for the annotation stage. An explanation of this ingenious process can be found on their website.

Recently, The Institute of Genome Research (TIGR) have announced that they are selecting *A. fumigatus* for the development of functional genomics resources and reagents via their Pathogen Functional Genomics Resource Centre. This will study a number of organisms using microarrays for in depth genomic comparison and study (http://pfgrc.tigr.org/).

As the sequence data is relatively new and there is little of it, direct genomic studies on *A. fumigatus* are few and far between. This will increase as the sequence data increases on various databases. This will open up the door to many types of experimentation. One important area to look at would be comparison studies using microarrays for expression studies. Using clinical isolates from patients and a cultured laboratory strain, a comparison of the two across its genome would be able to highlight the genes that are being expressed at various stages of infection in the host. Further studies, such as gene knockouts, could then be carried out to determine which genes are causing the infection

and what effect they have on the host. This would hopefully lead to the development of novel drugs, either targeted specifically to the host or to the organism to help combat the infection.

A full study of the organism's proteome would help understand its expression patterns, to better understand its biology and to help identify novel drug and vaccine targets. Proteomic profiles of the organism at different stages of its life cycle could be decoded, enabling scientists to learn more about its development, its host interaction and its general biology. It would also help scientists understand the continuing process of increased resistance to the already largely ineffective and toxic drugs that are on the market to combat the organism. It would help understand the mechanisms of how and why it becomes resistant, helping scientists to be one step ahead in their fight against infection.

Putative genes have usually been seen to be good starting points for developing drug targets. As annotated sequence becomes more available, putative genes would be highlighted, with their possible function being identified and experiments designed around this. With the information available to everyone, there is opportunity for many laboratories to tackle many different putative regions and functions.

Other studies, such as Single Nucleotide Polymorhism experiments can show the evolutionary divergence of the organism over time and how this divergence has affected its interaction with humans. This may lead to the identification of epidemiological patterns that can help scientists reduce areas of risk.

As the *Plasmodium falciparum* sequencing project has shown, the sequencing of the genome may not lead instantly to the answers of eradication or combat to diseases due to technology, funding and infrastructure, but at least the sequence is as good a starting point as any and will hopefully lead the way for a better understanding of the fungus.

This project and the ongoing whole genome sequencing project has been a huge undertaking, but one that ultimately will pay dividends when the genome releases its secrets to the world. With the rapidly increasing prevalence of this lethal fungus, work will continue on in the search to help either eradicate infection, help the symptoms of those who are already infected and also to fully understand the workings of the innocuous, yet highly dangerous mould.

Table 14

The 341 annotated genes from the 922kb contig of *Aspergillus fumigatus*

A. fumigatus CDS AfA34E6.080 hypothetical protein
A. fumigatus CDS AfA28D10.085c transporter, putative
A. fumigatus CDS AfA24A6.120c bgt1, beta-1,3-glucanosyltransferase
A. fumigatus CDS AfA31G4.020 isoflavone reductase, putative
A. fumigatus CDS AfA28D1.090c n-acetyltransferase, putative
A. fumigatus CDS AfA14E5.18c NADH-ubiquinone oxidoreductase, putative
A. fumigatus CDS AfA28D10.005c hypothetical protein, conserved
A. fumigatus CDS AfA28D1.050c hypothetical protein
A. fumigatus CDS AfA5A2.030c qutD, quinate permease, putative
A. fumigatus CDS AfA8D5.035 hypothetical protein
A. fumigatus CDS AfA34E6.065c tktA, transketolase, putative
A. fumigatus CDS AfA28D1.010c hypothetical protein
A. fumigatus CDS AfA6E3.195 DCG1-like protein, putative
A. fumigatus CDS AfA33H4.065c hypothetical protein
A. fumigatus CDS AfA24A6.115c myosin heavy chain-like protein, putative

A. fumigatus CDS AfA34E6.105 spindle-related protein, putative
A. fumigatus CDS AfA14E5.17c nucleoporin, putative
A. fumigatus CDS AfA8D5.025 adenylyl cyclase-associated protein, putative
A. fumigatus CDS AfA24A6.090c esterase/lipase/thioesterase family protein, putative
A. fumigatus CDS AfA34E6.060 spermidine synthase, putative
A. fumigatus CDS AfA6E3.170c DUF6-like integral membrane protein, putative
A. fumigatus CDS AfA28D1.125 hypothetical protein
A. fumigatus CDS AfA6E3.130c esterase/lipase/thioesterase family protein, putative
A. fumigatus CDS AfA14E5.20c gabA, GabA permease, putative
A. fumigatus CDS AfA28D1.005c hypothetical protein, conserved
A. fumigatus CDS AfA10A1.110c hypothetical protein, conserved
A. fumigatus CDS AfA28D1.085 hypothetical protein
A. fumigatus CDS AfA5C5.015c suppressor protein spt23-related, with ankyrin repeats
A. fumigatus CDS afa35g10.09C hypothetical protein, conserved
A. fumigatus CDS AfA28D1.080 hypothetical protein
A. fumigatus CDS AfA24A6.140 14-3-3-like protein, putative
A. fumigatus CDS AfA34E6.055 ubiquinol-cytochrome c reductase complex ubiquinone-binding protein qp-c precursor, putative
A. fumigatus CDS AfA33H4.080 nuclear segregation protein, putative
A. fumigatus CDS AfA8D5.010 hypothetical protein, conserved
A. fumigatus CDS AfA28D1.115 cation transport protein
A. fumigatus CDS afa35g10.12C transport protein, putative
A. fumigatus CDS AfA28D10.110c mitochondrial processing peptidase alpha subunit, putative
A. fumigatus CDS AfA24A6.045c 4-coumarate-coa ligase, putative
A. fumigatus CDS AfA14E5.15c hypothetical protein, conserved
A. fumigatus CDS AfA28D1.070 hypothetical protein
A. fumigatus CDS AfA24A6.135 phosphoesterase, putative
A. fumigatus CDS AfA10A1.001c hypothetical protein
A. fumigatus CDS AfA33H4.170c dioxygenase, putative
A. fumigatus CDS AfA34E6.045 hypothetical protein
A. fumigatus CDS AfA24A6.095 ubiquitin-conjugating enzyme e2, putative
A. fumigatus CDS AfA34E6.040 hypothetical protein
A. fumigatus CDS AfA6E3.200 mipA, gamma tubulin, putative
A. fumigatus CDS AfA33H4.130c possible epimerase
A. fumigatus CDS AfA6E3.165 actin-like protein, putative
A. fumigatus CDS AfA28D1.105 possible glycine cleavage system h protein
A. fumigatus CDS afa35g10.11C isoleucyl-trna synthetase, putative
A. fumigatus CDS AfA5C11.03 crnA, nitrate permease, putative
A. fumigatus CDS AfA6E3.060c hypothetical protein
A. fumigatus CDS AfA33H4.105 hypothetical protein
A. fumigatus CDS AfA10A1.040c 2-deoxy-D-gluconate 3-dehydrogenase, putative
A. fumigatus CDS afa35g10.07C hypothetical protein
A. fumigatus CDS AfA14E5.14c hypothetical protein, conserved
A. fumigatus CDS AfA28D1.062 hypothetical protein
A. fumigatus CDS AfA33H4.100 hypothetical protein, conserved

A. fumigatus CDS AfA24A6.125 endosomal p24b protein precursor, putative
A. fumigatus CDS AfA28D1.060 hypothetical protein
A. fumigatus CDS AfA6E3.020c hypothetical protein
A. fumigatus CDS AfA31G4.010c cipA, CipA protein, putative
A. fumigatus CDS AfA34E6.035 hypothetical protein
A. fumigatus CDS AfA8D5.001c hypothetical protein
A. fumigatus CDS AfA24A6.085 cytochrome p450, putative
A. fumigatus CDS AfA5A2.055 qutR, quinate repressor, putative
A. fumigatus CDS AfA28D1.110c possible RING finger protein
A. fumigatus CDS AfA6E3.155 possible transcription factor
A. fumigatus CDS AfA31E11.010c diphthine synthase, putative
A. fumigatus CDS afa35g10.10C hypothetical protein
A. fumigatus CDS AfA6E3.150 GTP cyclohydrolase ii, putative
A. fumigatus CDS AfA33H4.125c mitochondrial processing Peptidase beta subunit, mitochondrial precursor, putative
A. fumigatus CDS AfA28D10.040c hypothetical protein
A. fumigatus CDS afa35g10.06C phenol 2-monooxygenase, putative
A. fumigatus CDS afa35g10.05b hypothetical protein
A. fumigatus CDS AfA8D5.040c hypothetical protein
A. fumigatus CDS AfA10A1.035c freA, metalloreductase, putative
A. fumigatus CDS AfA24A6.110 RNA export mediator gle1 homologue, putative
A. fumigatus CDS AfA34E6.025 ao-I, copper amine oxidase 1, putative
A. fumigatus CDS AfA33H4.050 hypothetical protein
A. fumigatus CDS AfA24A6.075 vbS, versicolorin b synthase, putative
A. fumigatus CDS AfA5A2.045 qutE, catabolic 3-dehydroquinase, putative
A. fumigatus CDS AfA31G4.005c hypothetical protein
A. fumigatus CDS AfA6E3.145 hypothetical protein, conserved
A. fumigatus CDS AfA33H4.060c histidinol-phosphate aminotransferase
A. fumigatus CDS AfA34E6.020c hypothetical protein
A. fumigatus CDS AfA6E3.140 putative secreted protein
A. fumigatus CDS AfA28D1.045 hypothetical protein, conserved
A. fumigatus CDS afa35g10.05C hypothetical protein
A. fumigatus CDS AfA31E11.005c hypothetical protein
A. fumigatus CDS AfA14E5.12c ankyrin repeat protein, putative
A. fumigatus CDS AfA24A6.105 hypothetical protein
A. fumigatus CDS AfA5A2.060c cytosolic cu/zn superoxide dismutase-related protein, putative
A. fumigatus CDS AfA24A6.065 transporter, putative
A. fumigatus CDS AfA5C5.090c hypothetical protein
A. fumigatus CDS AfA5A2.035 qutB, quinate 5-dehydrogenase, putative
A. fumigatus CDS AfA34E6.010 hypothetical protein
A. fumigatus CDS AfA19D12.060c DNA repair helicase, putative
A. fumigatus CDS AfA28D1.040c dynamin-related protein, putative
A. fumigatus CDS AfA5A2.020c qutC, 3-dehydroshikimate dehydratase, putative
A. fumigatus CDS AfA33H4.055c integral membrane protein, putative
A. fumigatus CDS AfA6E3.095 hypothetical protein

A. fumigatus CDS AfA5C5.010c hypothetical protein, conserved
A. fumigatus CDS AfA28D1.035 pyridine nucleotide-disulphide oxidoreductase family protein, putative
A. fumigatus CDS AfA34E6.015c phosphotyrosyl phosphatase activator protein, putative
A. fumigatus CDS afa35g10.04C glycerate dehydrogenase, putative
A. fumigatus CDS AfA6E3.090 cytochrome p450 (E-class), putative
A. fumigatus CDS AfA14E5.11c hypothetical protein
A. fumigatus CDS AfA28D1.030 hypothetical protein
A. fumigatus CDS AfA34E6.005 hypothetical protein
A. fumigatus CDS AfA24A6.055 gluconolactonase precursor, putative
A. fumigatus CDS AfA5A2.025 qutH, QutH protein, putative
A. fumigatus CDS AfA19D12.095c hypothetical protein
A. fumigatus CDS AfA28D1.075c hypothetical protein, conserved
A. fumigatus CDS AfA24A6.050 2-dehydro-3-deoxyphosphoheptonate aldolase, putative
A. fumigatus CDS AfA6E3.125 hypothetical protein
A. fumigatus CDS AfA5C5.085c mdr4, ABC transporter, putative
A. fumigatus CDS AfA24A6.080c integral membrane protein, putative
A. fumigatus CDS AfA6E3.120 1-phosphatidylinositol-4,5-bisphosphatephosphodiesterase 1, putative
A. fumigatus CDS AfA6E3.160c hypothetical protein
A. fumigatus CDS AfA6E3.085 nucleobase permease, putative
A. fumigatus CDS afa5c11.22c regulator of nonsense transcripts, putative
A. fumigatus CDS AfA24A6.040c hypothetical protein
A. fumigatus CDS AfA6E3.080 conserved hypothetical protein
A. fumigatus CDS afa35g10.03C oligonucleotide transporter
A. fumigatus CDS AfA19D12.015c hypothetical threonine-rich protein
A. fumigatus CDS AfA14E5.10c nadA, NADH oxidase, putative
A. fumigatus CDS AfA28D1.020 serine/threonine-protein kinase, putative
A. fumigatus CDS afa5c11.18c sagA, sagA protein
A. fumigatus CDS AfA10A1.100c transcriptional regulator, putative
A. fumigatus CDS AfA5A2.015 ethanolamine kinase, putative
A. fumigatus CDS AfA6E3.115 hypothetical protein
A. fumigatus CDS AfA5A2.010 chord containing protein homologue, putative
A. fumigatus CDS AfA6E3.110 hypothetical protein
A. fumigatus CDS AfA19D12.085 hypothetical protein with DUF292 domain, putative
A. fumigatus CDS afa5c11.21c possible zinc finger protein
A. fumigatus CDS AfA6E3.070 hypothetical protein
A. fumigatus CDS AfA19D12.080 short-chain oxidoreductase, putative
A. fumigatus CDS AfA28D10.100c mitochondrial import receptor subunit tom22 homologue, putative
A. fumigatus CDS AfA33H4.015 hypothetical protein
A. fumigatus CDS AfA24A6.035c aflR, possible aflatoxin regulatory protein
A. fumigatus CDS AfA5A2.005 2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase
A. fumigatus CDS AfA31G4.001c hypothetical protein
A. fumigatus CDS AfA24A6.030 oxidoreductase, putative
A. fumigatus CDS AfA5A2.001 hypothetical protein

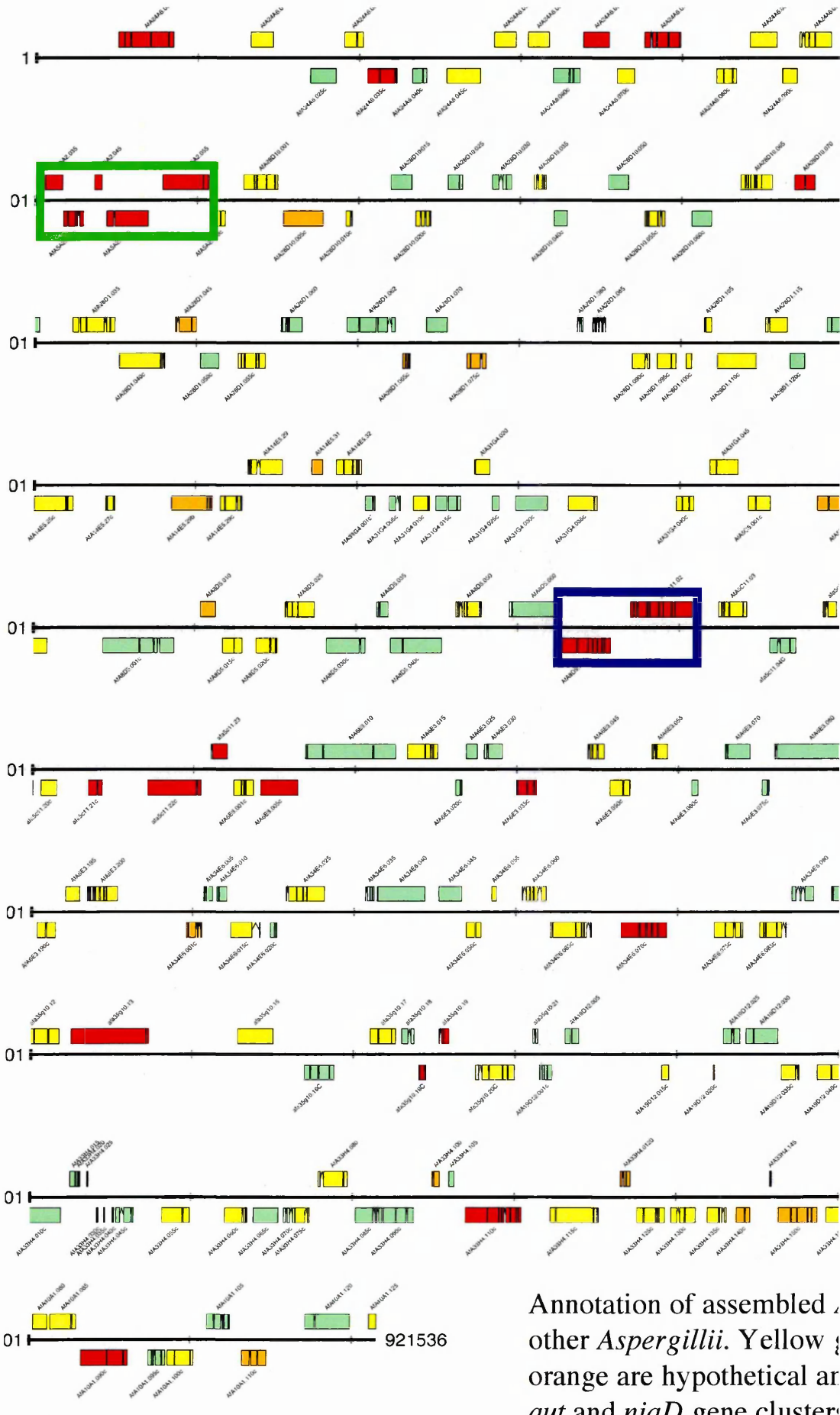
A. fumigatus CDS AfA19D12.075 SUN family protein, putative
A. fumigatus CDS afa5c11.20c dual specificity protein phosphatase 3, putative
A. fumigatus CDS AfA31E11.001c clathrin coat assembly protein, putative
A. fumigatus CDS AfA19D12.070 hypothetical protein, conserved
A. fumigatus CDS AfA6E3.050c geranylgeranyl pyrophosphate synthetase, putative
A. fumigatus CDS AfA28D1.001 serine palmitoyltransferase 2, putative
A. fumigatus CDS AfA33H4.005 possible zinc finger protein
A. fumigatus CDS afa5c11.16c hypothetical protein
A. fumigatus CDS AfA31G4.040c isoflavone reductase, putative
A. fumigatus CDS AfA19D12.105 hypothetical protein
A. fumigatus CDS AfA14E5.32 salicylate hydroxylase, putative
A. fumigatus CDS AfA28D1.100c mitochondrial 40s ribosomal protein mrp17, putative
A. fumigatus CDS AfA14E5.31 hypothetical protein, conserved
A. fumigatus CDS AfA33H4.155c outward-rectifier potassium channel tok1 homologue, putative
A. fumigatus CDS AfA6E3.055 cell division control protein cdc14, putative
A. fumigatus CDS AfA19D12.065 zinc finger protein, putative
A. fumigatus CDS AfA33H4.115c possible NOL1/NOP2/SUN family protein
A. fumigatus CDS AfA14E5.03c guanosine-diphosphatase, putative
A. fumigatus CDS AfA8D5.030c hypothetical protein
A. fumigatus CDS AfA24A6.015 peptidase, putative
A. fumigatus CDS AfA31G4.035c veA, veA protein
A. fumigatus CDS AfA14E5.29 hypothetical protein, conserved
A. fumigatus CDS AfA33H4.090c hypothetical membrane protein
A. fumigatus CDS AfA10A1.125 Ankyrin repeat protein, putative
A. fumigatus CDS AfA6E3.005c trpC, anthranilate synthase component ii, putative
A. fumigatus CDS AfA34E6.050c prohibitin, putative
A. fumigatus CDS AfA10A1.120 hypothetical protein
A. fumigatus CDS AfA33H4.0120 hypothetical protein, conserved
A. fumigatus CDS AfA14E5.21 hypothetical protein
A. fumigatus CDS AfA6E3.045 smr family protein, putative
A. fumigatus CDS AfA10A1.085 possible cation efflux protein
A. fumigatus CDS AfA19D12.055 DNA-directed RNA polymerase i, putative
A. fumigatus CDS AfA10A1.080 possible secreted cellulose-binding protein
A. fumigatus CDS AfA24A6.100c protein phosphatase ssd1 homologue, putative
A. fumigatus CDS AfA33H4.010c hypothetical protein
A. fumigatus CDS afa5c11.14c hypothetical protein, conserved
A. fumigatus CDS AfA5C5.001c odeA, oleate delta-12 desaturase
A. fumigatus CDS AfA8D5.065c niaD, nitrate reductase, putative
A. fumigatus CDS AfA24A6.005 3-hydroxy-3-methylglutaryl-coenzyme a reductase, putative
A. fumigatus CDS AfA19D12.090c possible transporter-like protein
A. fumigatus CDS AfA5A2.050c qutA, quinic acid utilization activator, putative
A. fumigatus CDS AfA5C5.080c GTPase activator protein, putative
A. fumigatus CDS AfA34E6.085c GTP-binding protein, putative
A. fumigatus CDS AfA14E5.13 peptide transporter, putative

A. fumigatus CDS AfA28D10.095 possible transcription factor IIIc-like protein
A. fumigatus CDS AfA19D12.050c u4/u6 small nuclear ribonucleoprotein hprp3-related protein, putative
A. fumigatus CDS AfA33H4.085c hypothetical protein
A. fumigatus CDS AfA10A1.075 hypothetical protein
A. fumigatus CDS AfA19D12.045 hypothetical protein, conserved
A. fumigatus CDS AfA5C5.040c la protein homolog, putative
A. fumigatus CDS afa35g10.21 hypothetical protein
A. fumigatus CDS AfA10A1.070 hypothetical protein
A. fumigatus CDS AfA6E3.030 hypothetical protein
A. fumigatus CDS AfA33H4.045c hypothetical protein
A. fumigatus CDS AfA14E5.01c possible translation initiation factor
A. fumigatus CDS afa5c11.09c glpV, glycogen phosphorylase 1, putative
A. fumigatus CDS AfA14E5.09 hypothetical protein, conserved
A. fumigatus CDS AfA14E5.08 hypothetical protein, conserved
A. fumigatus CDS AfA10A1.105 hypothetical protein
A. fumigatus CDS AfA14E5.07 hypothetical protein, conserved
A. fumigatus CDS AfA14E5.05 translation elongation factor tu precursor, mitochondrial
A. fumigatus CDS afa35g10.19 hhtA, histone h3, putative
A. fumigatus CDS AfA14E5.04 infC, translation initiation factor 3, putative
A. fumigatus CDS afa35g10.18 hypothetical protein
A. fumigatus CDS AfA6E3.190c aflatoxin b1 aldehyde reductase, putative
A. fumigatus CDS AfA28D1.065c hypothetical protein, conserved
A. fumigatus CDS afa35g10.17 transposase, putative
A. fumigatus CDS AfA14E5.02 hypothetical protein
A. fumigatus CDS afa35g10.16 hypothetical protein
A. fumigatus CDS afa35g10.15 zinc finger protein, putative
A. fumigatus CDS AfA10A1.065 60S ribosomal protein l17, putative
A. fumigatus CDS AfA6E3.025 hypothetical protein
A. fumigatus CDS afa35g10.13 aroM, pentafunctional arom polypeptide [includes: 3-dehydroquinate synthase], putative
A. fumigatus CDS AfA24A6.070c oxidoreductase, putative
A. fumigatus CDS afa35g10.12 dhp1, 5'->3'exoribonuclease, putative
A. fumigatus CDS AfA28D10.080 flavin-containing monooxygenase, putative
A. fumigatus CDS AfA10A1.060 NAD-dependant D-isomer specific 2-hydroxyacid dehydrogenase, putative
A. fumigatus CDS AfA28D1.025c ribosomal protein l15 homologue, putative
A. fumigatus CDS AfA19D12.030 hypothetical protein
A. fumigatus CDS afa5c11.12c culA, scf complex protein, putative
A. fumigatus CDS afa35g10.08 hypothetical protein
A. fumigatus CDS afa35g10.07 hypothetical protein
A. fumigatus CDS AfA6E3.015 4-coumarate coa--ligase, putative
A. fumigatus CDS AfA19D12.025 hypothetical protein
A. fumigatus CDS AfA28D10.070 possible bhlh transcription factor
A. fumigatus CDS afa35g10.02 proline-rich SH3 domain protein, putative
A. fumigatus CDS afa35g10.01 cell division protein kinase, putative

A. fumigatus CDS AfA6E3.185c malate permease, putative
A. fumigatus CDS AfA6E3.010 hypothetical protein
A. fumigatus CDS AfA5C5.075 hypothetical protein, conserved
A. fumigatus CDS AfA10A1.050 3-oxoacyl-[acyl-carrier protein] reductase, putative
A. fumigatus CDS AfA24A6.025c hypothetical protein
A. fumigatus CDS AfA6E3.001c brct domain protein, putative
A. fumigatus CDS AfA28D10.065 possible nucleoside hydrolase
A. fumigatus CDS AfA33H4.150c hypothetical protein, conserved
A. fumigatus CDS AfA10A1.045 nirA, nitrogen assimilation transcription regulator, putative
A. fumigatus CDS AfA34E6.110c hypothetical protein
A. fumigatus CDS AfA33H4.110c lacA orthologue, beta-galactosidase precursor, putative
A. fumigatus CDS afa5c11.10c ganB, G-protein alpha subunit, putative
A. fumigatus CDS AfA5C5.060 hypothetical protein, conserved
A. fumigatus CDS AfA31G4.030c hypothetical protein
A. fumigatus CDS AfA10A1.020c hypothetical protein, conserved
A. fumigatus CDS AfA19D12.005 hypothetical protein
A. fumigatus CDS AfA28D10.050 hypothetical protein
A. fumigatus CDS AfA5C5.055 mnn4, possible mannosylphosphorylation protein mnn4 protein
A. fumigatus CDS AfA10A1.030 atrE, ABC transporter, putative
A. fumigatus CDS AfA28D10.060c hypothetical protein
A. fumigatus CDS AfA10A1.095c hypothetical protein
A. fumigatus CDS AfA5C5.050 aflT, aflatoxin efflux pump AflT, putative
A. fumigatus CDS AfA34E6.001c hypothetical protein, conserved
A. fumigatus CDS AfA6E3.075c hypothetical protein
A. fumigatus CDS AfA33H4.001c hypothetical protein
A. fumigatus CDS AfA10A1.055c zinc-dependent alcohol dehydrogenase, putative
A. fumigatus CDS AfA28D10.020c 60S ribosomal protein l1-b, putative
A. fumigatus CDS AfA14E5.29c transcriptional regulator, putative
A. fumigatus CDS AfA14E5.29b hypothetical protein, conserved
A. fumigatus CDS AfA8D5.020c PfkB family carbohydrate kinase, putative
A. fumigatus CDS AfA6E3.035c fadA, guanine nucleotide-binding protein, putative
A. fumigatus CDS AfA31G4.025c hypothetical protein
A. fumigatus CDS AfA10A1.025 UBX-domain protein, putative
A. fumigatus CDS AfA5C5.047 hsp88, heat shock protein Hsp88, putative
A. fumigatus CDS AfA5C5.045 chsD, chitin synthase d
A. fumigatus CDS AfA24A6.130c hypothetical protein
A. fumigatus CDS AfA19D12.001c hypothetical protein
A. fumigatus CDS AfA28D10.055c surfeit locus protein 4 homologue, putative
A. fumigatus CDS AfA28D10.015c basic proline-rich protein
A. fumigatus CDS AfA28D10.035 adp-ribosylation factor, putative
A. fumigatus CDS AfA5A2.040c qutG, QutG protein, putative
A. fumigatus CDS AfA8D5.015c possible ribonuclease III
A. fumigatus CDS AfA10A1.015 fasciclin I family protein, putative

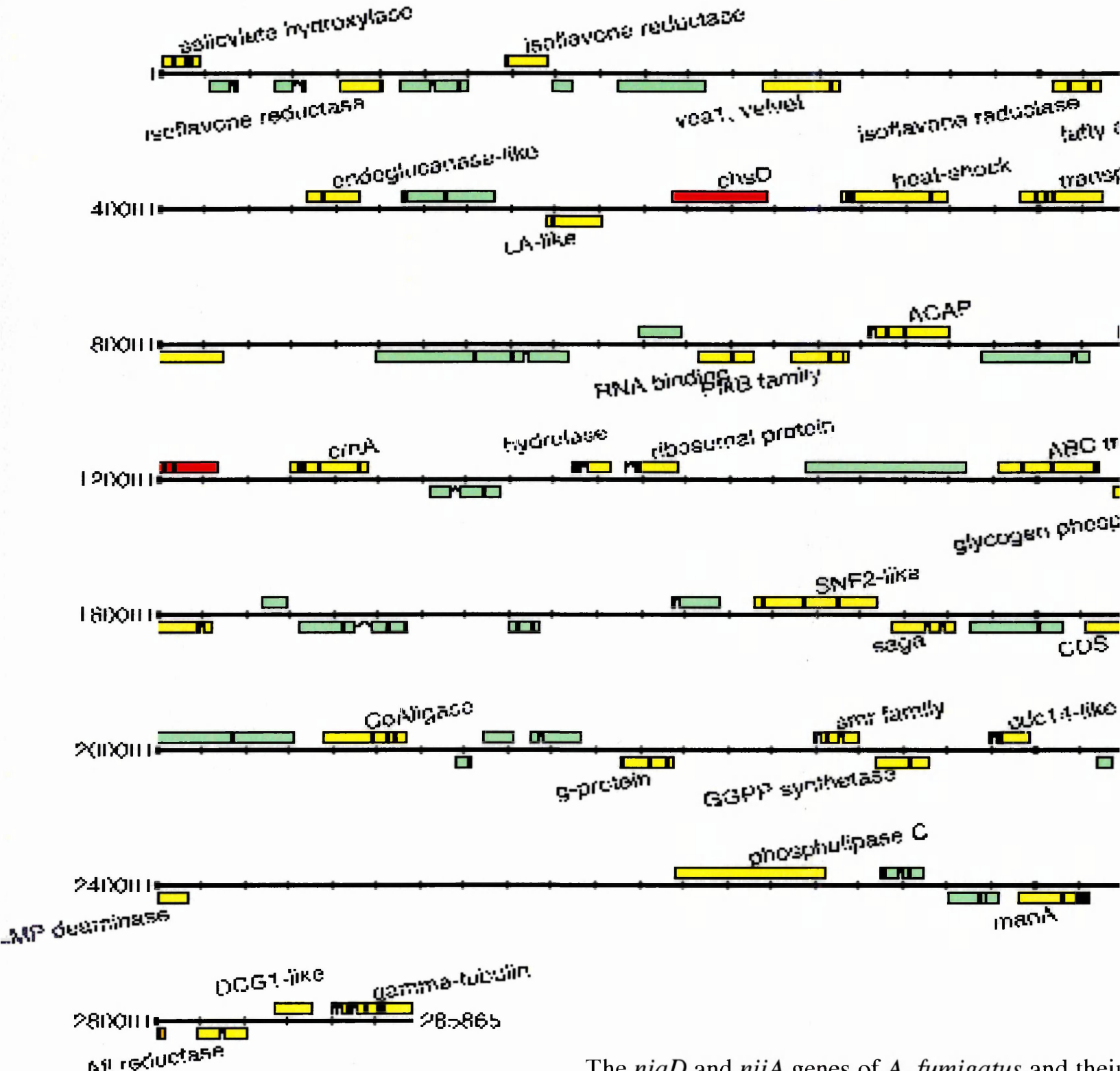
A. fumigatus CDS AfA5C5.070c hypothetical protein
A. fumigatus CDS AfA34E6.075c SET domain protein, putative
A. fumigatus CDS AfA28D10.030 hypothetical protein, putative
A. fumigatus CDS AfA19D12.040c glycosyl transferase, putative
A. fumigatus CDS AfA10A1.010 u6 snrna-associated sm-like protein, putative
A. fumigatus CDS AfA33H4.075c DNA-directed RNA polymerase, putative
A. fumigatus CDS AfA5C5.030 ligatin-like protein, putative
A. fumigatus CDS afa5c11.04C hypothetical protein
A. fumigatus CDS AfA14E5.27c possible pathogenesis-related protein precursor
A. fumigatus CDS AfA28D1.095c peroxisomal acyl-coenzyme a thioester hydrolase, putative
A. fumigatus CDS AfA28D10.025 hypothetical protein
A. fumigatus CDS AfA10A1.005 hypothetical protein
A. fumigatus CDS AfA6E3.180c hypothetical protein
A. fumigatus CDS AfA5C5.025 endoglucanase, putative
A. fumigatus CDS AfA28D1.055c oligouridylate binding protein, putative
A. fumigatus CDS AfA28D10.093c hypothetical protein
A. fumigatus CDS AfA5C5.065c 3-ketoacyl-coA thiolase, putative
A. fumigatus CDS AfA24A6.060c hypothetical protein
A. fumigatus CDS AfA19D12.035c hypothetical protein, conserved
A. fumigatus CDS AfA28D1.015c SH3-homology domain protein
A. fumigatus CDS AfA24A6.020c hypothetical protein
A. fumigatus CDS AfA6E3.100c adenosine deaminase, putative
A. fumigatus CDS AfA28D10.015 hypothetical protein
A. fumigatus CDS afa5c11.23 imp4 protein, putative
A. fumigatus CDS AfA31E11.015 transcription factor spt3, putative
A. fumigatus CDS afa14e5.22C possible transcription factor
A. fumigatus CDS AfA6E3.175c transporter, putative
A. fumigatus CDS afa35g10.18C histone h4, putative
A. fumigatus CDS AfA14E5.25c facC, carnitine acetyl transferase, putative
A. fumigatus CDS AfA6E3.135c manA, mannose 6 phosphate isomerase, putative
A. fumigatus CDS afa5c11.19 hypothetical protein
A. fumigatus CDS AfA28D10.001 flavin-containing monooxygenase, putative
A. fumigatus CDS afa5c11.17 possible swi2/snf2-like protein
A. fumigatus CDS afa5c11.16 hypothetical protein, conserved
A. fumigatus CDS AfA33H4.070c hypothetical protein
A. fumigatus CDS AfA33H4.140c hypothetical protein, conserved
A. fumigatus CDS afa5c11.13 hypothetical protein
A. fumigatus CDS afa5c11.11 hypothetical protein, conserved
A. fumigatus CDS afa5c11.10 osm1, osmotic sensitivity map kinase, putative
A. fumigatus CDS AfA34E6.100c HMG-like protein, putative
A. fumigatus CDS AfA10A1.090c exgO, exo-1,3-beta-D-glucanase, putative
A. fumigatus CDS AfA8D5.060 hypothetical protein
A. fumigatus CDS AfA31G4.045 rad57 protein, putative
A. fumigatus CDS AfA28D10.115c acetoacetyl-coa synthetase, putative

Supplementary Figure 2a



Annotation of assembled A
other *Aspergillii*. Yellow g
orange are hypothetical an
qut and *niaD* gene cluster

Supplementary figure 2b



The *niaD* and *niiA* genes of *A. fumigatus* and thei
sequence of five finished and completed overlapp

A. fumigatus CDS AfA33H4.165 tgf beta receptor associated protein 1 homologue, putative
A. fumigatus CDS afa5c11.08 ABC transporter, related to N. crassa adrenoleukodystrophy-related protein
A. fumigatus CDS afa5c11.07 hypothetical protein
A. fumigatus CDS afa5c11.06 60S ribosomal protein l5, putative
A. fumigatus CDS afa5c11.05 epoxide hydrolase, putative
A. fumigatus CDS AfA28D1.120c hypothetical protein
A. fumigatus CDS AfA34E6.095 hypothetical protein
A. fumigatus CDS afa5c11.02 niiA, nitrite reductase, putative
A. fumigatus CDS AfA28D10.090c qutD-like transporter, putative
A. fumigatus CDS AfA34E6.090 hypothetical protein
A. fumigatus CDS afa35g10.20C mdR, mfs-family multidrug resistance protein, putative
A. fumigatus CDS AfA31E11.020c Bli-3 protein, putative
A. fumigatus CDS AfA19D12.100c hypothetical protein, conserved
A. fumigatus CDS AfA8D5.050 isocitrate dehydrogenase, putative
A. fumigatus CDS AfA33H4.135c fibrillarin, putative
A. fumigatus CDS afa35g10.16C hypothetical protein
A. fumigatus CDS AfA14E5.23c jumonji family transcription factor, putative
A. fumigatus CDS AfA28D10.010c u6 snrna-associated sm-like protein, putative
A. fumigatus CDS AfA14E5.19c acyl CoA binding protein, putative
A. fumigatus CDS AfA34E6.070c facB, acetate regulatory DNA binding protein FacB, putative
A. fumigatus CDS AfA31G4.015c hypothetical protein

References

Adams, TH. & Weiser, JK. (1999), "Asexual sporulation: conidiation", in Oliver, RP. & Schweizer, M. (eds), *Molecular fungal biology*, Cambridge University Press, Cambridge

Allen, MJ., Voelker, DR. & Mason, RJ. (2001), "Interactions of surfactant proteins A and D with *Saccharomyces cerevisiae* and *Aspergillus fumigatus*", *Infect Immun*, vol. 69, no. 4, pp.2037-2044

Altschul, SF. *et al*,(1990), "Basic local alignment search tool", *J Mol Biol*, vol. 215, pp.403-410

Amaar, YG. & Moore, MM (1998), "Mapping of the nitrate-assimilation gene cluster (*crnA-niiA-niaD*) and characterisation of the nitrite reductase gene (*niiA*) in the opportunistic fungal pathogen *Aspergillus fumigatus*", *Curr Genet*, vol. 33, pp.206-215

Andrade, A. *et al*, (2002), 6th European Conference on Fungal Genetics, vol. Abstract IV, pp.3

Apweiler, R. *et al*, (2000), "InterPro--an integrated documentation resource for protein families, domains and functional sites", *Bioinformatics*, vol. 16, no. 12, pp.1145-1150

Arruda, LK. Mann, BJ. Chapman, MD. (1992), "Selective expression of a major allergen and cytotoxin, Asp f1, in *Aspergillus fumigatus*. Implications for the immunopathogenesis of *Aspergillus*-related diseases", *Journ Immunol*, vol.149, pp.3354-3359

Ashburner, M. *et al*, (2000), "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", *Nat Genet*, vol. 25, no. 1, pp.25-29

Badger, JH. & Olsen, GJ. (1999), "CRITICA: coding region identification tool invoking comparative analysis", *Mol Biol Evol*, vol. 16, pp.512-524

Bairoch, A. (1991), "PROSITE: a dictionary of sites and patterns of proteins", *Nucleic Acid Res*, vol. 30, pp.276-280

Balciunas, D. & Ronne, H. (2000), "Evidence of domain swapping within the jumonji family of transcription factors", *Trends Biochem Sci*, vol. 35, no. 6, pp.274-276

Banerjee, B. *et al*, (2002), "C-terminal cysteine residues determine the IgE binding of Aspergillus fumigatus allergen Asp f 2", *J Immunol*, vol. 169, no. 9, pp.5137-5144

Bennett, JH. (1842), "On the parasitic vegetable structures found growing in living animals", *Trans. R. Soc. Edinburgh*, vol. 15, pp.277-279

Benson, DA. *et al*, (2000), "GenBank", *Nucleic Acids Res*, vol. 28, no. 1, pp.15-18
Bentley, R. (1990), "The shikimate pathway--a metabolic tree with many branches", *Crit Rev Biochem Mol Biol*, vol. 25, no. 5, pp.307-384

Berkinshaw, JH. *et al*, (1931), "Studies in the biochemistry of micro-organisms. Part IX-On the production of mannitol from glucose by species *Aspergillus*", *Philos Trans R Soc Lond B Biol Sci*, vol, 22, pp.153-171

Bertrand, H. (1995), "Senescence is coupled to induction of an oxidative phosphorylation stress response by mitochondrial DNA mutations in *Neurospora*", *Canadian Journ Bot*, vol. 73, pp.189-204

Birch, M. *et al*, (1996), "Evidence of multiple phospholipase activities of *Aspergillus fumigatus*", *Infect Immun*, vol. 64, pp. 751-755

Birch, M. *et al*, (1997), "Prevalence of phthoic acid in *Aspergillus* spp.", *J Met Vet Mycol*, vol. 35, pp.143-154

Bonfield, JK., Smith, K & Staden, R. (1995), "new DNA sequence assembly program", *Nucleic Acids Res*, vol. 23, no. 24, pp.4992-4999

Borodovsky, M. & McIninch, J, (1993), "GeneMark: parallel gene recognition for both DNA strands", *Computers Chemistry*, vol. 17, pp.123-133

Bouchara, JP. *et al* (1993), "Extracellular fibrinogenolytic enzyme of *Aspergillus fumigatus*: substrate dependant variations in the proteinase synthesis and characterisation of the enxyme", *FEMS Immuno Med Mocrobiol*, vol. 7, pp.81-92

Bouck, J. *et al*, (1998), "Analysis of the quality and utility of random shotgun sequencing at low redundancies", *Genome Res*, vol. 8, no. 10, pp.1074-1084

Brajtburg, J. *et al*, (1985), "Involvement of oxidative damage in erythrocyte lysis induced by amphotericin B", *Antimicrob Agents Chemother*, vol. 27, no. 2, pp.172-176

Braun, BR., Kadosh, D. & Johnson, AD. (2001), "NRG1, a repressor of filamentous growth in *C. albicans*, is down regulated during filament induction", *EMBO J*, vol. 20, pp.4753-4761

Burke, DT., Carle, GF. & Olsen, MV, (1987), "Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors", *Science*, vol. 236, pp.806-812

Burke, DT. (1990), "YAC cloning: options and problems", *Genet Anal Tech Appl*, vol. 7, no. 5, pp.94-99

Cai, WW. *et al*, (1998), "An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization", *Genomics*, vol. 54, no. 3, pp.387-397

Calera, JA. *et al*, (1997), "Cloning and disruption of the antigenic catalase gene of *Aspergillus fumigatus*", *Infect Immun*, vol. 65, pp.4718-4724

Carpenter, EP. *et al*, (1998), "Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis", *Nature*, vol. 394, pp.299-302

Cawley, EP. (1947), "Aspergillosis and the aspergilli", *Arch Intern Med*, vol. 80, pp.423-434

Cawley, SE. Wirth, AI. & Speed, TP. (2001), "Phat--a gene finding program for Plasmodium falciparum", *Mol Biochem Parasitol*, vol. 118, no. 2, pp.167-174

Chazalet, V. *et al*, (1998), "Molecular typing of environmental and patient isolates of *Aspergillus fumigatus* from various hospital settings", *J Clin Microbiol*, vol. 36, no. 6, pp.1494-1500

Clements, JS. & Peacock, JE, (1990), "Amphotericin B revisited: reassessment of toxicity", *Am J Med*, vol. 88, no. 5, pp.22-27

Clemons, KV. *et al*, (2000), "Pathogenesis I: interactions of host cells and fungi", *Med Mycol*, vol. 38, Suppl. 1, pp.99-111

Collins, J. & Hohn, B. (1978), "Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads", *Proc Natl Acad Sci USA*, vol. 75, no. 9, pp.4242-4246

Deacon, JW (1997), "Modern Mycology", Blackwell Publishing, Oxford

Delcher, AL. *et al* (1999), "Improved microbial gene identification with GLIMMER", *Nucleic Acids Res*, vol. 27, pp.4636-4641

Del Sorbo, G. *et al*, (1997), "Multidrug resistance in *Aspergillus nidulans* involves novel ATP-binding cassette transporters", *Mol Gen Genet*, vol. 254, no. 4, pp.417-426

Denning, DW. (1987), "Aflatoxin and human disease. A review", *Adverse Drug Reactions and Acute Poisoning Reviews*, vol. 4, pp. 175-180

Denning, DW. *et al*, (1994), "NIAID Mycoses Study Group Multicenter Trial of Oral Itraconazole Therapy for Invasive Aspergillosis", *Am J Med*, vol. 97, no. 2, pp.135-144

Denning, DW. (1996), "Therapeutic outcome in invasive aspergillosis", *Clin Infect Dis*, vol. 23, no. 3, pp.608-615

Denning, DW. *et al*, (1997), "Itraconazole resistance in *Aspergillus fumigatus*", *Antimicrob Agents Chemother*, vol. 41, no. 6, pp.1364-1368

Denning, DW. (1998), "Invasive aspergillosis", *Clinical Infectious Disease*, vol. 26, no. 4, pp.781-803

Devlin, PF. & Kay, SA. (2001), "Circadian photoperception", *Annu Rev Physiol*, vol. 63, pp.677-694

Diamond, RD. *et al*, (1978), "Damage to hyphal forms of fungi by human leukocytes *in vitro*: a possible host defence mechanism in aspergillosis and mucormycosis", *Am J Pathol*, vol. 91, pp.313-328

Di Paolo, N. *et al*, (1993), "Acute renal failure from inhalation of mycotoxins", *Nephron*, vol. 64, no. 4, pp.621-625

Dunham, I. *et al*, (1997), "Bacterial Cloning Systems" in: Birren, B *et al* (eds), Genome Analysis: A Laboratory Manual, Vol. 3, Cloning Systems, Cold Spring Harbor Press, USA

Du Pont, B. (1990), "Itraconazole therapy in aspergillosis: study in 49 patients", *J Am Acad Dermatol*, vol. 23, no. 3, pp.607-614

Dziejman, M. *et al*, (2002), "Comparative genomic analysis of *Vibrio Cholerae*: genes that correlate with cholera endemic and pandemic disease", *PNAS*, vol. 99, no. 3, pp.1556-1561

Eaton, DL. & Gallagher, EP. (1994), "Mechanisms of aflatoxin carcinogenesis", *Annual Rev Pharmacol Tox*, vol. 34, pp.135-72

Ebina, K. *et al*, (1994), "Cloning and nucleotide sequence of cDNA encoding Asp-hemolysin from *Aspergillus fumigatus*", *Biochim Biophys Acta*, vol. 1219, pp.148-150

Esser, K. (1990), "Molecular aspects of ageing: facts and perspectives". In: Hawksworth, DL (ed) *Frontiers in Mycology*, CAB International, Wallingford

Falquet, L. *et al*, (2002), "The PROSITE database, its status in 2002", *Nucleic Acid Res*, vol. 30, pp.235-238

Fields, S. & Song, O. (1989), "A novel genetic system to detect protein-protein interactions", *Nature*, vol. 340, pp.245-246

Fluckiger, S. *et al*, (2002), "Comparison of the crystal structures of the human manganese superoxide dismutase and the homologous *Aspergillus fumigatus* allergen at 2-A resolution", *J Immunol*, vol. 168, no. 3, pp.1267-1272

Frengen, E. *et al*, (1999), "A modular, positive selection bacterial artificial chromosome vector with multiple cloning sites", *Genomics*, vol. 58, no. 3, pp.250-253

Frishman, D. *et al*, (1998), "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes", *Nucleic Acids Res*, vol.26, pp.2941-2947

Frisvad, JC. & Samson, RA. (1990), "Chemotaxonomy and morphology of *Aspergillus fumigatus* and related taxa", in: Samson RA, Pitt JI (eds), *Modern concepts in Penicillium and Apsergillus classification*, Plenum Press, New York

Gairdner, WT. (1856), "Clinical retrospect of cases treated during the session 1855-1856 (November-March); including remarks upon the more important fatal cases, and upon the cases of inflammation of the lungs treated during that period", *Edinburgh Med. Journ,* vol. 1, pp.969-984

Galagan, JE. *et al*, (2003), "The genome sequence of the filamentous fungus *Neurospora crassa*", *Nature*, vol. 442, pp.859-868

Gardner, M. *et al* (2002), "Genome sequence of the human malaria parasite *Plasmodium falciparum*", *Nature*, vol. 419, pp.498-511

Gavin, AC. *et al*, (2002), "Functional organization of the yeast proteome by systematic analysis of protein complexes", *Nature*, vol. 415, pp.141-147

Glayzer, DC. *et al*, (1995), "The isolation of Ant1, a transposable element from *Aspergillus niger*", *Mol Gen Genet*, vol. 249, no. 4, pp.432-438

Glockner, G. *et al*, (2002), "Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*", *Nature*, 418, pp.79-85

Goffeau, A. *et al*, (1996), "Life with 6000 genes", *Science*, vol. 274, pp.546, 563-567

Goodley, JM. Clayton, YM. & Hay, RJ. (1993), "Environmental sampling for aspergilli during building construction on a hospital site", *J Hosp Infect*, vol. 26, no. 1, pp.27-35

Gray, NS. *et al*, "Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors", *Science*, vol. 281, pp.533-538

Green, ED. *et al*, (1991), "Detection and characterization of chimeric yeast artificial chromosome clones", *Genomics*, vol. 11, no. 3, pp.658-669

Groll, AH. *et al*, (1996), "Trends in the postmortem epidemiology of invasive fungal infections at a university hospital", *J Infect*, vol. 33, no. 1, pp.23-32

Gunel-Ozcan, A. *et al*, (1997), "Salmonella typhimurium aroB mutants are attentuated in BALB/c mice", *Microb Pathog*, vol. 23, no. 5, pp.311-316

Hamer, L. *et al*, (2001), "Regions of microsynteny in *Magnaporthe grisea* and *Neurospora crassa*", *Fungal Genet Biol*, vol. 33, no. 2, pp.137-143

Hearn, VM.; Wilson, EV. & MacKenzie DWR. (1992), "Analysis of *Aspergillus fumigatus* catalases possessing antigenic activity", *J Med Microbiol*, vol. 36, pp.61-67

Hensel, M. *et al*, (1995), "Simultaneous identification of bacterial virulence genes by negative selection", *Science*, vol. 269, pp.400-403

Hinson, KFW, Moon, AJ & Plummer, NS. (1952) "Broncho-pulmonary aspergillosis. A review and a report of eight new cases", *Thorax*, vol. 7, pp.317-333

Hohn, B. & Murray, K. (1977), "Packaging recombinant DNA molecules into bacteriophage particles in vitro", *Proc Natl Acad Sci USA*, vol. 74, no. 8, pp.3259-3263

Hope, IA. (1994), "PES-1 is expressed during early embryogenesis in Caenorhabditis elegans and has homology to the fork head family of transcription factors", *Development*, vol. 120, no. 3, pp.505-514

Howell, AM. & Rose, AM. (1990), "Essential genes in the hDf6 region of chromosome I in *Caenorhabditis elegans*", *Genetics*, vol. 126, no. 3, pp.583-592

Ioannou, PA. *et al*, (1994), "A new bacteriophage P1-derived vector for the propagation of large human DNA fragments", *Nat Genet*, vol. 6, no. 1, pp.84-89

Ireland, LS. *et al*, (1998), "Molecular cloning, expression and catalytic activity of a human AKR7 member of the aldo-keto reductase superfamily: evidence that the major 2-carboxybenzaldehyde reductase from human liver is a homologue of rat aflatoxin B1-aldehyde reductase", *Biochem J*, vol. 332, pp.21-34

Jez, JM., Flynn, TG. & Penning, TM. (1997), "A new nomenclature for the aldo-keto reductase superfamily", *Biochem Pharmacol*, vol. 54, no. 6, pp.639-647

Kelkar, HS. *et al*, (2001), "The *Neurospora crassa* genome: cosmid libraries sorted by chromosome", *Nature*, vol. 422, pp.979-990

Keller, NP. & Hohn, TM. (1997), "Metabolic Pathway Gene Clusters in Filamentous Fungi", *Fungal Genet Biol*, vol. 21, no. 1, pp.17-29
Kelly, JD. *etal*, (1997), "Aflatoxin B1 activation in human lung", *Toxicol Appl Pharmacol*, vol. 144, no. 1, pp.88-95

Kim, UJ. *et al*, (1992), "Stable propagation of cosmid sized human DNA inserts in an F factor based vector", *Nucleic Acids Res*, vol. 20, no. 5, pp.1083-1085

Kishore, GM. & Shah, DM. (1988), "Amino acid biosynthesis inhibitors as herbicides", vol. 57, pp.627-663

Kodzius, R. *et al*, (2003), "Rapid identification of allergen-encoding cDNA clones by phage display and high-density arrays", *Comb Chem High Throughput Screen*, vol. 6, no.2, pp.147-154

Krogh, A. *et al*, (2001), "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes", *J Mol Biol*, vol. 305, no. 3, pp.567-580

Kurup, VP. *et al*, (2000), "Selected recombinant *Aspergillus fumigatus* allergens bind specifically to IgE in ABPA", *Clin Exp Allergy*, vol. 30, no. 7, pp.988-993

Larin, Z. Monaco, AP & Lehrach H. (1996), "Generation of large insert YAC libraries", *Methods Mol Biol*, vol. 54, pp.1-11

Latge, JP. *et al*, (1991), "The 18-kilodalton antigen secreted by *Aspergillus fumigatus*", *Infect Immun*, vol. 59, pp.2586-2594

Latge, JP. & Calderone, R. (2002), "Host-microbe interactions: fungi invasive human fungal opportunistic infections", *Curr Opin Microbiol*, vol. 5, no. 4, pp.355-358

Lowe, TM. & Eddy, SR. (1997), "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence", *Nucleic Acids Res*, vol. 25, no. 5, pp.955-964

Lukashin, AV. & Borodovsky, M. (1998), "GeneMark.hmm: new solutions for gene finding", *Nucleic Acids Res*, vol. 26, pp1107-1115

Link, HF. (1809), "Observationes in ordines plantarum naturales", *Igesellschaft Naturforschender Freunde zu Berlin*, Magazin.

Lorenz, J. (2002), "Genomic approaches to fungal pathogenicity", *Curr Opin Microbiol*, vol. 5, no. 4, pp.372-378

Lynch, AS., Briggs, D. & Hope, IA. (1995), "Developmental expression pattern screen for genes predicted in the C. elegans genome sequencing project", *Nat Genet*, vol. 11, no. 3, pp.309-313

Marshall, MA. & Timberlake, WE. (1991), "*Aspergillus nidulans* wetA activates spore-specific gene expression", *Moll Cell Biol,* vol.11, no.1, pp. 55-62

Markaryan, A. *et al* (1994), "Purification and characterisation of an elastinolytic metalloprotease from *Aspergillus fumigatus* and immunoelectron microscopic evidence of secretion of this enzyme by the fungus invading murine lung", *Infect Immun* vol. 61, pp.2149-2157

Mayer, AC. (1815), *Deutsch Arch Physiololgy*, Cited by Renon.

Megson, GM. *et al*, (1994), "The application of serum mannitol determinations for the diagnosis of invasive pulmonary aspergillosis in bone marrow transpant patients", *J Infect*, vol. 28, pp.58

Mendes-Giannini, MJ. *et al*, (2000), "Pathogenesis II: fungal responses to host responses: interaction of host cells with fungi", *Med Mycol*, vol. 38, Suppl. 1, pp.113-123

Micheli, PA. (1729), "Nova plantarum genera juxta Tournafortii methodum diposita" Flocrence:Bernado paperini

Monod, M. *et al*, (1995), "The secreted proteases of the pathogenic species of *Aspergillus* and their possible role in virulence", *Canadian Journal of Botany*, vol. 73, pp.1081-1086

Monod, M. *et al*, (2002), "Secreted proteases from pathogenic fungi", *Int J Med Microbiol*, vol. 292, no. 5-6, pp.405-419

Moore, CB. *et al*, (2000), "Antifungal drug resistance in Aspergillus", *J Infect*, vol. 41, no. 3, pp.203-220

Moss, MO. (1994), "Biosynthesis of *Aspergillus* toxins-non-aflotoxins", in: Powell KA, Peberdy JF and Renwick A, (eds), *The Genus Aspergillus*, Plenum Press, New York

Mosquera, J. *et al*, (2002), "In vitro interaction of terbinafine with itraconazole, fluconazole, amphotericin B and 5-flucytosine against Aspergillus spp", *J Antimicrob Chemother*, vol. 50, no. 2, pp.189-194

Muhlschlegal, F. *et al*, (1998), "Molecular mechanisms of virulence in fungus-host interactions for Aspergillus fumigatus and Candida albicans", vol. 36, Suppl 1, pp.238-248

Mulder, NJ. *et al*, (2003), "The InterPro Database, 2003 brings increased coverage and new features", *Nucleic Acids Res*, vol. 31, no. 1, pp.315-318

Mullbacher, A. Waring, .P & Eichner, RD, (1985), "Identification of an agent in cultures of *Aspergillus fumigatus* displaying anti-phagocytic and immunomodulating activity *in vitro*", *J Gen Microbiol*, vol. 131, pp.1251-1258

Nascimento, AM. *et al*, (2003), "Multiple resistance mechanisms among *Aspergillus fumigatus* mutants with high-level resistance to itraconazole", *Antimicrob Agents Chemother*, vol. 47, no. 5, pp.1719-1726

Neal, GE. *e al*, (1998), "Metabolism and toxicity of aflatoxins M1 and B1 in human-derived in vitro systems", *Toxicol Appl Pharmacol*, vol. 151, no. 1, pp.152-158

Neil, DL. *et al*, (1990), "Structural instability of human tandemly repeated DNA sequences cloned in yeast artificial chromosome vectors", *Nucleic Acids Res*, vol. 18, no. 6, pp.1421-1428

Neuveglise, C. *et al*, (1996), "Afut1, a retrotransposon-like element from *Aspergillus fumigatus*", *Nucleic Acids Res*, vol. 24, no. 8, pp.1428-1434

Ng, TT *et al* (1994), "Hydrocortisone-enhanced growth of *Aspergillus* spp.: implications for pathogenesis", *Microbiology*, vol. 140, pp. 2475-2479

Nielsen, H., Brunak, S. & von Heijne, G. (1999), "Machine learning approaches for the prediction of signal peptides and other protein sorting signals", *Protein Eng*, vol. 12, no. 1, pp.3-9

Oliver, RP. & Schwiezer, M. (2001), "Molecular Fungal Biology", Cambridge University Press, Cambridge

Osoegawa, K. *et al*, (1998), "An improved approach for construction of bacterial artificial chromosome libraries", *Genomics*, vol. 52, no. 1, pp.1-8

Paris, S. & Latge, JP. (2001), "Afut2, a new family of degenerate gypsy-like retrotransposon from *Aspergillus fumigatus*", *Med Mycol*, vol. 39, no. 2, pp.195-198

Parkhill, J. (2002), "Annotation of Microbial Genomes", *Meth Microbiol*, vol. 33, pp.3-26

Pearson, WR. & Lipman, DJ. (1988), "Improved tools for biological sequence comparison", *Proc Natl Acad Sci*, vol. 85, pp.2444-2448

Pederson, C., Wu, B. & Giese, H. (2002), "A Blumeria graminisf.sp. hordei BAC library--contig building and microsynteny studies", *Curr Genet*, vol. 42, no. 2, pp.103-113

Peterson, DG. *et al*, (2000), "Construction of Plant Bacterial Artificial Chromosome Libraries: An Illustrated Guide", *Journ Agricultural Genomics*.

Pierce, JC., Sauer, B & Sternberg, N. (1992), "A positive selection vector for cloning high molecular weight DNA by the bacteriophage P1 system: improved cloning efficacy", *Proc Natl Acad Sci USA*, vol. 89, no. 6, pp.2056-2060

Pitt, JI. (1994), "The current role of *Aspergillus* and *Penicillium* in human and animal health", *J Med Vet Mycol*, vol. 32, pp.17-32

Prieto, R., Yousibova, GL. & Woloshuk, CP. (1996), "Identification of aflatoxin biosynthesis genes by genetic complementation in an *Aspergillus flavus* mutant lacking the aflatoxin gene cluster", *Appl Environ Microbiol*, vol. 62, no. 10, pp.3567-3571

Rankin, NE. (1953), "Disseminated aspergillosis and moniliasis associated with agranulosytosis and antibiotic therapy", *BMJ*, vol. 1, pp.918-919

Rayer. (1942) Fioriep's N. Notixen.

Reichard, U. *et al* (1997), "Virulence of and aspergillopepson-deficient mutant of *Aspergillus fumigatus* and evidence for another aspartic Proteinase linked to the fungal cell wall", *J Med Vet Mycol*, vol. 35, pp.189-196

Robinson, C., Baker, SF. & Garrod, DR. (2001), "Peptidase allergens, occludin and claudins. Do their interactions facilitate the development of hypersensitivity reactions at mucosal surfaces?", vol. 31, no. 2, pp.186-192

Roilides, E. *et al*, (1994), "Antifungal activity of elutriated human monocytes against *Aspergillus fumigatus* hyphae: enhancement by granulocyte-macrophage colony-stimulating factor and interferon-γ", *J Infect Dis*, vol. 170, pp. 894-899

Salzberg, SL. *et al*, (1999), "Interpolated Markov models for eukaryotic gene finding", *Genomics*, vol. 59, no. 1, pp.24-31

Samson, RA. (1992), "Current taxonomic schemes of the genus *Aspergillus* and its teleomorphs", *Biotechnology*, vol. 23, pp.355.390

Schaffner, A.; Douglas, H & Braude, A. (1982), "Selectiveprotection against conidia by mononuclear and against mycelia by polymorphonuclear phagocytes in resistance to *Aspergillus*: observations on these two lines of defence *in vivo* and *in vitro* with human and mouse phagocytes", *J Clin Invest*, vol. 69, pp.617-631

Schein, JE. *et al*, (1993), "The use of deficiencies to determine essential gene content in the let-56-unc-22 region of *Caenorhabditis elegans*", *Genome*, vol. 36, no. 6, pp.1148-1156

Schmitt, HJ. *et al*, (1991), "Combination therapy in a model of pulmonary aspergillosis", *Mycoses*, vol. 34, no. 7-8, pp.281-285

Seoighe, C. *et al*, (2000), "Prevalence of small inversions in yeast gene order evolution", *Proc Natl Acad Sci USA*, vol. 97, no. 26, pp.14433-4437

Shizuya, H. *et al*, (1992), "Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector", Proc *Natl Acad Sci USA*, vol. 89, no. 18, pp.8794-8797

Schultz, J. *et al*, (1999), "SMART, a simple modular architecture research tool: identification of signalling domains", *Proc Natl Acad Sci USA*, vol. 95, no. 11, pp.5857-5864

Smith, JM. *et al*, (1993), "Construction and pathogenicity of *Aspergillus fumigatus* mutants that do not produce the ribotoxin restriction", *Mol Microbiol*, vol. 9, pp.1071-1077

Spellman, PT. *et al*, (1998), "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization", *Mol Biol Cell*, vol. 9, no. 12, pp.3273-3297

Strong, SJ. *et al*, (1997), "marked improvement of PAC and BAC cloning is achieved using electroelution of pulsed-field gel-separated partial digests of genomic DNA", *Nucleic Acids Res*, vol. 25, pp.3959-3961

Sutton, P. *et al*, (1996), "Exacerbation of invasive aspergillosis by the immunosuppressive fungal metabolite, gliotoxin", *Immno Cell Biol*, vol. 74, pp.318-322

Staib, F. (1984), "Ecological and epidemiological aspects of aspergilli pathogenic for man and animal in Berlin (West)", *Zentralbl Bakteriol Mikrobiol Hyg [A]*, vol. 257, no. 2, pp.240-245

Sturtevant, JE. & Latge, JP. (1992), "Interactions between conidia of *Aspergillus fumigatus* and human complement component C3", *Infect Immun*, vol. 60, no. 5, pp.1913-1918

Taksuka, T. *et al*, (1997), "Possible contribution of catalase to pathogenicity of *Aspergillus fumigatus*" in: Proceedings of the 13th Congress of the International Society for Human and Animal Mycology. Parma, Italy

Tavozoie, S. *et al*, "Systematic determination of genetic network architecture", *Nat Genet*, vol. 22, no. 3, pp.281-285

The International Human Genome Sequencing Consortium (2001), "Initial sequencing and analysis of the Human Genome", *Nature*, vol. 409, pp.860-921

Tobin, MB. Peery, RB & Skatrud PL (1997), "Genes encoding multiple drug resistance-like proteins in *Aspergillus fumigatus* and *Aspergillus flavus*", *Gene*, vol. 24, no. 200, pp.11-23

Toder, OC. & Turgeon, BG. (2001), "Fungal Genomics and Pathogenicity", *Current Opinion in Plant Biology*, vol. 4, no. 4, pp.315-321

Tronchin, G. *et al*, (1993), "Interaction between *Aspergillus fumigatus* and basement membrane laminin: binding and substrate degradation", *Biol Cell*, vol. 77, pp. 201-208
Tronchin, G. *et al*, (2002), "Purification and partial characterization of a 32-kilodalton sialic acid-specific lectin from *Aspergillus fumigatus*", *Infect Immun*, vol. 70, no. 12, pp.6891-6895
Uetz, P. *et al*, (2000), "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*", *Nature*, vol. 403, pp.623-627

Venter, JC., Smith, HO. & Hood, L. (1996), "A new strategy for genome sequencing", *Nature*, vol. 381, pp.364-366

Vertut-Doi, A., Ohnishi, SI. & Bolard, J. (1994), "The endocytic process in CHO cells, a toxic pathway of the polyene antibiotic amphotericin B", *Antimicrob Agents Chemother*, vol. 38, no. 10, pp.2372-2379

Virchow, VR. (1856), "Britrage zur lehre con den beim menschen corkommenden pflanzlichen parasiten", *Virchow's Archive*, vol. 557.

Waring, P. (1990), "DNA fragmentation induced in macrophages by gliotoxin does not require protein synthesis and is preceded by raised inositol triphosphate levels", *J Biol Chem*, vol. 265, pp.14476-14480

Warris, A., Weemaes, CM. & Verweij, M. (2002), "Multidrug resistance in *Aspergillus fumigatus*", *N Engl J Med*, vol. 347, no. 26, pp.2173-2174

Wasan, KM. *et al* (1993), "Roles of liposome composition and temperature in distribution of amphotericin B in serum lipoproteins", *Antimicrob Agents Chemother*, vol. 37, no. 2, pp.246-250

Waterson, R. & Sulston, J. (1995), "The genome of *Caenorhabditis elegans*", *Proc Natl Acad Sci USA*, vol. 92, no. 24, pp.10836-10840

Wheaton, SW. (1890), "Case primarily of tubercle, in which a fungus (aspergillus) grew in the bronchi and lung, stimulating actinomycosis", *Path Trans*, vol. 41, pp.34-37

Wilson, DK. *et al*, (1993), "Refined 1.8 A structure of human aldose reductase complexed with the potent inhibitor zopolrestat", *Proc Natl Acad Sci USA*, vol. 90, no.21, pp.9847-9851

Wojnarowski, C. *et al*, (1997), "Sensitization to *Aspergillus fumigatus* and lung function in children with cystic fibrosis", *Am J Respir Crit Care Med*, vol. 155, no. 6, pp. 1902-1907

Woloshuk, CP. & Prieto, R. (1998), "Genetic organization and function of the aflatoxin B1 biosynthetic genes", *FEMS Microbiol Lett*, vol. 160, no. 2, pp.169-176

Wong, B. *et al*, (1989), "Increased amounts of the *Aspergillus* metabolite D-mannitol in tissue and serum of rats with experimental aspergillosis", *J Infect Dis*, vol. 160, pp.95-103

Yu, JH.& Leonard, TJ. (1995), "Sterigmatocystin biosynthesis in *Aspergillus nidulans* requires a novel type I polyketide synthase", *J Bacteriol*, vol. 177, no. 16, pp.4792-4800

Zhu, H. *et al*, (2000), "Analysis of yeast protein kinases using protein chips", *Nat Genet*, vol. 28, pp.283-289

Appendix 1

Year	Author	Report
1729	Micheli	First description of Aspergillosis
1815	Mayer	Air sac and pulmonary infection in a jackdaw
1842	Bennett	Aspergilloma complicating tuberculosis with *Aspergillus* sputum
1842	Rayer	Pleural aspergillosis
1844	Fresenius	Clear species description of *A. fumigatus* (isolated from air sacs and bronchi of a great bustard)
1879	Leber	Aspergillus keratitis following chaff entering the eye
1886	Bostroem	Cutaneous aspergillosis
1887	Popoff	Allergic pulmonary disease caused by *Aspergillus*
1890	Wheaton	Aspergillus tracheobronchitis
1891	Zarniko	Maxillary sinus aspergillosis
1897	Oppe	Sphenoid sinus aspergillosis
1931	Just	Cerebral aspergillosis
1936	Shaw and Warthen	Aspergillosis of the bone
1947	Cawley	Invasive aspergillosis (probable) complicating chronic granulomatous disease (with meningitis)
1950	Zimmerman	Native valve aspergillus endocharditis
1952	Hinson, Moon and Plummer	Allergic bronchopulmonary aspergillosis, definition of disease
1953	Rankin	Invasive pulmonary aspergillosis as an opportunistic infection (during neutropenia)
1955	Zimmerman	Invasive aspergillosis in infancy
1959	Finegold, Will and Murray	Classification of invasive aspergillosis
1964	Newman and Coredell	Post operative aspergillus endocarditis
1966	Milosev *et al*	Paranasal aspergillus granuloma in Sudan
1970	Young *et al*	Comprehensive description of pathology of invasive aspergillosis, with clinical correlation
1983	Katzenstein, Sale and Greenbereger	Allergic aspergillus sinusitis

Appendix 2

Materials and solutions

- GTE: 227ml 20% glucose, 494ml of 0.1M EDTA, 128.5ml 1M Tris-HCl pH 7.4 to 5l with Double De-ionised Water (DDW)
- Solution 2:
- LB Broth: 10g Tryptone, 5g Yeast Extract, 5g Sodium Chloride made to 1ltr with DDW
- TE (10:10): 10mM Tris-HCl pH8.0, 10mM EDTA
- T0.1E: 10mM Tris-HCl pH8.0, 1mM EDTA
- TBE: Tris-Borate pH8.4, 1mM EDTA pH8.0
- 1xTAE (Tris, 0.1M EDTA, Glacial Acetic Acid)
- Ficoll Loading Dye: 5mg Bromophenol Blue (BDH), 0.5g Ficoll 400 (Sigma), 0.5ml 10xTBE, 4.5ml DDW.
- SDS/NaOH (4N Sodium Hydroxide (BDH, AnalaR Grade) 25ml, 20% SDS solution (100g Sodium Dodecyl Sulphate in 500ml DDW)
- 10% SDS (50g Sodium Dodecyl Sulphate in 500ml DDW)
- Guys buffer (1M KCl, 1M Tris-HCl pH8.3, 1M $MgCl_2$, double deionised water (DDW))
- 20xSSC (1 Ltr=175.3g NaCl, 88,1g Hydrated Trisodium Citrate made to 1 ltr with DDW)
- TYE Agar plates (15g Agar, 8g NaCl, 10g Bacto Tryptone, 5g Yeast Extract made to 1 ltr with DDW)
- 40% sucrose/creosol red (40g sucrose made to 100ml with DDW, creosol red (Sigma) to a final concentration of 0.147mg/ml, filter sterilised)

- 2xTY (16g Tryptone, 10g Yeast Extract, 5g NaCl made to 1ltr with DDW)
- (T7: TAATACGACTCACTATAGGG; SP6: ATTTAGGTGACACTATAG) (Sigma Genesis 120,000pmol/ml) (Tm: 55°C)
- Precipitation mix (95% Ethanol, 0.1M EDTA, 1M Sodium Acetate)
- Solution 1: 4.504g of Glucose (50mM final conc.), 10ml of 0.5M EDTA (10mM final conc.), 12.5ml of 1M Tris pH8.0 (5mM final conc.), made up to 500ml with double distilled (dd) water and filter sterilised.
- Solution II (made fresh each time): 8.6ml of (dd) water, 400µl of 5M NaOH (2N final conc.), 1ml of 10% SDS (2N final conc.)
- Solution III: 3M KOAc pH5.5 (stored at 4°C)
- Fingerprint marker ladder : T0.1E 19.2µl, Promega ladder 1.5µl, Boehringer-Mannheim Molecular Weight V 0.1µl, 6X Buffer Dye 4.2µl.
- 6X Buffer II Dye: 0.25% bromophenol Blue, 0.25% xylene cyanol, 15% Ficoll {Type 400: Pharmacia}

Equipment

- BioRad Clamped Homogenous Electric Field gel apparatus. Gels usually run with 1.0% high gelling temperature agarose and 0.5xTBE
- New Brunswick Scientific Innova 4300 shaking incubators
- Eppendorf 5810R centrifuge
- Sorval RT-6000D centrifuge with microtitre plate rotor
- Sorval RT7 centrifuge
- Owl Scientific Gator Wide Format System A3-1 gel tank and bed

- Molecular Dynamics Phosphoimaging Scanner and NT Fragment software supplied with machine

Appendix 3

Disease	Symptoms	Prognosis	Comments	Misc. information
Allergic Bronchoplumonary Aspergillosis (ABPA)	Recurrent fever, coughing, sputum plugs infested with *Aspergillus*	Broncho-spasm and Atelectasis. Airway obstruction and alveolar collapse	Close relationship to asthma (5% pop.) Found in 10-11% cystic fibrosis patients	Due to Type I, III and IV allergic reactions
Extrinsic Allergic Alveolitis (EAA)	No previous allergic history in patient, but patient exposed to high levels of conidia. Rapid onset of fever, dyspnea and depression	Basis of "Farmers Lung". Pulmonary fibrosis may occur		Usually symptoms set in with a few hours of exposure.
Aspergilloma	Usually develops in old, tubercular or disease caused cavities. Chronic cough, weight loss and lethargy	Haemoptysis is main complication (50-80% of patients)	10% patients show spontaneous lysis in lungs. Fungal ball evident on X-ray	Massive haemoptysis can be fatal. Starts with granuloma (solid grouping of inflammatory cells
Acute Invasive Pulmonary Aspergillosis	Bronchopneumonia, fever, haemorrhagic pulmonary infarction, thromboembolism	Lack of response to treatment. Also can lead to massive Haemoptysis	Mortality of about 95%. Can be seen in various forms	Main infection of immunocompromised patients. Can go on to infect other organs
Central Nervous System Infection	Mood swings and behavioural changes, confusion and reduces consciousness. Can disseminate into cerebrospinal fluid (CSF)	50% patients with CSF infection appear healthy	Can lead to mental impairment	Damage usually irreversible and can lead to death
Chronic Necrotising Aspergillosis	Patient usually been treated for previous lung disease or already have chronic lung disease. Chronic cough, fever, weight loss and lethargy	Usually in middle –aged or older men	Can be sometime before diagnosis is made while symptoms persist	
Infection of the paranasal sinuses	Nasal discharge, fever, facial pain and headache. Facial disfigurement can occur after	Most common infection is aspergillosis. Aspergillus sinusitis is extremely	Four groups: allergic aspergillus sinusitis, fulminant invasive	Have clinical pulmonary parallels such as ABPA,

	necrotic lesions have taken hold. Brain and eye orbit also susceptible	common around world. Bone marrow transplant patients contracting sinus aspergillosis have 100% mortality	aspergillosis, saprophytic aspergillus colonisation and chronic sinusitis	Aspergilloma etc
Cerebral Aspergillosis	Cerebral infarction and thrombosis	High prevalence and highly fatal	Disseminated from other infected regions of the body	
Occular infections	Corneal ulcers, endopthalmitis and orbital aspergillosis	Causes three types of eye infection	25% cases reported, spreads to brain and causes death	Usually spreads from paranasal sinuses, described earlier
Osteomyelitis	Spinal infection and paralysis	Rare infection		
Otomycosis	Infection of the auditory canal.	Usually chronic, but can be cleared with sporadic recurrence		
Endocarditis and Myocarditis	Embolism usually common	Myocarditis is found in 15% of patients dying with disseminated aspergillosis	Usually occurs in open heart surgery patients, but also drug users	
Skin Infections	Lesions on skin, become hardened, blue and violet in colour, covered by scab	Cutaneous Aspergillosis occurs with local or general immunodeficiency e.g. scalds or burns. Has mortality rate of 50% in these types of infections	Reported as two defined types.	Can lead to disseminated aspergillosis
Infections of the Gastrointestinal tract	Intestinal ulcers, internal bleeding and perforation	Oesophagus most common site for infection	Found in 40-50% patients dying from disseminated infection	
Liver and Spleen infections	Most patients do not report symptoms, but some observation of jaundice, liver and abdominal pain has occurred	Seen in 30% of patients with disseminated aspergillosis		
AIDS patients and invasive	60-80% patients present	Most present with	Complications	AIDS is a very

aspergillus infection	pulmonary infection. Small skin lesions are also very common. Disease can be seen in all organs of body in various forms. Due to the nature of the disease, the patient can present with one, a few or all of the types of infection, causing multiple symptoms	pulmonary disease, but 10-15% have no lung involvement at all. Central Nervous System is the most common infected site outside the heart and lungs	always lead to more than one eventual presentation. Any treatment is only useful in prolonging the inevitable	complex disease that can embrace all of the above diseases in various forms.
Solid Organ Transplant susceptibility to Aspergillosis	Sub-dermal legions, skeletal damage and infection of the CNS. Fever, cough, chest pain and dyspnoea. Can cause peritonitis of the kidney in kidney transplant patients. Gut and paranasal sinuses are also targets leading to catastrophic dissemination	Infections of the pulmonary tracts dominate.	More non-specific than other infections. Symptoms tend to be observed with other symptoms from both infectious and non-infectious illnesses.	
Invasive pneumonia	Respiratory failure, fever	Does not respond to antibiotics and gives rise to classic symptoms of pneumonia	Most feared by doctors. Almost always fatal, particularly when in severe neutropenia	Disease CAN be treated, but has to be caught very early in onset, with treatment of high doses of AmB

Aspergillus is also known to be virulent pathogen in other vertebrate animals, as well as non-vertebrates. These diseases can present themselves as pulmonary and air sac infections in many birds, especially those kept in captivity and in new hatched chicks. Long nose dogs are also prone to sinusitis and pregnancy termination and horses are known to be prone to catastrophic haemorrhaging due to guttural pouch aspergillosis infection. Infections can occur in other animals such a whales and dolphins. *A. fumigatus*

is thought to be the contributor to most of these diseases, with the exception of sinusitis, which is often caused by *A. flavus*, but this does not mean to say that they do not cause the disease in any form.