Genome sequencing of the extremophile *Thermus scotoductus* SA-01 and expression of selected genes

by

KAMINI GOUNDER

Submitted in accordance with the requirements for the degree of Philosophiae Doctor

in the
Faculty of Natural Sciences
Department of Microbial, Biochemistry and Food Biotechnology
University of Free State
May 2009

SUPERVISOR : **Prof. D. Litthauer**

CO-SUPERVISORS : **Prof. E. van Heerden**

Dr. L.A. Piater

DECLARATION

I declare that this thesis hereby submitted by me for the Doctor of Philosophy degree at the University of the Free State is my own independent work and has not previously been submitted by me at another university/faculty. I further cede copyright of the thesis in favour of the University of the Free State.

_____

Kamini Gounder (2005169780)

May 2009

My humble pranams at the lotus feet of my divine Lord Sri Sathya Sai Baba
...I offer to thee.

For my Dad, who I miss so dearly.

And my Mum and brothers, who I cannot live without.

ACKNOWLEDGEMENTS

I wish to extend my appreciation and gratitude to the following individuals:

My supervisor, Prof. D. Litthauer, for guidance, constructive criticism and encouragement during the course of this study.

To Esta van Heerden and Lizelle A. Piater, for their input in this study.

Thank you to the National Research Foundation and the Metagenomics Platform for financial assistance.

To Inqaba Biotec for accommodating us while the pyrosequencing was being carried out.

Prof. Fourie Joubert (University of Pretoria) for your advice and assisting with the STADEN software.

Thank you to TIGR for providing the Annotation Engine Service and the manual annotation tool Manatee.

Thank you to Prof. G. Gottschalk, Prof R. Daniel and the students for accommodating us at the Göttingen Genomics Laboratory in Göttingen, Germany for 3 months. Lots of appreciation and thanks especially to Elzbieta Brzuszkiewicz, Sonja Voget and Heiko Leisegang for taking time out of your own work so that I could learn as much as possible from you. Many thanks for the expert mentorship and support. Thank you to Melissa Büngener for all the lab work. My time with you all was invaluable. A very special thank you as well to Antje Wolherr for performing the BiBLAST.

Prof. H-G. Patterton, for financial assistance for the Germany trip as well as use of the Bioinformatic lab.

Staff and postgraduate students at the Extreme Biochemistry Group and the Department of Microbial, Biochemical and Food Biotechnology for any assistance offered during the course of this study.

Walter Muller, for his friendship and translation of the summary.

My friends Nathlee, Landi and Godfrey for their continuous support, encouragement, friendship and the many constructive brainstorming sessions during the course of this study. Thank you!

To my parents, Nivan, Shivan for your love, unending encouragement and support. Thank You! I could not have done this without you all.

And finally, to Swami and my angels, thank you!

INDEX

Chapter 2

Whole-genome sequencing of the extremophile *Thermus scotoductus* SA-01

Chapter 3

Cloning and Expression of the DNA polymerase I (DNAPolI) and single-stranded DNA-binding (SSB) protein from *T. scotoductus* SA-01 to enhance the efficiency of PCR.

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

A	adenine
ATP	adenosine triphosphate
BCA	bicinchoninic acid
BLAST	Basic Logical Alignment Search Tool
bp	base pairs
BSA	bovine serum albumin
°C	degrees Celsius
C	cytosine
DGGE	Denaturing Gradient Gel Electrophoresis
dH_20	distilled water
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
dNTPs	deoxyribonucleoside triphosphates
E. coli	*Escherichia coli*
EDTA	ethylene diamine tetra acetate
e.g.	for example
***et al*.**	et alei (and others)
Fig.	figure
g	gram
g	gravitational force
G	guanine
Gb	gigabase
gDNA	Genomic DNA
hr	hour(s)
i.e.	that is
IPTG	Isopropyl β-D-thiogalactoside
KB	kilo bases
kDa	kilo Daltons
LB	Luria-Bertani broth
min	minute(s)
ml	millilitres

mM	millimolar
MOPS	3-(N-morpholino)propanesulfonic acid
NaCl	sodium chloride
NADH	Nicotinamide adenine dinucleotide (reduced)
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced)
NCBI	National Center for Biotechnology Information
ng	nanogram
nm	nanometer
OD	optical density
ORF	open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PFGE	Pulsed Field Gel Electrophoresis
psi	pounds per square inch
RNA	ribonucleic acid
rRNA	Ribosomal Ribonucleic acid
rpm	revolutions per minute
sec	second(s)
SDS	sodium dodecyl sulphate
sp	species (singular)
TAE	Tris, Acetic acid, EDTA
TE	Tris, EDTA
U	Units
UV	ultraviolet
µg	microgram
µl	microlitres
X-Gal	5-bromo-4-chloro-3-indolyl β-D-galactoside

ABSTRACT

Thermus scotoductus SA-01 is an extremophile that was isolated from groundwater samples at 3.2 km depth in a South African gold mine and has previously been shown to grow using nitrate, Fe(III), Mn(IV) or Sº as terminal electron acceptors and to be capable of reducing Cr(VI), U(VI), Co(III) and the quinone-containing compound anthraquinone-2,6-disulfonate. This study reports the sequencing of the *T. scotoductus* SA-01 genome using a strategy involving the GS20 and FLX pyrosequencing, which is a novel, rapid method of high-throughput sequencing, as well as Sanger technology. The GS20 and FLX pyrosequencing data was assembly into 35 contigs using the Newbler Assembly software. Mapping attempts using various software against the reference strain *T. thermophilus,* proved unsuccessful due to low levels of synteny and extensive rearrangement noticed between the two organisms.

After using various strategies to close the gaps between the 35 contigs with Sanger sequencing, the complete chromosome sequence was obtained. The genome consists of a 2 346 803 bp chromosome and a plasmid, which could not be closed with all sequencing attempts. The draft plasmid sequence consists of 8 383 bp with about 65% in agreement with the chromosome sequence. Automatic annotation of the complete chromosome and draft plasmid sequence performed by TIGR (J. Craig Venter Institute formerly known as The Institute for Genome Research) revealed the presence of 2482 and 12 ORFs, respectively. ORF correction was performed using the Artemis software package. Manual annotation was performed using the ERGO Tool software on half of the genome using various public databases and criteria. BLAST results of the plasmid sequence against the chromosome show that four ORFs present on the draft plasmid are also present in an identical copy (one or more than one copy) on the *T. scotoductus* SA-01 chromosome, providing evidence of genetic exchange between the chromosome and the extrachromosomal element.

Comparative genome analysis was done using strains that are related (3 genomes) to *T. scotoductus*, isolated from a South African goldmine (1 genome) and metal reducing organisms (2 genomes). Using this data, analysis of metabolism and thermophily of *T. scotoductus* SA-01 could be comparatively elucidated as well as determining the orthologous gene content. The complete chromosome and draft sequence of *T. scotoductus*

SA-01 not only provides valuable basic data in terms of the organism's lifestyle and capabilities but is also consists of many genes of potential interest for biotechnological applications.

Due to its thermophilic nature, *T. scotoductus* SA-01 would contain many thermostable enzymes, which possess qualities that make them more robust and better suited for use in molecular applications. The DNA polymerase I and single stranded DNA binding (SSB) protein was chosen for expression studies for their potential use in a PCR reaction. A partially purified DNA polymerase I protein was obtained; however the protein was found to be non-functional in a PCR. Expression of the SSB was performed, but the protein could not be purified for further analysis. Obtaining expression at higher levels and complete protein characterization would be required in order to understand the capabilities of these proteins.

Chapter 1

Literature Review

1. Introduction

The study of microbial evolution and ecology has been revolutionized by DNA sequencing and analysis (Tyson *et al*., 2004). The knowledge of an entire genome sequence not only provides a wealth of data, but also specific information that cannot be obtained from other approaches. Only after the completion of genome projects did it become obvious that many genes had not been identified by classical genetics (Frangeul *et al*., 1999).

In a few years we will all have access to over a thousand sequenced genomes (Overbeek *et al*., 2005). At the moment, the Genomes OnLine Database (GOLD) currently has 992 complete genomes in their database. Every newly sequenced genome will add valuable information, allowing conclusions to be made concerning new metabolic pathways, infection mechanisms, evolution of microorganisms etc. Also, comparative genomics will benefit from the increasing number of genomes that will be sequenced in the future, which will deepen our understanding of this exciting field (Schuster and Gottschalk, 2005).

Recently a new approach for high-throughput DNA sequencing has been described using pyrophosphate sequencing (Margulies *et al*., 2005). The 454 pyrosequencing technology [both the Genome Sequencer (GS) 20 and FLX generation system] has proven very successful for a number of applications such as complete microbial genome sequencing, metagenomic and microbial diversity analysis, ChiP sequencing and epigenetic studies, genome surveys, gene expression profiling and even for sample sequencing fragments of Neanderthal DNA that were extracted from ancient remains (Quinn *et al*., 2008).

In addition to its metal reduction capabilities, the thermophile *Thermus scotoductus* SA-01 is particularly interesting to study with regards to its choice of environment, the deep subsurface of the Witwatersrand Goldfields. Thus the genome structure, function and evolution of this organism can only be studied through its complete genome sequence and detailed bioinformatic analysis.

1.1 Genomics

'Genomics' is used to describe a field of science different from genetics in its focus on the study of DNA from a broader standpoint, that of the entire complement of genetic material (Venter *et al*., 2003). Originally, the term was used to describe a specific discipline in genetics that deals with mapping, sequencing and analyzing genomes. However, an increasing number of people have expanded its use to include functional analysis of entire genomes as well, including whole genome RNA transcripts (called transcriptomics), proteins (proteomics) and metabolites (metabolomics) (Xu, 2006).

The year 1995 marked the publication of two human pathogenic bacterial genomic sequences: *Heamophilus influenzae* (Fleischmann *et al*., 1995) and *Mycoplasma genitalium* (Fraser *et al*., 1995). Within 5 years of these publications, numerous other bacteria were sequenced, including *Mycobacterium tuberculosis*, one of the most important human bacterial pathogens, *Escherichia coli* and the first archaeon, *Archaeoglobus fulgidus* (Hall, 2007). The variation in microbial genome size is incredibly large, ranging from ~ 0.5 Mbp to more than 10 Mbp (Schuster and Gottschalk, 2005). Large genomes of mammals such as human and chimpanzee have led to the massive expansion of sequence data (Hall, 2007). In 2006, Poinar *et al*., sequenced 28 million base pairs of DNA in a metagenomics approach, using a woolly mammoth (*Mammuthus primigenius*) sample from Siberia. Using DNA from an exceptionally preserved sample, sequence data showed a 98.55% identity to the African elephant (*Loxodonta africana*). In addition, using high-throughput sequencing, Neanderthal genomic data has also been obtained and has been compared to human and chimpanzee genomes (Noonan *et al*., 2006 and Green *et al*., 2006).

The total number of completed bacterial genome sequences has more than doubled over the last two years and there are 855 publicly listed bacterial and archaeal genome projects that are in various stages of progress (Binnewies *et al*., 2006). Overbeek *et al* (2005) predicted that the 1000th genome would be sequenced at some point during 2007 (Fig 1.1). However to date, according to the GOLD database, 978 genomes have been completed and published. Currently there are 2497 ongoing bacterial genomes, 101 archaeal and 1029 eukaryotic genomes.

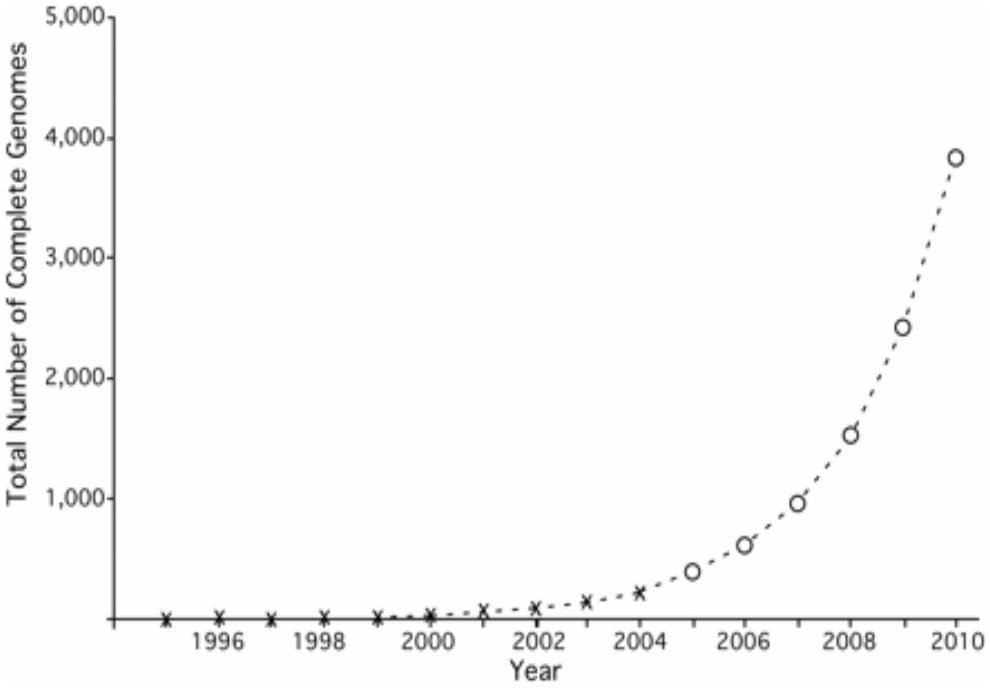


Fig 1.1 Accumulation of complete archaeal and bacterial genome sequences at NCBI 1994-2004, and prediction of the release of genomes through 2010. Data from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi was extracted and plotted by year as shown with the crosses. Data from 2004-2010 is projected by the power law and is represented by open circles. At this current rate of growth, the 1000th complete genome should have been released by late 2007 or early 2008.

Completion of genome projects could not have been accomplished without major innovations in recombinant protein engineering, fluorescent dye development, capillary electrophoresis, automation, robotics, informatics and process management (Metzker, 2005). Comparable breakthroughs have also been achieved in closure strategies in centres such as The Institute for Genome Research (TIGR) and the Pathogen Sequencing Unit at the Sanger Centre, which routinely produce complete microbial genome sequencing data, and closure and annotation can usually be accomplished in a matter of a few months (Fraser *et al*., 2002). The most significant technical advance in genomics has been the development of efficient, high-throughput DNA sequencing techniques and instruments. While the basic principle for DNA sequencing was established in the mid-1970s, it was not until the mid 1990s when efficient automated DNA sequencers with fluorescent dyes to tag dideoxyribonucleotides (with one colour for each of the four types of nucleotides) were developed (Xu, 2006). In addition, there are several commercial next-

generation sequencing technologies that have become available in recent years (Shendure *et al*., 2004).

1.2 DNA Sequencing Technologies

Advances in genome sequencing technologies have similarly had great impact on microbial biology, providing new insights into microbial evolution, biochemistry, physiology and diversity (DeLong, 2005). In addition, the need for sequencing has never been greater than it is today, with applications spanning diverse research sectors including comparative genomics and evolution, forensics, epidemiology and applied medicine for diagnostics and therapeutics (Metzker, 2005).

Due to the overwhelming number of ongoing genome projects there is a growing demand for even greater speeds and lower costs in the development of new sequencing technologies, which are starting to make way into the marketplace (Bonetta, 2006). Large-scale sequencing projects, including whole genome sequencing, have usually required the cloning of DNA fragments into bacterial vectors, amplification and purification of individual templates, followed by Sanger sequencing using fluorescent chain-terminating nucleotide analogues and either slab gel or capillary electrophoresis (Margulies *et al*., 2005). Though the majority of DNA sequencing techniques are gel-based and electrophoretic, there are high-throughput techniques that are more suitable for other applications than long sequence reads (Gharizadeh *et al*., 2006). Thus there is a need for a more efficient and cost-effective approach for genome sequencing that can maintain the high quality of data produced by conventional Sanger sequencing (Goldberg *et al*., 2006).

1.2.1 Older sequence techniques

1.2.1.1 Sanger sequencing

The existing Sanger sequencing methods have served as the cornerstone for genome sequencing, including microbial sequencing, for over a decade (Goldberg *et al*., 2006). This method of DNA sequencing and subsequent developments in automation and computation revolutionized the world of biological sciences and eventually led to the sequencing of the consensus human genome (Braslavsky *et al*., 2003).

Conventional DNA sequencing relies on the elegant principal of the dideoxynucleotide, chain-termination technique first described more than two decades ago. This multi-step principle has gone through major improvements during the years to make it a robust technique that has been used for the sequencing of several different bacterial, archeal and eukaryotic genomes (Ronaghi, 2001).

The Sanger sequencing method is based on the incorporation of 2′, 3′-dideoxynucleotide triphosphates (ddNTPs) – similar to the dideoxynucleotides (dNTPs), but with a chain-terminating hydrogen atom instead of a hydroxyl group attached to the 3′ carbon – to a growing DNA chain. In a sequencing reaction a single-stranded DNA fragment is combined with the appropriate sequencing primer; a ddNTP (for example, ddTTP); and the normal dNTPs (dTTP, dCTP, dATP and dGTP), one of which is labelled. When DNA polymerase is added to the mix, it begins to synthesize the corresponding DNA strand. DNA synthesis will stop every time the ddTTP is added, resulting in many labelled DNA fragments of varying lengths but always with a T residue at the end. In this older method the reaction is carried out four times using a different ddNTP in each reaction. After gel electrophoresis and autoradiography, the arrangement of the nucleotides in the DNA can be determined by placing the fragments in the four lanes in order (Bonetta, 2006).

Improvements were made in the 1990s with the use of different coloured fluorescent dyes to label terminators so that all of the terminators can be incorporated in a single reaction. The first sequencing machines used this technology in combination with devices to automatically read fragments as they were separated on a polyacrylamide gel. Later, capillaries replaced the gels, which simplified the separation step and increased the read lengths. Within a period of 10 years, the average read length of a sequencing read has increased from around 450 bp to 850 bp (Hall, 2007). Although sample preparation and sequencing reactions are still mostly done by hand, automated sequencers these days take care of loading and running the gels and reading the results. The market leader is Applied Biosystems (ABI)'s flagship 3730xl sequencer (Fig 1.2). The machine contains a capillary array – with each capillary not wider than a human hair and equivalent to one slab gel lane – that can run 96 sequencing reactions, each generating some 800 bases, in parallel (Bonetta, 2006). The instrument now has an increased throughput of more than 1.6 million bp/day (Chan, 2005).

Fig 1.2 The high-throughput 3730 & 3730*xl* DNA Analyzers were developed to meet the growing needs of institutions ranging from core and research labs in academia, government, and medicine to biotechnology, pharmaceuticals and genome centers (Applied Biosystems).

1.2.1.1 Maxam and Gilbert Sequencing

This method was presented in 1977 and is based on sequencing by chemical cleavage. In this technique, the DNA fragments are generated either by digestion of the sequencing template by restriction enzymes or PCR amplification with the ends of the fragments labelled, traditionally by radioactivity. Single-stranded DNA fragments radioactively labelled at one end are isolated and subjected to chemical cleavage of base positions. Four parallel cleavage reactions are performed, each one resulting in cleavage after one specific base. The sequence is deduced from the sequence gel separation pattern like in the Sanger DNA sequencing method. A read length of up to 500 bp has been achieved with this method. However, the chemical reactions in this technique are slow and involve hazardous chemicals that require special handling in the DNA cleavage reactions (Ahmadian *et al*., 2006).

1.2.2 New Sequencing Techniques

1.2.2.1 Sequencing by Hybridization (SBH)

This method utilizes a large number of short, nested oligonucleotides immobilized on a solid support to which the labelled sequencing template is hybridised (Ahmadian *et al*., 2006). One approach is to immobilize the DNA that is to be sequenced on a membrane or glass chip and then to carry out serial hybridisations with short probe oligonucleotides (for example, 7 bp oligonucleotides). The extent to which specific probes bind the target DNA can be used to infer the unknown sequence (Shendure *et al*., 2004). The target sequence is deduced by computer analysis of the hybridisation pattern of the sample DNA. DNA sequences can also be analysed by sequence by synthesis. This method is mainly suitable for detection of genetic variations within known DNA sequences and re-sequencing. SBH may also be employed for certain applications such as genotyping samples with well-characterised genetic variations such as single nucleotide polymorphisms (SNPs) (Ahmadian *et al*., 2006).

For each base pair of a reference genome to be resequenced, there are four features on the chip. The middle base pair of these four features is either an A, C, G or T. The sequence that surrounds the variable middle base is identical for all four features and matches the reference sequence. By hybridising labelled sample DNA to the chip and determining which of the four features yields the strongest signal for each base pair in the reference sequence, a DNA sample can be rapidly resequenced. This technique can be used to obtain an impressive amount of sequence, i.e. > 10^9 bases. The primary challenges that SBH faces are to design probes or strategies that avoid cross-hybridisation of probes to the incorrect targets as a result of repetitive elements or chance similarities. Also, SBH still requires sample preparation steps, as the relevant fraction of the genome must be amplified by PCR before hybridisation (Shendure *et al*., 2004).

1.2.2.2 Pyrosequencing

This most current sequencing technology is a modification of the classical Sanger method called pyrosequencing (Edwards *et al*., 2006) that reads the DNA sequence as the DNA strand is synthesized (Fig 1.3) (Bonetta, 2006).

Fig 1.3 The Genome Sequencer and FLX Instrument features a groundbreaking combination of long reads, exceptional accuracy and high throughput (Roche Applied Sciences, 454 Life Sciences).

In a cascade of enzymatic reactions, visible light is generated that is proportional to the number of incorporated nucleotides. The cascade starts with a nucleic acid polymerisation reaction in which inorganic pyrophosphate (PP_i) is released as a result of nucleotide incorporation by polymerase. The released PP_i is subsequently used to synthesise ATP by ATP sulfurylase, which provides the energy to luciferase to oxidize luciferin and generate light. Because the added nucleotide is known, the sequence of the template can be determined.

Three different versions of pyrosequencing have been reported thus far. However, the four-enzyme system of pyrosequencing has been the most popular version (Langaee and Ronaghi, 2005). The 4 enzymes included in the pyrosequencing system are the Klenow fragment of DNA Polymerase I, ATP sulfurylase, luciferase and apyrase (Ahmadian *et al*., 2006). The Klenow fragment of *E. coli* DNA Pol I is a relatively slow polymerase. The ATP sulfurylase is a recombinant version from the yeast *Saccharomyces cerevisiae* and the luciferase is from the American firefly *Photinus pyralis*. The overall reaction from polymerisation to light detection takes place within 3-4 sec at room temperature. One pmol of DNA in a pyrosequencing reaction yields 6 X 10^{11} ATP molecules, which in turn, generates more than 6 X 10^{9} photons at a wavelength of 560 nanometers. This amount of light is easily detected

by a photodiode, photomultiplier tube or a charge-coupled device (CCD) camera (Ronaghi, 2001).

Steps in the Pyrosequencing reaction:

1. The DNA polymerisation occurs if the added nucleotide forms a base pair with the sequencing template and thereby is incorporated into the growing strand.
2. The inorganic pyrophosphate, PP_i, released by the Klenow DNA polymerase serves as substrate for ATP sulfurylase, which produces ATP.
3. The ATP is converted to light by luciferase and the light signal is detected. Hence, only if the correct nucleotide is added to the reaction mixture, light is produced.
4. Apyrase removes unincorporated nucleotides and ATP between the additions of different bases (Fig 1.4) (Ahmadian *et al*., 2006).

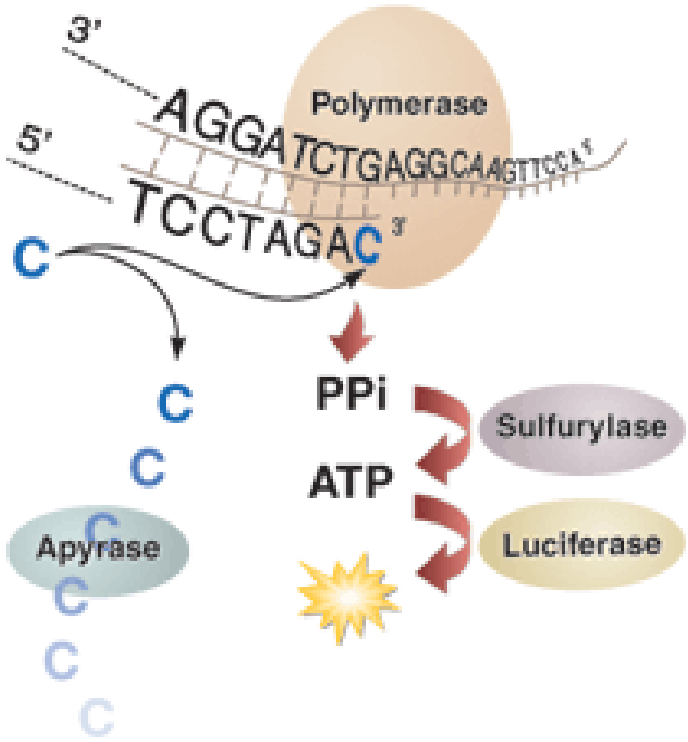


Fig 1.4 Schematic representation of the pyrosequencing enzyme system. Of the added dNTP forms a base pair with the template, Klenow Polymerase incorporates it into the growing DNA strand and pyrophosphate (PP_i) is released. ATP sulfurylase converts the PP_i into ATP, which serves as a substrate for the light producing enzyme Luciferase. The light produced is detected as evidence of that nucleotide incorporation has taken place (Ahmadian *et al*., 2006).

Pyrosequencing also eliminates the need for cloning, thus removing the potential for both aberrant recombinants in the surrogate host and for cloning-related artefacts such as counter selection against potentially toxic genes such as those found on phages. For environmental microbiology there are two main approaches that are currently using pyrosequencing. The first is whole genome random sequencing. In this approach community genomic DNA is extracted and sequenced as is. The second is to sequence 16S rDNA libraries to extinction. In this approach, 16S rDNA genes are amplified by PCR, but instead of cloning, the genes are sequenced with pyrosequencing (Edwards *et al*., 2006).

Pyrosequencing has the potential advantages of accuracy, flexibility, parallel processing and can be easily automated. Furthermore, it dispenses the need for labelled primers, labelled nucleotides and gel electrophoresis. The method is broadly

applicable for analysis of short DNA sequences used in bacterial, fungal and viral typing, scanning for undefined mutations, bacterial genotyping and tag sequencing (Ahmadian *et al*., 2006; Gharizadeh *et al*., 2006). The methodological performance of pyrosequencing in determination of difficult secondary DNA structures, mutation detection, cDNA analysis, resequencing of disease-associated genes, microbial typing and single nucleotide polymorphism (SNP) analysis has been shown (Langaee and Ronaghi, 2005). In addition to the raw-sequencing cost factor, the different methods developed for pyrosequencing have eliminated the need for PCR-amplification, library construction, cloning, colony picking and arraying. A new pyrosequencing technology is the 454 GS20 or GS FLX Sequencer (454 Life Sciences). Recently, the GS FLX Titanium series was introduced producing individual sequencing reads with an improved Q20 length of 400 base pairs (99 per cent accuracy at the 400th base and higher for preceding bases) and a five-fold increase in throughput to 400-600 million base pairs per instrument run. It is a highly parallel non-cloning pyrosequencing based system capable of sequencing 100 times faster than current state-of-the-art Sanger sequencing and capillary electrophoresis platforms. The major concerns have been relatively short read lengths (i.e. as of 2007 an average of 100-200 nt compared to 800-1 000 nt for Sanger sequencing), a lack of a paired end protocol and the accuracy of individual reads for repetitive DNA, particularly in the case of monopolymer repeats. Combined, these factors often make it impossible to span repetitive regions, which therefore collapse into single consensus contigs during sequence assemblies and leave unresolved sequence gaps. These issues have recently been addressed with the release of the GS FLX system as well as the Long Paired End sequencing platform. The GS FLX system provides longer read lengths and lower per-base error rates than the previous system. This system currently offers the longest read length of any of the next generation sequencing systems currently available (Quinn *et al*., 2008).

The main concerns for this technique are the short length of sequence fragments and the requirement to use whole genome amplification to generate sufficient DNA for sequencing from environmental libraries (Edwards *et al.*, 2006). Single-stranded DNA binding protein (SSB) is highly recommended for primer and template complications in pyrosequencing. However, SSB has shown limitations in resolving strong secondary structures or primer related self/cross-hybridisations in challenging regions (Gharizadeh *et al*., 2006). The principle problem with this approach is the short sequence fragments that are generated. This, of course, limits the ability of

most bioinformatics analyses that are currently used such as gene finding, protein similarity searches and sequence assembly (Edwards *et al.*, 2006).

According to Margulies *et al* (2005), a high-density pyrosequencing is 99.96% accurate when compared with DNA sequenced by conventional sequencing methods and capillary electrophoresis. A study done by Huse *et al* (2007) also showed that by using objective criteria to eliminate low quality data, the quality of individual GS20 sequence reads in molecular ecological applications can surpass the accuracy of traditional capillary methods. Gharizadeh *et al* (2006), compared pyrosequences with Sanger dideoxy methods for 4 747 templates. Comparisons of the traditional capillary sequences with the 25-30 nucleotide pyrosequence reads demonstrated similar levels of read accuracy. Smith *et al* (2007), performed large numbers of parallel sequencing runs of *Acinetobacter baumanii* with a genome sequence coverage of 21.1 times. The authors found that when combined with conventional gap filling, the accuracy of the sequence and assembly are comparable to the whole genome shotgun sequencing methods that have become the gold standard of bacterial genomic sequencing. Another particular study was done to determine the optimal combination of 454 and Sanger sequencing data that would produce the best possible high quality genome assembly in the most timely and cost effective manner for marine microbial genomes. The results showed that 8 X Sanger sequencing to be the most cost effective approach and for organisms with a large genome size, many sequencing gaps and/or hard stops, results showed initial sequencing of 5.3 X Sanger data followed by the addition of two 454 runs to be the most cost-effective approach. By increasing the amount of 454 sequencing data at any ratio to Sanger sequencing, results showed an improvement to the final draft genome in terms of coverage, reduction of gaps and reduction of poorly sequenced regions that degrade the value of an assembly (Goldberg *et al.*, 2006). Jeong and Kim (2008), determined that 454 pyrosequencing at a 20 X sequencing coverage is usually enough to produce a high quality draft. For a conventional microbial genome project that employs paired-end Sanger sequencing on genomic libraries, end sequences from a fosmid library that has a 10 X clone coverage is sufficient for generating scaffolds. The authors also suggest that this would be an appropriate choice when both 454 pyrosequencing and fosmid end sequencing with Sanger chemistry are utilized. However, Aury *et al*., (2008), compared the assemblies obtained using Sanger data with those from different inputs from the latest new sequencing technologies (454 GSFLX and Solexa/Illumina). The authors concluded from the study that a combination of the two new sequencing technologies allows production of a high-

quality draft of at least a comparable quality to those obtained with Sanger data alone.

With respect to *de novo* assembly of a complex genome, the most relevant test to date of the capability of the 454 pyrosequencing technology (GS20 system) involved sequencing four Bacterial Artificial Chromosome (BAC)s containing inserts of the barley genome, two of which had previously been sequenced using the traditional Sanger approach (Quinn *et al*., 2008). It was found that all gene-containing regions were covered efficiently and at high quality with 454 sequencing whereas repetitive sequences were more problematic with 454 sequencing than with Sanger sequencing. 454 sequencing provided a much more even coverage of the BAC clones than Sanger sequencing, resulting in almost complete assembly of all genic sequences even at only 9 to10-fold coverage (Wicker *et al*., 2006). Given the significant and ongoing improvements in the 454 technology since the barley BAC analysis, Quinn *et al* (2008), presented the results of the first use of the GS FLX paired-end reads for *de novo* sequence assembly of a 1 Mb region of Atlantic salmon DNA covered by a minimum tiling path comprising of 8 BACs. The data demonstrated that this improved the GS FLX assemblies, however, with respect to *de novo* sequencing of complex genomes, the GS FLX technology is limited to gene mining and establishing a set of ordered sequence contigs. The results from the study also showed that for a salmon reference sequence, it appears that a substantial portion of sequencing should be done using Sanger technology.

The first metagenomic analysis performed using pyrosequencing was done on environmental samples from the Soudan Mine. The authors concluded that by combining pyrosequencing, subsystems analysis and comparative metagenomics the microbiology of different environments could be correlated with the chemistry and hydrogeology of those environments to identify significant ecological differences between them (Edwards *et al*., 2006).

1.2.2.3 Cyclic array sequencing on single molecules

Previous methods are based on *in vitro* or *in situ* amplification step, so that the DNA to be sequenced is present at sufficient copy numbers to achieve the required signal. A method for directly sequencing single molecules of DNA would eliminate the need for costly and often problematic procedures, such as cloning and PCR amplification. Several groups are developing cyclic-array methods that are related to those

methods discussed above. Each method relies on the extension of a primed DNA template by a polymerase with fluorescently labelled nucleotides, but they differ in the specifics of their biochemistry and signal detection. An advantage of this method is that they might require less starting material than other ultra low cost sequencing contenders and conventional sequencing. This feature is relevant to all technologies and methods for amplifying large DNA molecules by multiple displacement amplification or whole genome amplification are improving rapidly. This will enhance our ability to obtain a complete sequence from single cells even when they are dead or difficult to grow in culture (Shendure *et al.*, 2004).

1.2.2.4 Nanopore sequencing

This method is a creative single-molecule approach unlike others. As DNA passes through a 1.5 nm nanopore, different base pairs obstruct the pore to varying degrees, resulting in fluctuations in the electrical conductance of the pore. The pore conductance can be measured and used to infer the DNA sequence (Fig 1.5). The accuracy of base calling range from 60% for single events to 99.9% for 15 events. However, the method has so far been limited to the terminal base pairs of a specific type of hairpin. This method has a great deal of long-term potential for extraordinary rapid sequencing with little to no sample preparation. However, it is probable that significant pore engineering will be necessary to achieve single-base resolution. Rather than engineering a pore to probe single nucleotides, Visigen and Li-cor are attempting to engineer DNA polymerases or fluorescent nucleotides to provide real-time, base specific signals while synthesising DNA at its natural place (in other words, a non-cyclical sequencing-by-extension system) (Shendure *et al.*, 2004). This approach is conceptually appealing as it does not require fluorescent labelling and is fast. However, there are some daunting challenges. To practically implement this approach, solid-state nanopores need to be fabricated; in this manner, denaturing conditions can be used and measurements can be more robust. Solid-state pores have yet to demonstrate discrimination of different nucleotides in DNA. Therefore, nanopore sequencing hurdles need to be addressed before it can routinely sequence DNA. Accomplishments in the nanopore sequencing field include rapid discrimination between pyrimidine and purine segments. Applications of this technique include detection of single nucleotide polymorphisms with oligonucleotides immobilised in the nanopore and analysis of DNA heterogeneity and phosphorylation. Currently, the approach calls for the use of single-stranded DNA for sequencing. The longest single-stranded DNA molecules that have been measured are approximately

100 bp. Double-stranded DNA, however, have fared better in solid-state nanopores; DNA lengths up to 48.5 kb have been demonstrated to pass through solid-state nanopores. Furthermore, a sequencing strategy for double-stranded DNA has yet to be articulated for nanopore sequencing (Chan, 2005).

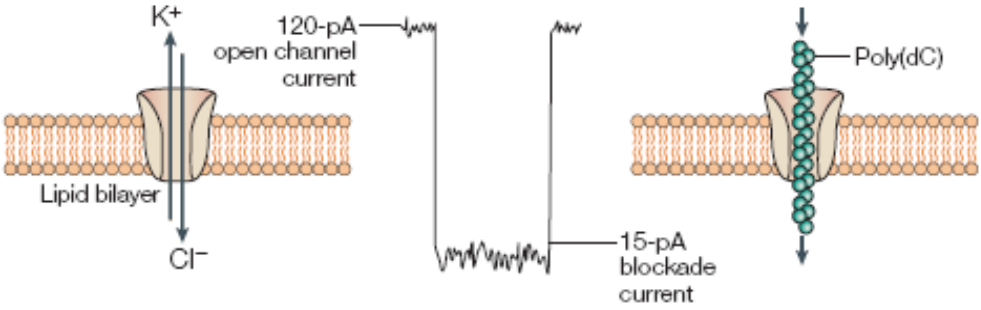


Fig 1.5 Nanopore sequencing, left, single-stranded polynucleotides can only pass single-file through a hemolysin nanopore. Right, the presence of the polynuceotide in the nanopore is detected as a transient blockade of the baseline ionic current, pA, pico-Ampere (Shendure *et al.*, 2004).

1.2.2.5 Solexa Sequencing

A massively parallel sequencing by synthesis from amplified fragments has recently been developed by a company called Solexa. This technology differs from 454 sequencing as it amplifies the DNA on a solid surface followed by synthesis by incorporation of modified nucleotides linked to coloured dyes. The company has since released their first instrument that is capable of sequencing over 1 Gb in a single run and is likely to have a major impact on the genomics field (Hall, 2007). Read lengths are 30-50 bases, which are of sufficient length for re-sequencing applications (Bentley, 2006). It should be noted that this platform has recently shown dramatic and rapid increases in total yield, sequence quality and read length such that the sequencer is capable of yielding over 100 million high-quality short reads (up to 76 bases) per three to five day run totalling several gigabases of aligned sequence (Lister *et al*., 2008).

Another, new technology released from Helicos (BioSciences Corporation) is the HeliScope™ Single Molecule Sequencer. The sequencer images billions of single

molecules per run and produces over one gigabase of usable sequence data per day – a throughput performance almost 100 X greater than Sanger methods, and faster than any of the "next-generation" methodologies.

1.3 Bioinformatic Analysis

Of all the methods mentioned, none would be successful in microbial research without bioinformatic tools. Broadly defined, bioinformatics refers to the use of computers to seek patterns in the observed biological data and to propose mechanisms for such patterns (Xu, 2006). The choice of appropriate bioinformatic packages should be made at the beginning of the project, since changing to another package generally leads to a vast amount of additional work (Franguel *et al*., 1999).

1.3.1 Assembly Phase

One of the most complex and computationally intensive tasks of genome sequence analysis is genome assembly (Pop *et al*., 2004). The new DNA sequencing techniques demand new assembly software to stitch together short strings of nucleotide bases, as determined by a sequencer, called reads (Miller *et al*., 2008).

The assembly phase is composed of three major steps: the conversion of the data from automated sequencers to nucleotide sequences, the utilisation of these sequences in the assembly process and the continuous assessment of this process (Frangeul *et al*., 1999). Some of the major assemblers used today are for example : PCAP (parallel contig assembly program), capable of assembling tens of millions of reads into long sequences (Huang *et al*., 2003); Atlas (Havlak *et al*., 2004); Arachne (Jaffe *et al*., 2003) and Celera Assembler, which has been modified for combinations of ABI 3730 and 454 FLX reads (Miller *et al*., 2008). One of the first assemblers introduced by Staden in 1980 was a computer program developed to store and manipulate DNA gel reading data obtained from the shotgun method of DNA sequencing (Staden, 1980).

Essentially, the basic principle steps in assembly consists of the following:

- Sequence and quality data are read and the reads are cleaned.
- Overlaps are detected between reads. False overlaps, duplicate reads, chimeric reads and reads with self-matches (including repetitive sequences) are also identified and left out for further treatment.

- The reads are grouped to form a contig layout of the finished sequence.
- A multiple sequence alignment of the reads is performed and a consensus sequence is constructed for each contig layout (often along with a computed quality value for each base).
- Possible sites of misassembly are identified by combining manual inspection with quality value validation (Scheibye-Alsing *et al*., 2009).

The pyrosequencing platform produced by 454 Life Sciences is sold with Newbler, an assembler specifically for 454's medium-length reads (Margulies *et al*., 2005). One may mix traditional Sanger-type sequences, usually generated from fosmid libraries, for scaffolding purposes, together with 454 pyrosequencing data to produce more accurate data. Among the SFF tools that Roche Applied Science provides for the handling of raw data files, SFFINFO can generate FASTA and quality score files from an SFF file. Although the converted files can be assembled using PHRAP, it does not ensure correct assembly because the quality scores that are generated from 454 data are not compatible with those from Sanger reads. Further, PHRAP has problems with handling massive reads, which is usually hundreds and thousands from an SFF file (Jeong and Kim, 2008).

New assemblers such as Velvet offer functionality specifically for short-read sequencing technologies such as Solexa (Miller *et al*., 2008). Although some simple bacterial genome assemblies have been carried out on reads of less than 50 bp, for the vast majority of genomes, assembly would be impossible. The ability to generate read pairs is also vital for assembly of large genomes as it allows distant regions of the genome to be linked (Hall, 2007). According to Salzberg and Yorke (2005), there are hundreds, and sometimes thousands of mis-assemblies. These include regions where a genome is incorrectly rearranged as well as places where large chunks of DNA are simply deleted and the surrounded sequences just crunched together. The source of most mis-assemblies is, as it has always been, repeats. Genomes vary in their repeat content; however, large genomes are filled with repeats of all shapes and sizes. However, there are several software tools such as CONSED, which are dedicated to editing the assembly results (Frangeul *et al*., 1999).

According to Jeong and Kim (2008), a recent report demonstrated that the GS-assembler program (gsAssembler for *de novo* assembly and gsMapping for reference guided assembly, supplied by Roche Applied Science) is ideal for correct

assembly of 454 data that are short and inherently error-rich. Also, the recent versions of GS assembler programs support mixed assembly with Sanger-type reads, however, its performance is not well known at present.

In 1995, Bonfield *et al.*, described a program called Genome Assembly Program (GAP), which can be used for DNA sequence assembly and is suitable for large and small projects, a variety of strategies and can handle data from a range of sequencing instruments.

1.3.2 Closure phase

Generally, a thorough analysis done on the sequences obtained by the assembly software allows for an effective choice of the moment to begin the closure phase (Frangeul *et al*., 1999). A complete genome sequence represents a finished product in which the order and accuracy of every base pair have been verified. In contrast, a draft sequence, even one of high coverage, represents a collection of contigs of various sizes, with unknown order and orientation that contain sequencing errors and possible mis-assemblies (Fraser *et al*., 2002).

In the late stages of a whole genome shotgun (WGS) sequencing project, most DNA sequences can be assembled into large contiguous blocks or contigs. As the project nears completion, the number of contigs grows smaller as the size of contigs grows larger (Tettelin *et al*., 1999). Linkage information for contigs can be derived from the genomic sequences of related organisms. As new genome sequences are released on a weekly basis, the chance increases for matching of an unfinished genome with a related genome (van Hijum *et al*., 2005). Due to randomness in the library and unclonable sequences, some regions of the genome are not represented in the contigs, resulting in gaps. Such gaps are called sequence gaps and they can be 'walked' by synthetic primers using the shotgun clone as a template. Some of the physical ends of contigs can be extended by primer walking directly on genomic DNA (Tettelin *et al*., 1999). The resulting contigs that are still unlinked can be extended using methods described in Fig 1.6 below:

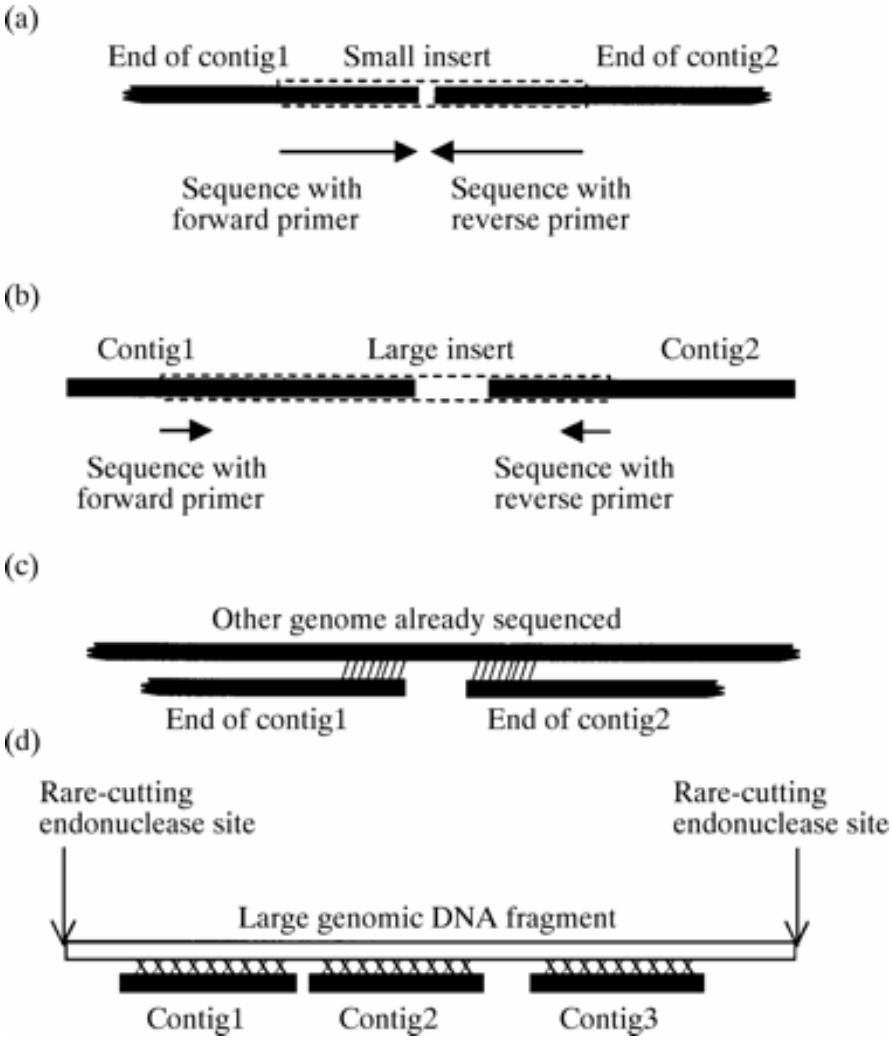


Fig 1.6 Methods for the construction of supercontigs. (a) Contigs sharing sequences with a linking small-insert clone. (b) Contigs sharing the end sequences of a linking clone from a large-insert library. (c) Contigs sharing the same operon (or gene) in another entirely sequenced genome. (d) Contigs identified by hybridization to be located on the same large genomic fragment. The symbols used are: cloned insert of the linking clone (rectangle with dotted lines); sequences performed on these clones (arrows); known sequences (black boxes); unknown sequences (white boxes); similarity detected by hybridization (xxxxxxx); similarity detected by BLAST (///////) (Franguel *et al*., 1999).

Most large-scale genome sequencing projects today employ the whole-genome shotgun sequencing strategy, in which a genome is shattered into numerous small fragments and the fragments are then sequenced from both ends. The resulting sequences, ranging from 650 to 850 bp in length using the latest sequencing technology, must then be assembled to reconstruct the chromosomes of the target organism (Pop *et al*., 2004). During the closure phase, both ends of the small and

large insert clones are sequenced. Thus, if the terminal sequences of a single clone belong to different contigs, it is highly probable that these two contigs are neighbours (Fig 1.6a and b). However, the orientation of the sequences and the distance to the end of the contig should be compatible with the size range of the inserts. These clones are called 'linkage clones' and the resulting contigs are 'supercontigs'. The probability that one clone can be used as a linking clone increases with its size. Therefore large-insert libraries are more useful at this stage in the assembly process (Frangeul *et al*., 1999).

However, some regions that cause gaps are unclonable or unstable in bacterial cells and hence are under-represented in libraries. Sequencing using DNA polymerase-based extension products can be hindered by motifs that form secondary structures or other structural forms. These motifs are often GC-rich sequences with high thermal and structural stability, presumably because the high duplex melting temperature permits stable secondary structures to form, thus preventing completion of a sequencing reaction or causing band compression in completed reactions (Keith *et al*., 2004). Other gaps result from extremely GC-rich or GC-poor regions and large repeat sequences. Significant effort is needed to close these gaps to finish the project (Tettelin *et al*., 1999). Some examples of methods to reduce the stability of duplex DNA are the inclusion of denaturing chemicals, sulfones, or dimethylsulfoxide (DMSO); shearing of DNA into smaller pieces to disrupt the motif, and the introduction of non-mutagenic, strand-destabilizing nucleotide analogues into the sequencing reactions (Kieth *et al*., 2004).

Draft sequence data can also be useful in comparative studies when a complete genome sequence for a closely related strain or species exists and can be used as a scaffold to order and orient contigs (Fraser *et al*., 2002). In addition, the gene order in certain operons may also be conserved between the genome being sequenced and the sequenced genome of related organisms, called a control genome (Fig 1.6c). In this way, a control genome may be of help in the completion of the genome under study (Frangeul *et al*., 1999). Unfortunately, physical ends (and gaps) are frequently the result of repetitive sequences that cannot be resolved by sequence assembly algorithms. Generally, the repeat regions are usually longer than the average sequence read, and walking using a primer located outside the repeat will not get across the repeat and therefore will not extend the physical end into the gap. Generating PCR products across each gap using unique primers located outside the repeat region can circumvent this problem. These PCR products can then be

subsequently walked using the product itself as a template, where the repeats do not cause a problem because they are unique within the PCR product (except in the case of a long tandem repeats). In addition, PCR products do not need to be cloned prior to sequencing and therefore regions potentially toxic to the host (another cause of gaps in a shotgun sequencing project) will be sequenced (Tettelin *et al*., 1999). Another method would be using multiplex PCR, utilizing multiple PCR primers in a single sequencing reaction. Thus, primer design takes into account that in many cases contig ends contain unreliable DNA sequences and repetitive sequences (van Hijum *et al*., 2005).

If computational approaches (e.g. Mapping) fail to predict links between the remaining contigs, experiments may be required. Southern-hybridization experiments, using probes corresponding to different contigs, may show that some of these contigs are located on the same large restriction fragment (Fig 1.6d). Such fragments can be obtained after digestion of chromosomal DNA with a rare-cutting endonuclease and separated by pulse-field gel electrophoresis (PFGE) (Frangeul *et al*., 1999). Another technique for rapid closure of genomes is optical mapping. This method permits the assembly of whole-genome restriction endonuclease maps by digesting immobilized DNA molecules and determining size and order of fragments (Latreille *et al*., 2007).

In the closure phase, the assembled sequence emanating from the shotgun approach is analysed and refined, with additional sequence data typically generated to attain long-range continuity and to improve accuracy. Sequence finishing is a low-throughput, craftsman-like process that involves highly skilled personnel performing both computational and experimental procedures in a customized fashion and, as a result, it is also relatively expensive (Blakesley *et al*., 2004).

1.3.3 Genome Annotation

The process of genome annotation can be defined as assigning meaning to sequence data that would otherwise be almost devoid of information. By identifying regions of interest and defining putative functions for those areas, the genome can be understood and further research initiated (Meyer *et al*., 2003). However, genome annotation is only made more complete when individual genes are placed in context of metabolic pathways, coordinated cellular activities of cellular structures (Haft *et al*., 2005). A number of genome annotation systems intended for the analysis of prokaryotic and eukaryotic organisms have been designed and presented in the last few years. The first generation was published in 1996 and consisted of the MAGPIE, GeneQuiz and Pendant systems. Since then, a second generation of mostly commercial genome annotation systems has been published including ERGO (Integrated Genomics, Inc.), Pendant-Pro (successor to Pedant), Phylosopher (Gene Data, Inc), BioScout (Genequiz, Lion AG), WIT and the open source Artemis. Some systems (MAGPIE, Artemis, and Phylosopher) contain extensive visualizations or include multiple genome comparison-based annotation strategies (most notably by ERGO). With the exception of Artemis, all systems provide an automatic annotation feature. Lately, the Manatee system has been made public by TIGR (Meyer *et al*., 2003).

In response to the challenge of the rapid release of genome data and the need for high-throughput annotation systems, the Fellowship for Interpretation of Genomes (FIG) launched the 'Project to Annotate a 1000 Genomes'. The project is built around the principle that the key to improved accuracy in high-throughput annotation technology is to have experts annotate single subsystems (a set of functional roles that together implement a specific biological process or structural complex) over the complete collection of genomes rather than having an annotation expert attempt to annotate all of the genes in a single genome (Overbeek *et al*., 2005). Meyer *et al.,* (2003), also presented an open source genome annotation system for prokaryotic genomes called GenDB, which supports manual as well as automatic annotation strategies. However, annotation generally is thought to be of best quality when performed by a human expert. Due to this, software assistance for computation, storage, retrieval and analysis of relevant data has become essential for the success of any genome project (Venter *et al*., 2003).

1.4 Whole-Genome Comparison

Genome sequence comparison has been an important method for understanding gene function and genome evolution since the early days of gene sequencing. The pairwise sequence comparison methods implemented in BLAST and FASTA have proved invaluable in discovering the evolutionary relationships and functions of thousands of proteins from hundreds of different species (Kurtz *et al*., 2004). When a closely related, fully sequenced genome is available, comparative assembly can be easily performed by extracting the homologous sequence and assembling it with either a comparative assembler or an alignment program that can handle draft sequence (Chen and Pachter, 2005). In 1999, Delcher *et al.*, described a system for pairwise alignment and comparison of very large scale DNA sequences. The algorithm assumes the sequences are closely related and using this assumption can quickly compare sequences that are millions of nucleotides in length. It is also able to compare entire chromosomes as large as human chromosome (i.e. several hundred million base pairs), and in the process identify all differences between two different individuals. This alignment system, called MUMmer, is capable of aligning complete bacterial genomes in <1 min on a standard desktop computer. This system has now been redesigned to require far less memory and in the process run faster, as well as align either protein or DNA sequences (Delcher *et al*., 2002). Until 1999, each new genome published was so distant from all previous genomes that aligning them would not yield interesting results. However, related to the growing number of closely related species that have been sequenced is a rapid growth in the number of known species whose genomes are similar but have undergone significant rearrangement (Kurtz *et al*., 2004). Alignments of related bacterial species led to the discovery that chromosome-scale inversions are a common evolutionary phenomenon in these species, and that the inversions are nearly always symmetric about the origin of replication. These inversions show up as X-shaped alignments in the dot plot of all the DNA sequences conserved between the two species. Many comparative genome analyses have been carried out using the MUMmer software package with many observing the X-alignment (Delcher *et al*., 2002). Another such example of comparison was using *C. kluyveri* with other clostridial genomes. Only a few regions of synteny on protein level were found, in particular with the genomes of *Clostridium acetobuylicum* and *Clostridium tetani*. This particular example of low conservation of genome organization underlines the heterogeneity of the genus *Clostridium* (Seedorf *et al*., 2008).

Not only do genomes allow for the discovery of more genes but they also help us to understand how genes and genomes are evolving, as this can provide clues to gene function. Using genome comparison has led to the concept of 'Pan-genome', which refers to the full repertoire contained within a species. The Pan-genome theory predicts that any bacterial species will be made up of core set of genes that is found in all individuals and a dispensable set of genes that may or may not be present in any particular individual. By sequencing more and more individuals, the scale of the Pan-genome can be estimated. Therefore a single genome may give a very poor representation of the genetic potential of the species (Hall, 2007). The results from a number of completed genome projects have demonstrated that information on overall genome organization can provide biological insights. Whole genome sequencing represents the most powerful approach to identification of genomic diversity among closely related strains or isolates. However, such intergenome comparisons are greatly facilitated if at least one of the genomes is completely finished to a high degree of accuracy, rather than in the multiple unordered assemblies typical of a draft project (Fraser *et al*., 2002).

Chapter 2

Whole-genome sequencing of the extremophile *Thermus scotoductus* SA-01

2.1 Introduction

South African mines provide ready access to some of the world's deepest subterrestrial extreme environments. These mines allow investigators to collect water, rock and air samples for microbial and geochemical examination (Pfiffner *et al*., 2006) at depths of up to 4 kilometres below surface. In 1999, Kieft *et al.,* described the isolation and characterization of a facultatively anaerobic *Thermus* strain that is capable of dissimilatory iron reduction as well as growth with oxygen and nitrate as terminal electron acceptors. The authors obtained the strain by collecting rock and groundwater samples from the Witwatersrand Supergroup at a 3.2 km depth below surface in a South African gold mine operated by Western Deep Levels Inc. (currently Anglogold Ashanti). The Witwatersrand Supergroup is a 2.9 billion year old formation of low-permeability sandstone and shale with minor volcanic units and conglomerates. The ambient temperature of the rock at levels in excess of 3 km is approximately 60ºC (Kieft *et al*., 1999, Balkwill *et al*., 2004). The *Thermus* SA-01 strain is closely related to *Thermus* strains NMX2 A.1 and VI-7 (previously isolated from thermal springs in New Mexico, USA and Portugal, respectively). *Thermus* strains SA-01 and NMX2 A.1 have also previously been shown to grow using nitrate, Fe(III), Mn(IV) or Sº as terminal electron acceptors and to be capable of reducing Cr(VI), U(VI), Co(III) and the quinone-containing compound anthraquinone-2,6-disulfonate. Phylogenetic analyses of 16S rDNA sequences, BOX PCR genomic fingerprinting and DNA-DNA reassociation analyses indicated that these strains belong to the previously described genospecies *T. scotoductus* (Balkwill *et al*., 2004).

Although the physiology and genetics of the genus *Thermus* have been studied for three decades, *T. scotoductus* SA-01 is being used as a model organism in order to study the metabolic versatility of proteins and enzymes which are possibly partaking in energy conservation (van Heerden *et al*., 2008). In 2006, Möller *et al.*, reported the isolation of a soluble and membrane-associated Fe(III) reductase. In 2007, Opperman and van Heerden found that apart from *T. scotoductus* SA-01's ability to reduce Cr(VI) through a strictly anaerobic membrane-bound mechanism, it also has a second enzyme localized in the cytoplasm that can reduce Cr(VI) aerobically. The

membrane-associated chromate reductase has been purified to apparent homogeneity and shown to couple the reduction of Cr(VI) to NAD(P)H oxidation, with a preference toward NADH (Opperman and van Heerden, 2008). In addition, the cytoplasmic chromate reductase has been shown to be related to the Old Yellow Enzyme (Opperman *et al*., 2008).

Many thermophiles and hyperthermophiles have been isolated from hot springs and other thermal environments. By 2004, the complete genome sequences of 19 thermophilic or hyperthermophilic prokaryotic species have been determined (Takami *et al*., 2004). According to the Genamics GenomeSeek database, the genome sequence of 105 thermophilic and 10 hyperthermophiles have been sequenced to date (http://genamics.com/genomes/index.htm). In 1974, Oshima and Imahori isolated a non-sporulating, thermophilic bacterium from a hot spring in Japan. One strain was capable of growing at over 80ºC and was tentatively placed in the genus *Flavobacterium*. This thermophilic organism was then transferred to the genus *Thermus* as *T. thermophilus*. The genome sequence of *T. thermophilus* HB27 has been completed and published, the first for the genus *Thermus* (Henne *et al*., 2004). Genome sequencing studies performed on extremophilic organisms have already made an impact on the research community as well as opened a detailed study on the physiology of extreme thermophilic bacteria (Park *et al*., 2003) and resulted in the discovery of a number of new genes with potential interest for biotechnological interest (Lioliou *et al*., 2004). Six species belonging to the genus *Thermus* have been validly described i.e. *T. aquaticus*, *T. filiformis*, *T. thermophilus*, *T. scotoductus*, *T. brockianus* and *T. oshimai*) [Moreira *et al*., 1997]. According to DSMZ - Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (German Collection of Microorganisms and Cell Cultures), since then other *Thermus* sp. have been identified such as *T. antranikianii, T. chliarophilus, T. igniterrae, T. profundus, T. ruber* and *T. silvanus.* There is also a high frequent occurrence of small plasmids in *Thermus* but megaplasmids can also be present in this genus (Moreira and Sá-Correia, 1997).

The extreme thermophile *T. thermophilus* HB27, an aerobic, rod-shaped, gram-negative bacterium that grows at temperatures between 50ºC and 82ºC, is known to exhibit high frequencies of natural transformation (Friedrich *et al*., 2001). Despite the substantial evidence that there has been massive DNA transfer between archaea and hyperthermophilic bacteria and the assumption that life originated in hot ecosystems, information on DNA transfer in hot environments and on the surface and

the function of transformation systems in extreme environments is scarce (Freidrich *et al*., 2003). Thermophilic bacteria were also found to clearly stand out in terms of interdomain DNA transfer, for example, 24% and 16% of the genes in the hyperthermophilic bacteria *Thermotoga maritima* and *Aquifex aeolicus*, respectively, are suggested to be transferred from archaeal hyperthermophiles. Moreover, adaptation to life at high temperatures appears to be closely linked to horizontal gene transfer (Schwarzenlander and Averhoff, 2006).

Another example of comparative genomics is between the genus *Thermus* and *Deinococcus*. After the radiation from their common ancestor, these two lineages have taken divergent paths towards their distinct lifestyles. In addition to their extensive gene loss, *Thermus* seems to have acquired numerous genes from thermophiles, which likely was the decisive contribution to its thermophilic adaptation. However, by contrast, *Deinococcus* lost few genes but seems to have acquired many bacterial genes that apparently enhanced its ability to survive different kinds of environmental stress. Also, the single megaplasmids of *Thermus* and the DR177 megaplasmid of *Deinococcus* are homologous and probably were inherited from the common ancestor of these bacteria (Omelchenko *et al*., 2005). Surprisingly, *T. thermophilus* takes up DNA not only from thermophilic and mesophilic bacteria but also from various members of the Archaea and Eukarya (Schwarzenlander and Averhoff, 2006). Thus, comparative genomics is a useful approach for extracting candidate genes associated with thermophily (Takami *et al*., 2004).

Microorganisms are able to exploit very different environments and therefore must have evolved phenotypic traits allowing adaptation and survival under very different environmental conditions (Averhoff, 2004). In the last decade considerable effort has been expended in research on the biochemistry and physiology of thermophilic bacteria. The molecular parameters of different proteins, including enzymes that are operative under extreme conditions, are still under active investigation. According to Pantazaki *et al*., 2002, the fundamental cell processes such as replication, transcription, translation, secretion, cell-signalling etc., in thermophilic bacteria is yet to be understood. Based on the current knowledge of *T. scotoductus* SA-01's interaction with metals, thereby opens the debate on whether such organisms are able of conserving energy by not employing elaborate processes for every metal encountered, but rather multitasking or moonlighting of metabolic proteins in order to sustain themselves in these environments (van Heerden *et al*., 2008), led to the aim of this study, which was to determine the complete genome sequence of this

organism. Every genome that has been sequenced to date has been able to provide new insight into the biological processes, activities and potential of these species, which had not been evident before (Lioliou *et al*., 2004). Studying the operons present in the complete genome of the extremophile *T. scotoductus* SA-01 will form the basis for further investigation into genome evolution and gene regulation in extreme environments.

2.2 Materials And Methods

2.2.1 Culture Preparation

Thermus scotoductus SA−01 (ATCC 700910) was routinely cultured aerobically in a complex organic medium, TYG [5 g tryptone, 3 g yeast extract and 1 g glucose in 1 L dH_2O] at 65°C, pH 7.0, with shaking (200 rpm). The strain was examined for purity by streaking onto TYG medium solidified with 2% agar. Frozen stocks were maintained in 15% glycerol at −80°C.

Escherichia coli TOP10 (Invitrogen) competent cells were used as cloning host and was grown in Luria-Bertani (LB) medium [10 gL^{-1} tryptone, 5 gL^{-1} yeast extract and 5 gL^{-1} NaCl (pH 7)] at 37°C with aeration (200 rpm). Ampicillin (100 $\mu g.ml^{-1}$) was added when required.

2.2.2 Genomic DNA extraction using commercial kits

Genomic DNA Isolations were done using the Wizard Purification Kit (Promega) and ZR Soil Microbe DNA Kit™ (Zymo Research) and performed according to the manufacturer's instructions. DNA concentration readings were taken using the Nanodrop 2000 and the Nanodrop 3300 Flourospectrophotometer that utilizes a PicoGreen stain that specifically detects double-stranded DNA.

2.2.3 Strain verification

Strain identity was verified by amplification of the 16S rRNA gene using genomic DNA as template. RFLP analysis of the PCR product using specific restriction enzymes *Eco*RI, *Sma*I and *Bsr*DI and using WebCutter for analysis (www.firstmarket.com/cutter/cut2.html). PCR products were also cloned and a number of clones were subsequently sequenced.

2.2.4 Cloning and Screening of 16S rRNA PCR products

2.2.4.1 PCR amplification of 16S rRNA (Prokaryotes)

Bacterial-specific primers 27F and the universal 1492R primer (Lane, 1991) were used to amplify the 1 500 bp 16S rRNA genes. A Hotstart PCR amplification approach was carried out in a PxE Thermal Cycler (Thermo Electron) in a total reaction mixture volume of 50 μl.

The PCR reaction mixture contained 1.0 μl of template DNA, 1 μl of the 27F primer (10 μM), 1 μl of the 1492R primer (10 μM), 2.0 μl of 20 mM deoxynucleoside triphosphates (dNTPs), 0.25 μl (5 $U.\mu l^{-1}$) of Taq DNA polymerase, 5.0 μl of 10 X buffer, 3 μl of 25 mM $MgCl_2$, 1 μl of 10% BSA and 35.75 μl of sterile distilled water.

The reaction mixture was incubated at 95°C for 5 mins to denature the DNA. This was followed by 30 cycles of amplification, each of which consisted of three steps in the following order: denaturation at 95°C for 30 sec, annealing at 52°C for 45 sec and extension of the primers at 72°C for 1 min 30 sec. Final extension was at 72°C for 10 mins. Amplification products were visualized in an ethidium bromide containing 1% agarose gel using an UV transilluminator after electrophoresis at 100 V for 90 mins.

2.2.4.2 Ligation of DNA fragments

PCR products were purified from the agarose gel using the GFX™ PCR, DNA and Gel Band Purification Kit (Amersham Biosciences). Ligation of cloning vectors (pGEM®-T Easy Vector [50 ng]) and PCR products were performed using T4 DNA ligase. Typical ligation reactions contained 1 μl vector, 2 μl deionised water, 5 μl of 2X ligation buffer, 1 μl PCR product and 1 U of T4 DNA ligase in a total volume ranging from 10 μl. Reactions were incubated overnight at 4°C.

2.2.4.3 Bacterial Transformation

Competent *E. coli* TOP10 cells were transformed with 10 μl of the ligation mixture. Cells were allowed to slowly thaw on ice followed by the addition of the ligation mixture and gentle mixing. Cells were then incubated for 30 min on ice to allow the DNA to bind to the cells, heat shocked for 40 sec in a 42°C water bath to allow the DNA to enter the cells and immediately placed on ice for 2 min. The mixture was

suspended in 800 μl aliquot of LB mix (5 ml LB + 100 μl + 1 M glucose + 50 μl 2 M Mg^{2+}) and incubated at 37°C with shaking for one hour at 200 rpm. Transformed cells were then centrifuged at 4000 rpm for 1 min and 650 μl of the resulting supernatant was removed. The cells were resuspended in the remaining medium and 100 μl of the cell suspension then spread on LB agar plates containing ampicillin, X-gal and IPTG. Blue-white selection was carried out to identify transformants.

2.2.4.4 Screening of transformed cells

Ligation of the PCR fragment and the pGEM®-T Easy Vector was done using the pGEM®-T Easy Vector System (Promega). These ligation mixtures were then transformed into *E. coli.* This allowed early differentiation of *E. coli* cells harboring recombinant plasmids (white colonies) from those carrying only pGEM®-T (blue colonies). White colonies on LB agar plates were isolated and sub-cultured on master plates.

Individual colonies, from master plates with the region containing the putative PCR product, in pGEM®-T, were inoculated into 5 ml of LB broth (supplemented with 100 $\mu g.ml^{-1}$ ampicillin). Cell suspensions were incubated for 16 hrs at 37°C in an orbital shaker at 200 rpm. *E. coli* cells were tested for the presence of recombinant plasmids by plasmid isolation and restriction analysis. Plasmid DNA was isolated using the GeneJet™ Plasmid Miniprep Kit according to the manufacturers instructions.

2.2.4.5 Restriction Fragment Length Polymorphism (RFLP) and Sequence Analysis

Restriction analysis of isolated plasmids was done using the restriction enzymes *Eco*RI, *Sma*I and *Bsr*DI.

2.2.4.6 Sequencing

Plasmid inserts were sequenced by Inqaba Biotechnical Industries (South Africa) using the ABI 3130xl genetic analyzer (Applied Biosystems, Foster City, CA), incorporating the ABI Big Dye Terminator Cycle Sequencing kit version 3.1 (Applied Biosystems, Foster City, CA) using the universal SP6 and T7 promoter primers. Electropherograms of the generated sequences were inspected with FinchTV software (Geospiza) and Vector NTI (Invitrogen). Sequence alignments were performed using the DNAssist program (Patterton and Graves, 2000).

2.2.5 High-throughput 454-pyrosequencing (GS20/FLX)

2.2.5.1 Library construction and DNA pyrosequencing

Two 454-pyrosequencing runs were performed during the project. The first made use of the Roche GS20 system (producing 100 bp reads). A second run was then done using the latest FLX system (producing 200 bp reads).

A *T. scotoductus* DNA library was constructed and sequenced at Inqaba Biotec, Pretoria, South Africa. The methods were all standard protocols developed for the Roche GS20/FLX sequencer (Fig 2.1). The library was constructed by shearing the isolated genomic DNA into fragments, which were blunt-ended and phosphorylated by enzymatic polishing using T4 DNA polymerase, T4 polynucleotide kinase and Klenow DNA polymerase. The polished DNA fragments were then subjected to adapter ligation followed by isolation of the single-stranded template DNA (sstDNA). The quality and quantity of the sstDNA library was assessed using the Agilent 2100 Bioanalyzer. The sstDNA library fragment was captured onto a single DNA capture bead and clonally amplified within individual emulsion droplets. The emulsions were disrupted using isopropanol, the beads without an amplified sstDNA fragment were removed, and the beads with an amplified sstDNA fragment were recovered for sequencing. The recovered sstDNA beads were packed onto a 70-75 mm PicoTitrePlate™ and loaded onto the GS20/FLX Sequencing System (454 Life Sciences). The raw reads obtained were assembled into contigs using the Newbler Assembly software.

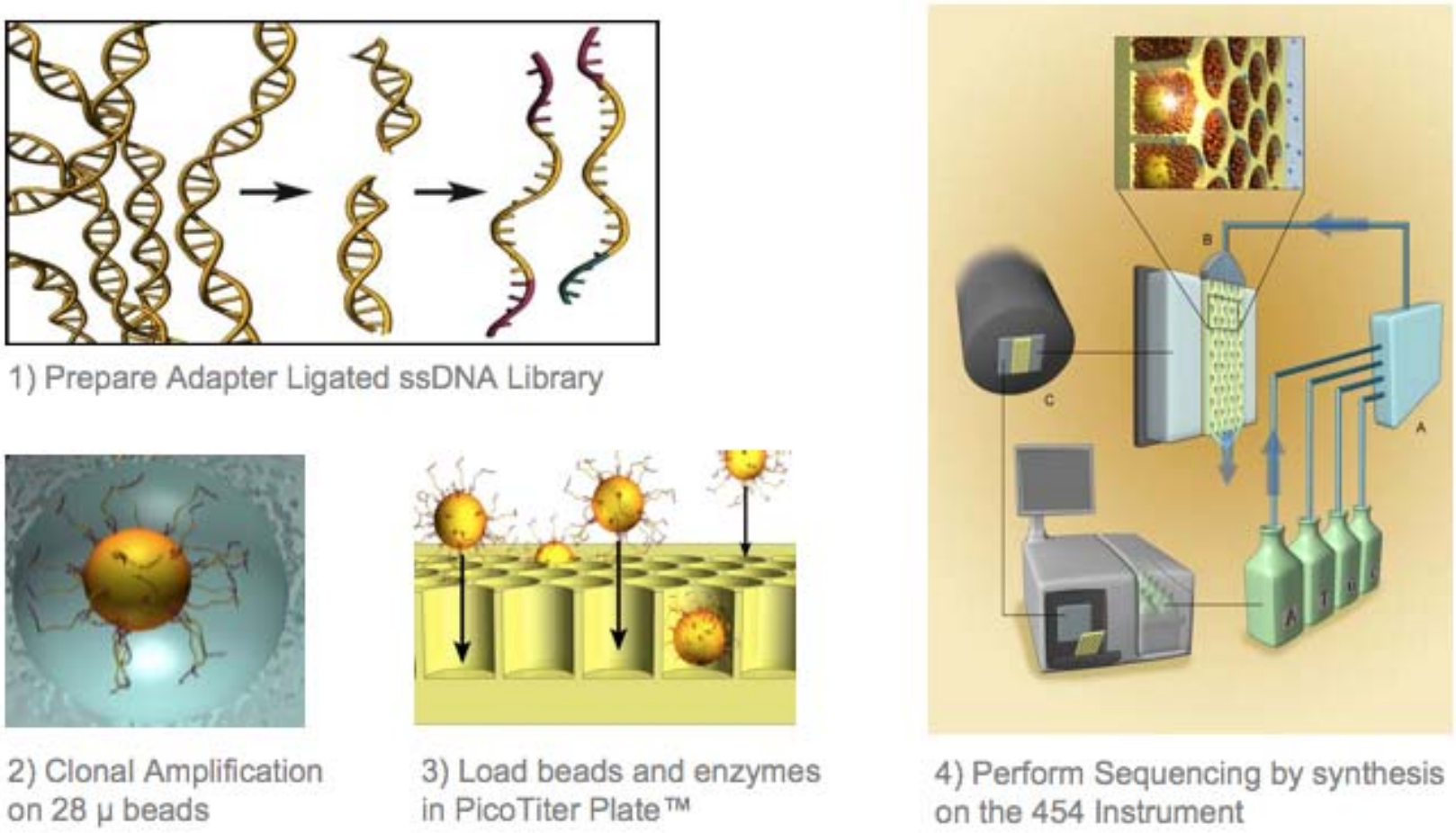


Fig 2.1 Steps involved in the library construction and sequencing of DNA using the GS20/FLX pyrosequencing system (Roche Applied Science).

2.2.6 Assembly analysis

Assembly of the sequence reads were performed using Newbler RunAssembly software with different input parameters to determine the reproducibility. This was done by placing the input files in ascending and descending order according to file size. The resulting contigs were compared to the original data obtained from Inqaba Biotec using BLASTn. Assembly of the reads by mapping onto the *Thermus thermophilus* HB27 chromosome sequence was also performed using RunAssembly software.

The raw reads obtained from both the GS20 and FLX pyrosequencing runs were assembled into contigs using the Newbler RunAssembly software using the GS20 version and the latest GS/FLX version.

2.2.7 Genome Alignment

Alignment of the contigs (draft genome) and the completely assembled chromosome of *T. scotoductus* SA-01 were performed using various programs (Table 2.1).

2.2.8 Reverse-BLAST Analysis

A reverse-BLASTn (Altschul *et al*., 1990) was done with the *T. scotoductus* SA-01 contigs against the *T. thermophilus* HB27 and HB8 genome sequences as well as the megaplasmid (pTT27) and plasmid sequences (pTT8). This was done to determine the presence of a megaplasmid or plasmid sequence present in the draft genome of *T. scotoductus* SA-01.

Table 2.1 Genome alignment using various programs.

Genome Mapping of *T. scotoductus* SA-01		
Software Used	**Reference strain**	**Reason**
1. MUMmer (Kurtz *et al*., 2004)	*Thermus thermophilus* HB27 and *Thermus thermophilus* HB8	MUMmer is generally used for the alignment of 2 genomes whether complete or draft sequences. Here the complete genome of *T. thermophilus* HB27 and *T. thermophilus* HB8 were first aligned to test the comparison of 2 closely related organisms using MUMmer. The contigs and the completed chromosome of *T. scotoductus* SA-01 were then used and compared to the sequences with *T. thermophilus* HB27 for overall alignment.
2. WebACT (Carver *et al*., 2005)	*Thermus thermophilus* HB27	The web-based program ACT: DNA Sequence Comparison Viewer, found at http://www.sanger.ac.uk/software/ACT, was used to align the complete chromosome of *T. thermophilus* HB27 and the assembled contigs and completed chromosome of *T. scotoductus* SA-01 genome. This was done in order to confirm why mapping could not be used as an aid in assembly of the reads and closing the chromosome.

2.2.9 Fosmid Library Construction for *T. scotoductus* SA-01

Fosmid Library construction performed according to protocol (Fig 2.2) obtained from DOE Joint Genome Institute, Department of Energy, Office of Science (USA).

2.2.9.1 Shearing of gDNA using Hydroshear

The Hydroshear (Gene Machines) was washed before and after shearing the sample with 60 μl of 0.2 M HCl (3 times), 0.2 M NaOH (2 times) and TE buffer (5 times). The sample was passed through the syringe for 4 cycles at a medium speed. Genomic DNA (~20 μg) was added into an Eppendorf tube and brought to a final volume of 200 μl with 10 mM Tris, if required. The gDNA was then sheared for 25 cycles at a speed code of 17. Thereafter, the gDNA was collected in a clean tube and concentrated to a volume of 60 μl using the rotary evaporator (Savant) and finally placed on ice immediately.

2.2.9.2 Blunt End Repair

The reaction was performed on ice by adding the following reagents from the Epicentre Fosmid Library Kit:

	1X
Sheared DNA	50-60 μl
10 X Buffer	8 μl
2.5 mM dNTPs	8 μl
10 mM ATP	8 μl
ER enzyme mix	4 μl
Total volume	**88 μl**

The Eppendorf tube was capped, rapidly votexed and spun down. The mixture was then incubated at room temperature for 45 mins where after the sample was heat inactivated at 70°C for 10 mins.

2.2.9.3 Phenol Extraction

Samples were spun down at 10 000 rpm for 2 mins. The volume of the sample was measured and an equal volume of phenol (Sigma) was added to the sample. The

sample was vortexed for 15-30 sec and then spun for 5 mins at 10 000 rpm. The top aqueous layer was carefully pipetted into a new 2 ml round bottom tube.

2.2.9.4 Ethanol Precipitation

To the measured sample volume, 1/10 volume of 1 M NaCl, 1.5 μl pellet paint and 2.5 volumes of 96% absolute ethanol was added. The mixture was vortexed well and tubes placed at -80°C for at least 30 mins. The tubes were then spun down for 20 mins at 13 000 rpm at 4°C. The supernatant was discarded, keeping an eye on the pink pellet. The pellet was then washed twice with 200 μl cold 96% ethanol. Using a pipette, the supernatant was carefully removed being careful of "the wiley pellet". The pellet was vacuum dried for approximately 5 mins with no heat, resuspended in 30 μl Tris-HCl, vortexed, spun down and placed at 25°C to fully resuspend the pellet. The sample concentration was measured and stored at -20°C until needed.

2.2.9.5 Ligation Reaction

The following was combined in an Eppendorf tube:

	2 X Ligation	**Fosmid Control DNA**
10 X Ligation Buffer	2	1
ATP (10 mM)	2	1
pCC1FOS Vector	2	1
Insert DNA (40 kb)	6	2.5 (Fosmid Control DNA)
DNA Ligase	2	1
dH_20	6	3.5
Total Volume	**20 μl**	**10 μl**

Tubes were incubated at 16°C overnight for primary ligation. Following overnight ligation, an additional 0.5 μl of DNA ligase was added to the ligation mix and incubated for a further 90 mins at room temperature. Samples were then heat inactivated for 10 mins at 70°C and subsequently cooled on ice for 10 mins. Samples were stored at -20°C until required.

2.2.9.6 Preparation of Infection Cells

The day before plating, 1 μl of *E. coli* EPI300 cells were inoculated into a 500 ml flask containing 100 ml LB broth and 1 ml 1M $MgSO_4$. The culture was incubated overnight at 37°C with shaking at 200 rpm. The next morning, 5 ml of overnight grown culture was inoculated into a 500 ml flask containing 100 ml LB broth and 1ml 1 M $MgSO_4$. The flask was incubated at 37°C with shaking until an OD_{600} of between 0.8 – 1. The culture was stored at 4°C for a maximum of 5 days.

2.2.9.7 Packaging

One tube of packaging extract (1 tube = 2 reactions, 50 μl) was thawed on ice. Once completely thawed, 25 μl of packaging extract was added to 10 μl of ligation mix. The sample was mixed gently, quickly spun down and incubated at 30°C for 90 mins. After incubation, the remaining 25 μl of packaging extract was added to the sample and further incubated for another 90 mins at 30°C. Phage dilution buffer was then added to a final volume of 1 ml followed by 25 μl of chloroform. The sample was stored at 4°C until further use.

2.2.9.8 Infection

During infection, 3 X 15 μl of packaging/phage dilution buffer solution was added to 150 μl of infection cells (OD=0.9) and incubated at 37°C for 90 mins. After incubation, 100 μl of the transformation mix was plated out on LB agar plates containing 12.5 μg chloramphenicol and incubated at 37°C overnight. In addition, a 10% glycerol transformation stock (157 μl 80% glycerol + full transformation = 1257 μl glycerol transformation stock) was made and mixed by inverting the tube several times. The glycerol stock was then stored at -80°C.

2.2.9.12 Fosmid Control DNA

A 1:100 and 1:1000 dilution was made of the Fosmid control DNA using the Phage Dilution Buffer. Ten microlitres of the diluted and undiluted Fosmid control DNA was then added to 100 μl of the Epi300 T1 *E. coli* cells and grown to an OD of 0.9 while being incubated at 37°C for 90 mins. After incubation the total volume of 110 μl was plated out onto LB agar plates containing 12.5 μg chloramphenicol and incubated at 37°C overnight.

2.2.9.10 Induction of clones

For the pre-culture, colonies were picked up into flat-bottom blocks containing LB broth with 12.5 μg chloramphenicol and incubated at 37°C with shaking (350 rpm). Thereafter, 150 ml 2 X LB, 75 μl chloramphenicol and 1.5 ml Induction solution was mixed together and 1060 μl dispensed into each well. To each well, 240 μl of the pre-culture was added and further incubated for 5 hrs at 37°C.

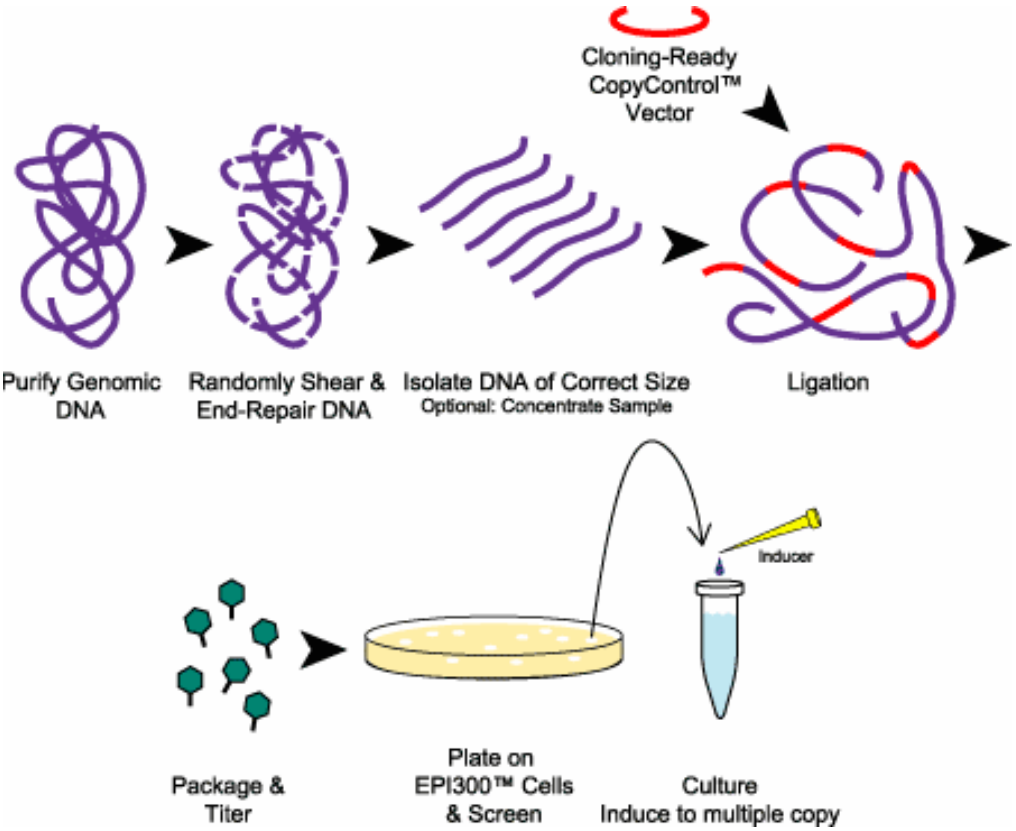


Fig 2.2 Schematic representation of steps used for a fosmid library preparation (Taken from Epicentre Biotechnologies).

2.2.9.11 Plasmid DNA isolation

Plasmid DNA isolation from fosmid clones was performed using the Automated JANUS Workstation from Perkin Elmer. This automated robot system is designed to perform pDNA isolations from a 96-well plate.

2.2.9.12 DNA sequencing with the ABI 3730xl Automated Sequencer (Applied Biosystems)

Fosmid clones were picked up into 4 X 96 well plates and sequenced using the ABI 3730xl, 96-capillary electrophoresis system that is a sensitive detection system and gives long sequence reads (up to a 1000 bases for high quality DNA).

According to the ABI manufacturers specifications the sequence reaction contained the following:

DNA	12 μl
Big Dye	4 μl
5 X Seq Buffer	2 μl
Primer (pCC1f/pCC1r), 5 pmol	2 μl
Total	**20 μl**

Table 2.2 ABI-Plasmid-Cycle programme.

ABI cycle sequencing conditions		
Lid	110°C On	
Pause	95°C forever	
Hold	95°C for 3 mins	
Hold (Denaturation)	95°C for 30 sec	x 25 cycles
Hold (Annealing)	53°C for 30 sec	
Temp (Elongation)	72°C for 1 min (1 min/kbp)	
Hold (Elongation)	72°C for 10 mins	
Temp	4°C forever	
End		

2.2.10 16S rRNA Library Construction for determining RNA clusters

2.2.10.1 Prokaryotic 16S rRNA PCR

PCR reactions were performed on undiluted and diluted 1:10, 1:100, and 1:1000 gDNA samples of *T. scotoductus* SA-01 as per protocol 2.2.4.1.

The PCR conditions differed slightly from previously mentioned as the standard protocol of the Göttingen Genomics Laboratory was used at this time. The reaction mixture was incubated at 94°C for 5 mins to denature the DNA. This was followed by 30 cycles of amplification, each of which consisted of three steps in the following order: denaturation at 94°C for 1 min, annealing at 45°C for 45 sec and extension of the primers at 72°C for 1 min 30 sec. Final extension was at 72°C for 5 mins. Amplification products were visualized on an ethidium bromide containing 1% agarose gel using an UV transilluminator after electrophoresis at 100 V for 90 mins.

2.2.10.2 Ligation of DNA fragments

PCR products were purified from the agarose gel using the peqGOLD Gel Extraction Kit (peQLab.Biotechnologie GmbH). The TOPO TA Cloning Kit (Invitrogen), which provides a highly efficient, 5 minute, one-step cloning strategy for the direct insertion of Taq polymerase amplified PCR products into a plasmid vector, was used. The linearised plasmid vector contains a single 3′-thymidine (T) overhangs for TA cloning and the topoisomerase I from *Vaccinia* virus covalently bound to the vector (Fig 2.3).

The ligation reaction contained 2 μl of the PCR product, 0.5 μl salt solution and 0.5 μl TOPO vector. The ligation mix was mixed gently and incubated at room temperature for 5 mins.

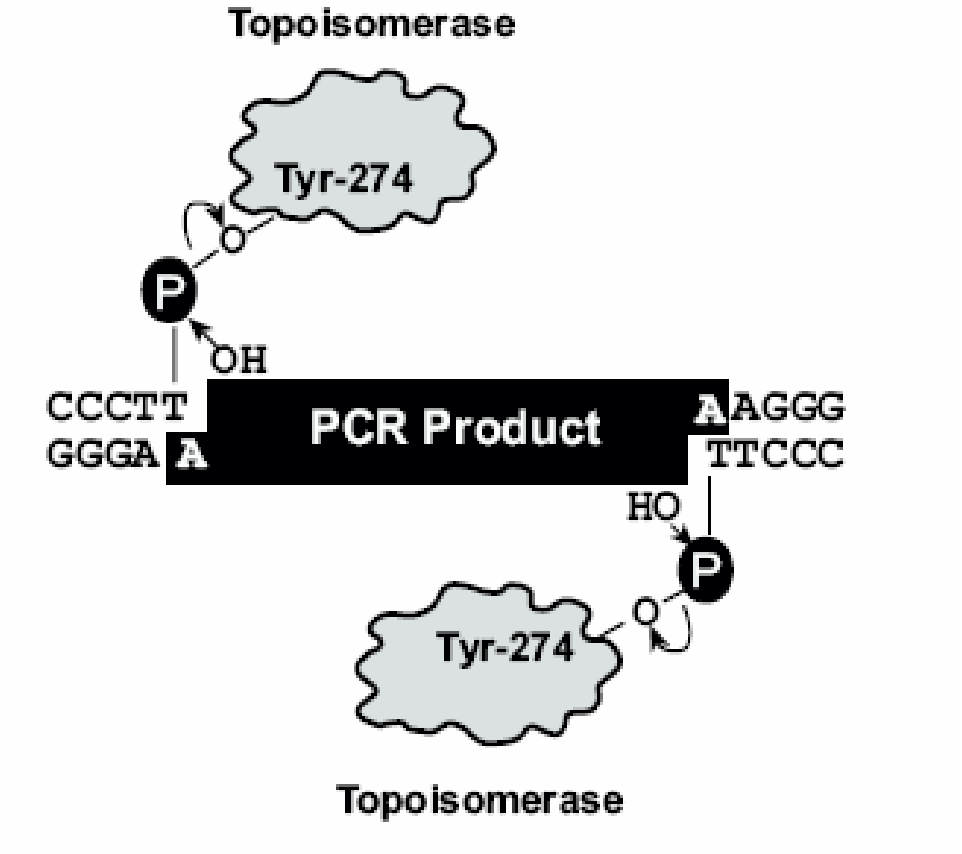


Fig 2.3 The TOPO TA cloning system from Invitrogen, containing the topoisomerase I for 5 minute cloning of *Taq* polymerase-amplified PCR products.

2.2.10.3 Bacterial Transformation and Screening

Blue-white selection of transformants was carried out to identify transformants as previously described in 2.2.4.3.

Plasmid DNA was isolated using the peqGOLD Plasmid Miniprep Kit (peQLab Biotechnolgie GmbH) according to manufacturer's instructions.

2.2.11 Sequence Analysis

Sequencing of 10 clones was done using the ABI 3730xl DNA Analyzer (Applied Biosystems).

2.2.12 Raw Data Processing

Sequences were processed first with Phred (Ewing and Green, 1998) and Pregap4, joined to assembled contigs and edited with Gap4, which is a part of the Staden v4.11 package software (http://staden.sourceforge.net/) (Staden *et al*., 2000).

2.2.13 Order of Contigs for Whole Genome

The order of the contigs was determined by adding the fosmid library clone sequences containing inserts of approx. 40 kb in size to the assembled contigs. Using the Gap4 program, a contig order was established by checking for fosmid spanning gaps and in this way many supercontigs could be obtained.

2.2.14 Gap Closure Strategies

2.2.14.1 Gap Closure by BLASTn Analysis

In order to close the remaining gaps, one strategy used was to BLASTn approximately 100 bases of the ends of the contigs against the public database (Altschul *et al*., 1990), to determine the order of the contigs.

2.2.14.2 Gap Closure using PCR

Primers were designed at the ends of each contig using STADEN. Using the results from the contig order, PCR was done using primers for 2 specific ends of 2 contigs that had a fosmid-spanning gap.

Depending on the predicted length of the gap, a standard or long-range PCR was performed (Table 2.3). Annealing and extension times were optimised to remove non-specific bands. In addition different polymerase systems [e.g. Accuprime GC-rich polymerase (Invitrogen), 5′-Prime Extender system (Eppendorf)] were also used for optimum PCR results.

Table 2.3 Standard and Long range PCR conditions for gap closure.

Standard PCR (1-2 kb)	Long Range PCR (5 kb)
Heat lid to 98°C Pause @ forever	Heat lid to 98°C Pause @ forever
Hold 98°C for 5 mins	Hold 98°C for 5 mins
Hold 98°C for 20 sec Hold 53/55/57°C for 10 sec Hold 72°C for 1 mins } X 30 cycles	Hold 98°C for 20 sec Hold 53/55/57°C for 10 sec Hold 72°C for 1 mins } X 30 cycles
Deactivate lid heating	Deactivate lid heating
Hold @ 20°C forever	Hold @ 20°C forever

2.2.14.3 Gap Closure using Fosmid Walking

Sequencing was also done on fosmid clones as DNA template, which spanned the region containing the gap using the respective primers designed from the ends of the contigs.

2.2.15 ORF Corrections

The automatically annotated ORFs from TIGR (www.tigr.org/AnnotationEngine) were loaded into Artemis (Rutherford *et al*., 2000) for manual checking and correction of open reading frames. Artemis is a free genome viewer (Sanger Centre website: http://www.sanger.ac.uk/Software/Artemis/) and annotation tool that allows visualisation of sequence features and the results of analyses within the context of the sequence, and its six-frame translation.

2.2.16 Annotation

2.2.16.1 Automatic Annotation

Two sets of automatic annotation were performed. The first set was the remaining 301 contigs obtained after the GS20 run from an initial 371 contigs, 70 of which were manually joined using the Gap4 program. The second automatic annotation was performed on the 35 contigs obtained after the combined GS20/FLX pyrosequencing run. The contigs were concatenated together and were joined with the following sequence:
"NNNNNCACACACTTAATTAATTAAGTGTGTGNNNNN". This inserts stop codons in all six reading frames. The DNA sequence was submitted to Craig Venter Institute for Genome Research (TIGR/JCVI) Annotation Engine (www.tigr.org/AnnotationEngine), where it was run through TIGR's prokaryotic annotation pipeline. Included in the pipeline is gene finding with Glimmer, BLAST-extend-repraze (BER) searches, HMM searches, TMHMM searches, SignalP predictions and automatic annotations from AutoAnnotate. All of this information is stored in a SQL database and associated files, which was downloaded to our site. The manual browsing and annotation tool Manatee was downloaded from SourceForge (manatee.sourceforge.net) and used to manually review the output from the prokaryotic pipeline of the Annotation Engine.

2.2.16.2 Manual Annotation

The resulting automatic annotation done by TIGR on the 35 contigs was then manually checked using the ERGO Tool (Integrated Genomics, Inc., http://www.integratedgenomics.com). Each ORF prediction was verified and modified manually by searching derived protein sequences against public nucleotide or protein domain databases such as GenBank or PFAM, respectively. The wide range of software and databases were used which are shown in Table 2.4.

Table 2.4 List of databases and software used for Manual Annotation.

Program	Description
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool, or **BLAST**, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. http://www.ncbi.nlm.nih.gov/BLAST
COGs	**C**lusters of **O**rthologous **G**roups of proteins, **(COGs)** were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. http://www.ncbi.nlm.nih.gov/COG
PFAM	The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Proteins are generally composed of one or more functional regions, commonly termed domains. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function. http://pfam.sanger.ac.uk/
PROSITE	PROSITE is a database of protein families and domains. It is based on the observation that, while there is a huge number of different proteins, most of them can be grouped, on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor.

	PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins. http://www.expasy.ch/prosite/
LipoP 1.0 Server	The LipoP 1.0 server produces predictions of lipoproteins and discriminates between lipoprotein signal peptides, other signal peptides and n-terminal membrane helices in Gram negative bacteria http://www.cbs.dtu.dk/services/LipoP
SignalP 3.0 Server	SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models. http://www.cbs.dtu.dk/services/SignalP/
TMpred	The TMpred (Prediction of transmembrane regions and orientations) program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occuring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. http://www.ch.embnet.org/software/TMPRED_form.html
InterPro	InterPro is a database of protein families, domains, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences. http://www.ebi.ac.uk/interpro/

ExPASy	The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE. The UniProt Knowledgebase consists of: • **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. • **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. These databases are developed by the Swiss-Prot groups at SIB and at EBI.
Manatee	Manatee is a web-based gene evaluation and genome annotation tool that can view, modify, and store annotation for prokaryotic and eukaryotic genomes. The Manatee interface allows biologists to quickly identify genes and make high quality functional assignments using a multitude of genome analyses tools. These tools consist of, but are not limited to GO classifications, BER and BLAST search data, paralogous families, and annotation suggestions generated from automated analysis. The Manatee project was created by the bioinformatics department at The Institute for Genomic Research (TIGR) in Rockville, MD. http://manatee.sourceforge.net/

2.2.17 Polishing of Genome Sequence

Resequencing of certain areas in the genome was redone due to the quality of sequences. Polishing was performed by checking the Confidence Value graph using the Gap4 program, which had a cut-off value of 45. Sequences with an average cut-off of <45 had to be resequenced to determine high-level accurate base-calling. The Confidence Value consensus algorithm produces a consensus sequence for which the expected error rate for each base is known and would calculate the expected number of errors in a particular consensus sequence. For example, if 50 bases in the consensus had confidence 10, we would expect those 50 bases (with an error rate of 1/10) to contain 5 errors; and if 200 bases had confidence 20, we would expect them to contain 2 errors. If these 50 bases with confidence 10, and 200 bases with confidence 20 were the least accurate parts of the consensus, they are the bases which we should check and edit first. In so doing we would be dealing with the places most likely to be wrong, and would raise the confidence of the whole consensus.

2.2.18 Insertion Sequence (IS) Search

An insertion sequence (IS) search was performed to determine if the genome sequence contained any possible IS elements. The search was done using the web-based IS Finder program at http://www-is.biotoul.fr/is.html.

2.2.19 Bi-directional BLAST

The proteins obtained from the genome sequence of *T. scotoductus* SA-01 were bi-blasted against four chosen organisms using BLASTp. The software performing the bi-directional BLAST is implemented in Java for platform-independency and easy expansion. The NCBI-BLAST suite and the EMBOSS package are used for the BLAST search and the global Needleman-Wunsch-Similarities are generated (Wollherr and Liesegang, 2008, Poster presentation, unpublished).

The organisms of choice for the bi-directional BLAST with *T. scotoductus* SA-01 were the related *Thermus thermophilus* HB8 (Oshima & Imahori, 1974), *Thermus thermophilus* HB27 (Henne *et al*., 2004), *Deinococcus radiodurans* (Makarova *et al.,* 2001), the *Desulforudis audaxviator* (Chivian *et al*., 2008) genome sequence obtained by metagenome sequencing and the metal reducers *Shewanella oneidensis* MR-1 (Heidelberg *et al*., 2002) and *Geobacter sulfurreducens* PCA (Methe *et al*.,

2003). Bi-directional BLAST results were visualized with MS Excel and a colour-key chosen for mapping different levels of similarities.

2.3 Results And Discussion

2.3.1 Isolation of genomic DNA using Commercial Kits

In order to do a run of the library preparation genomic DNA was isolated using the Wizard Kit (Promega) and ZR Soil Microbe DNA Kit™ (Zymo Research) and quality controls of *T. scotoductus* DNA were done (Fig 2.4). The gDNA that was isolated was found to be pure and the total DNA yields using the commercial kits were > 10 μg (Fig 2.4(i)). The restriction analysis on the 16S rRNA gene PCR product also confirmed the strain was *T. scotoductus* (Fig 2.4(ii)). Fig 2.5 shows the alignment of the sequence obtained with the partial 16S ribosomal RNA gene of *Thermus* SA-01 (NCBI: AF020205; Kieft *et al.*, 1999).

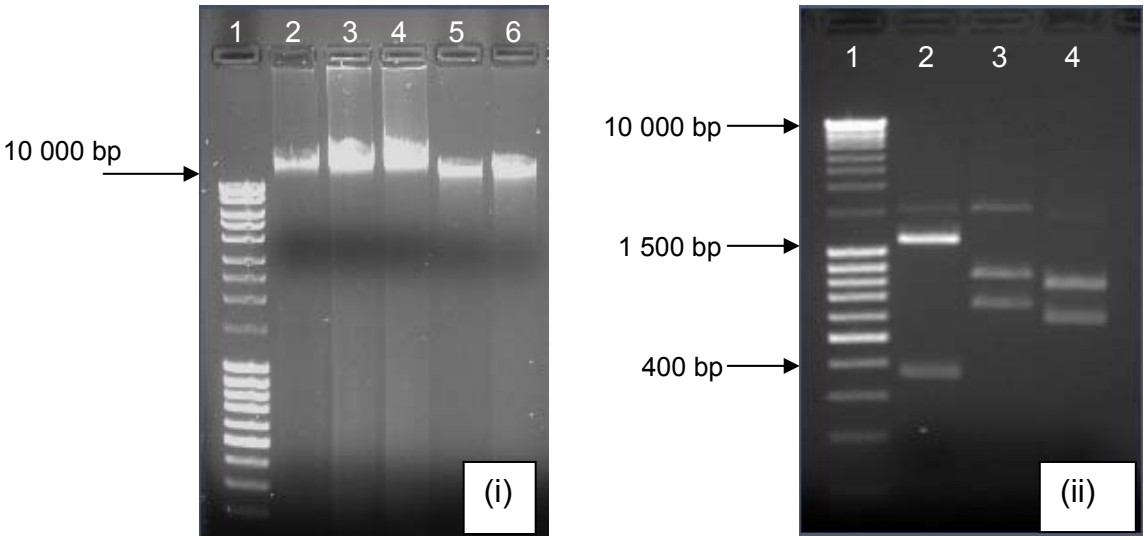


Fig 2.4 *Thermus scotoductus* SA-01 strain quality controls. (i) DNA isolations of *T. scotoductus* SA-01 strain using 2 commercial kits. Lanes 1: MassRuler, Lane 2-4: Genomic DNA isolated using Wizard kit (Promega) and lanes 5-6: genomic DNA isolated using ZR Soil Microbe DNA Kit (Zymo Research). (ii) Agarose gel showing restriction patterns of *T. scotoductus* 16S rDNA PCR product using 3 different enzymes. Lane 1: MassRuler, lane 2: 16S PCR product of *T. scotoductus* digested with *Bse*MI, lane 3: 16S PCR product of *T. scotoductus* digested with *Eco*RI and lane 4: 16S PCR product of *T. scotoductus* digested with *Sma*I.

```
Kieft et al,1999   1    AAGAGTTTGATCGTGGCTCAGGGTGAACGCTGGCGGCGTGCCTAAGACATGCAAGTCGAGCGGGGCAGGTTTATACCTGTCCAG 84
T.scotSA01         1     AGAGTTTGATCCTGGCTCAGGGTGAACGCTGGCGGCGTGCCTAAGACATGCAAGTCGAGCGGGGCAGGTTTATACCTGTCCAG 83

Kieft et al,1999   85   CGGCGGACGGGTGAGTAACGCGTGGGTGACCTACCCGGAAGAGGCGGACAACCTGGGGAAACCCAGGCTAATCCGCCATGTGGT 168
T.scotSA01         84   CGGCGGACGGGTGAGTAACGCGTGGGTGACCTACCCGGAAGAGGCGGACAACCTGGGGAAACCCAGGCTAATCCGCCATGTGGT 167

Kieft et al,1999   169  CCTGTCCTGTGGGGCAGGACTAAAGGGTGAATAGCCCGCTTCCGGATGGGCCCGCGTCCCATCAGCTAGTTGGTGGGGTAAAGG 252
T.scotSA01         168  CCTGTCCTGTGGGGCAGGACTAAAGGGTGAATAGCCCGCTTCCGGATGGGCCCGCGTCCCATCAGCTAGTTGGTGGGGTAAAGG 251

Kieft et al,1999   253  CCCACCAAGGCGACGACGGGTAGCCGGTCTGAGAGGATGGCCGGCCACAGGGGCACTGAGACACGGGCCCCACTCCTACGGGAG 336
T.scotSA01         252  CCCACCAAGGCGACGACGGGTAGCCGGTCTGAGAGGATGGCCGGCCACAGGGGCACTGAGACACGGGCCCCACTCCTACGGGAG 335

Kieft et al,1999   337  GCAGCAGTTAGGAATCTTCCGCAATGGACGGAAGTCTGACGGAGCGACGCCGCTTGGAGGAGGAAGCCCTTCGGGGTGTAAACT 420
T.scotSA01         336  GCAGCAGTTAGGAATCTTCCGCAATGGACGGAAGTCTGACGGAGCGACGCCGCTTGGAGGAGGAAGCCCTTCGGGGTGTAAACT 419

Kieft et al,1999   421  CCTGAACTGGGGACGAAAGCCCCGTGTAGGGGGATGACGGTAACCAGGTAATAGCGCCGGCCAACTCCGTGCCAGCAGCCGCGG 504
T.scotSA01         420  CCTGAACTGGGGACGAAAGCCCCGTGTAGGGGGATGACGGTACCCAGGTAATAGCGCCGGCCAACTCCGTGCCAGCAGCCGCGG 503

Kieft et al,1999   505  TAATACGGAGGGCGCGAGCGTTACCCGGATTTACTGGGCGTAAAGGGCGTGTAGGCGGCCTGAGGCGTCCCATGTGAAAGGCCA 588
T.scotSA01         504  TAATACGGAGGGCGCGAGCGTTACCCGGATTTACTGGGCGTAAAGGGCGTGTAGGCGGCCTGGGGCGTCCCATGTGAAAGGCCA 587

Kieft et al,1999   589  CGGCTCAACCGTGGAGGAGCGTGGGATACGCTCAGGCTAGAGGGTGGGAGAGGGTGGTGGAATTCCCGGAGTAGCGGTGAAATG 672
T.scotSA01         588  CGGCTCAACCGTGGAGGAGCGTGGGATACGCTCAGGCTAGAGGGTGGGAGAGGGTGGTGGAATTCCCGGAGTAGCGGTGAAATG 671

Kieft et al,1999   673  CGCAGATACCGGGAGGAACGCCGATGGCGAAGGCAGCCACCTGGTCCACTTCTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAA 756
T.scotSA01         672  CGCAGATACCGGGAGGAACGCCGATGGCGAAGGCAGCCACCTGGTCCACTTCTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAA 755

Kieft et al,1999   757  CCGGATTAGATACCCGGGTAGTCCACGCCCTAAACGATGCGCGCTAGGTCTTTGGGGTTTACCTGGGGGCCGAAGCCAACGCGT 840
T.scotSA01         756  CCGGATTAGATACCCGGGTAGTCCACGCCCTAAACGATGCGCGCTAGGTCTTTGGGGTTTACCTGGGGGCCGAAGCCAACGCGT 839

Kieft et al,1999   841  TAAGCGCGCCGCCTGGGGAGTACGGCCGCAAGGCTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGG 924
T.scotSA01         840  TAAGCGCGCCGCCTGGGGAGTACGGCCGCAAGGCTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGG 923

Kieft et al,1999   925  TTTAATTCGAAGCAACGCGAAGAACCTTACCAGGCCTTGACATGCTAGGGGACCTAGGTGAAAGCCTGGGGTACCCGCGAGGGA 1008
T.scotSA01         924  TTTAATTCGAAGCAACGCGAAGAACCTTACCAGGCCTTGACATGCTAGGGGACCTAGGTGAAAGCCTGGGGTACCCGCGAGGGA 1007
```

```
Kieft et al,1999   1009 GCCCTAGCACAGGTGCTGCATGGCCGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTGCC 1092
T.scotSA01         1008 GCCCTAGCACAGGTGCTGCATGGCCGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCTGCC 1091

Kieft et al,1999   1093 CTTAGTTGCCAGCGGGATAGGCCGGGCACTCTAAGGGGACTGCCTGCGAAAGCAGGAGGAAGGCGGGGACGACGTCTGGTCATC 1176
T.scotSA01         1092 CTTAGTTGCCAGCGGGATAGGCCGGGCACTCTAAGGGGACTGCCTGCGAAAGCAGGAGGAAGGCGGGGACGACGTCTGGTCATC 1175

Kieft et al,1999   1177 ATGGCCCTTACGGCCTGGGCGACACACGTGCTACAATGCCCACTACAGAGCGAGGCGACCCAGTGATGGGGAGCGAATCGCAAA 1260
T.scotSA01         1176 ATGGCCCTTACGGCCTGGGCGACACACGTGCTACAATGCCCACTACAGAGCGAGGCGACCCAGTGATGGGGAGCGAATCGCAAA 1259

Kieft et al,1999   1261 AAGGTGGGCGTAGTTCGGATTGGGGTCTGCAACCCGACCCCATGAAGCCGGAATCGCTAGTAATCGCGGATCAGCCATGCCGCG 1344
T.scotSA01         1260 AAGGTGGGCGTAGTTCGGATTGGGGTCTGCAACCCGACCCCATGAAGCCGGAATCGCTAGTAATCGCGGATCAGCCATGCCGCG 1343

Kieft et al,1999   1345 GTGAATACGTTCCCGGGCCTTGCACACACCGCCCGT                                                1380
T.scotSA01         1344 GTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACGCCATGGGAGCGGGTTCTACCCGAAGTCGCCGGGAGCCTTAGGGC 1427

Kieft et al,1999   1381                                                    1380
T.scotSA01         1428 AGGCGCCGAGGGTAGGGCTCGTGACTGGGGCGAAGTCGTAACAAGGTAGCC 1478
```

Fig 2.5 Alignment of the 16S rRNA sequence obtained with *Thermus scotoductus* SA-01 NCBI Accession number: AF020205 (Kieft *et al.*, 1999).

2.3.2 High-throughput GS20/FLX 454-pyrosequencing

2.3.2.1 Genomic DNA preparation

The DNA concentration of gDNA sent for the first GS20 454-pyrosequencing run was 944 ng/ul, which was sufficient for the library preparation and subsequent sequencing.

2.3.2.2 Library Construction

The pooled nebulised material was run on an Agilent BioAnalyzer DNA 1000 LabChip in order to determine the relative size and distribution of fragments generated. The graph in Fig 2.6 indicates that the nebulisation experiment was successful as fragments of the correct size and distribution were obtained.

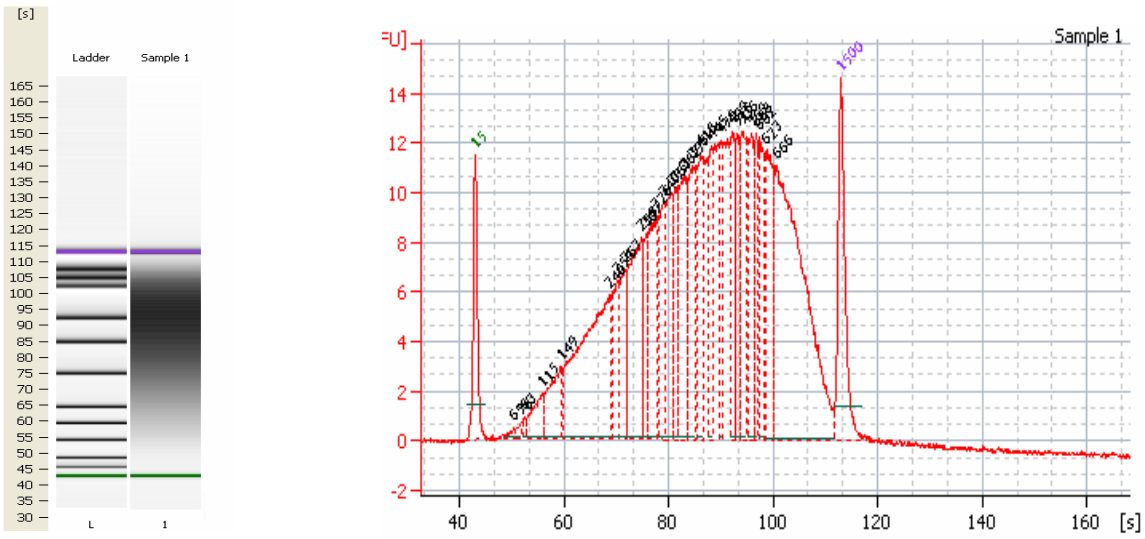
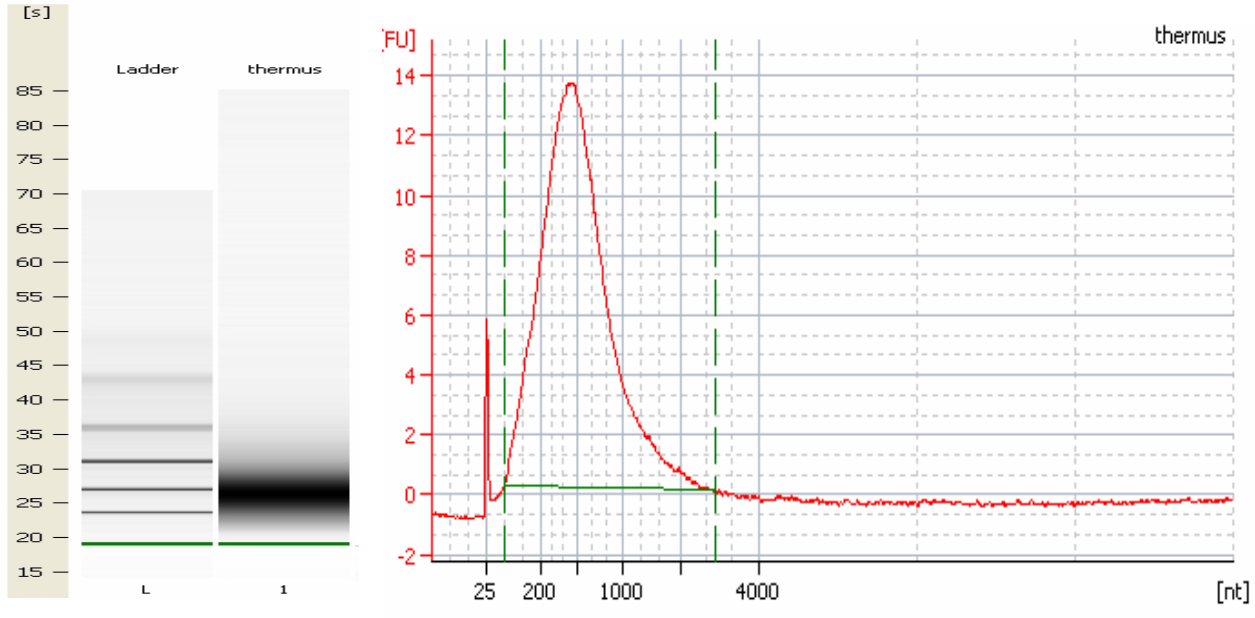
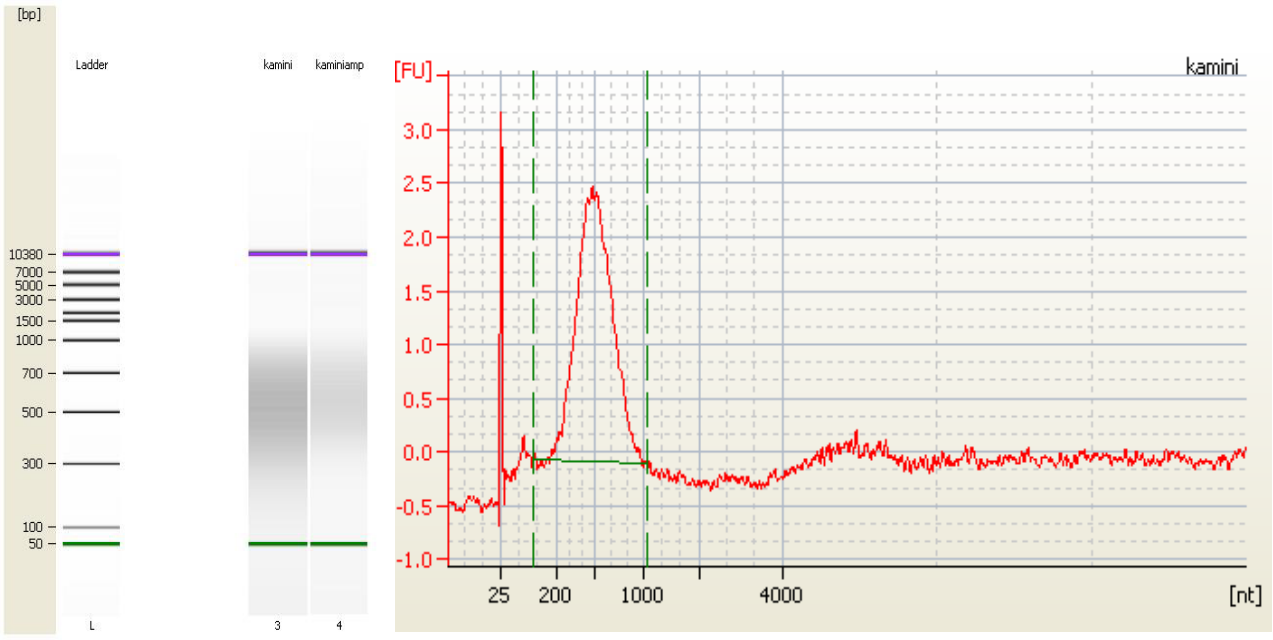


Fig 2.6 Graphical representation of the relative size distribution and yield of fragments generated after nebulization of genomic DNA.

The sstDNA library was then isolated and again run on the Agilent RNA Pico 6000 LabChip to determine the size distribution and yield. The graph in Fig 2.7 (a and b) indicates the successfully isolation of the sstDNA of an appropriate size and yield for both the GS20 and FLX run respectively. The pyrosequencing protocol was then continued with the emPCR and finally the actual pyrosequencing run.



a.



b.

Fig 2.7 Graphical representation of the relative size distribution and yield of fragments generated of a sstDNA library preparation for the GS20 (a) and FLX run (b).

Preliminary BLAST analysis of some of the contigs received after the first pyrosequencing confirmed that the data obtained was that of *T. scotoductus*, as parts of the 16S rRNA gene sequence that has a significant identity (98%) to *T. scotoductus* SA-01 (Accession no: AF020205) was obtained (Fig 2.8).

```
>  gb|AF020205.1|  Thermus sp. SA-01 16S ribosomal RNA gene, partial sequence
Length=1380

 Score =  701 bits (379),  Expect = 0.0
 Identities = 388/392 (98%), Gaps = 1/392 (0%)
 Strand=Plus/Minus

Query  1466  ACGGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGCGGCATGGCTGATCCGCGATTA  1525
              ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1380  ACGGGCGGTGTGTGCAAGGCCCGGGAACGTATTCACCGCGGCATGGCTGATCCGCGATTA  1321

Query  1526  CTAGCGATTCCGGCTTCATGGGGTCGGGTTGCAGACCCCAATCCGAACTACGCCCACCTT  1585
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1320  CTAGCGATTCCGGCTTCATGGGGTCGGGTTGCAGACCCCAATCCGAACTACGCCCACCTT  1261

Query  1586  TTTGCGATTCGCTCCCCATCACTGGGTCGCCTCGCTCTGTAGTGGGCATTGTAGCACGTG  1645
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1260  TTTGCGATTCGCTCCCCATCACTGGGTCGCCTCGCTCTGTAGTGGGCATTGTAGCACGTG  1201

Query  1646  TGTCGCCCAGGCCGTAAGGGCCATGATGACCAGACGTCGTCCCCGCCTTCCTCCTGCTTT  1705
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1200  TGTCGCCCAGGCCGTAAGGGCCATGATGACCAGACGTCGTCCCCGCCTTCCTCCTGCTTT  1141

Query  1706  CGCAGGCAGTCCCCTTAGAGTGCCCGGCCCAACCCGCTGGCAACTAAGGGCAGGGGTTGC  1765
              |||||||||||||||||||||||||||| | ||||||||||||||||||||||||||||
Sbjct  1140  CGCAGGCAGTCCCCTTAGAGTGCCCGGCCTATCCCGCTGGCAACTAAGGGCAGGGGTTGC  1081

Query  1766  GCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACGGCCATGCAGCAC  1825
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1080  GCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACGGCCATGCAGCAC  1021

Query  1826  CTGTGCTAGGGCTCCCTCGCGGGTACCC-AGG  1856
              |||||||||||||||||||||||||||| |||
Sbjct  1020  CTGTGCTAGGGCTCCCTCGCGGGTACCCCAGG  989
```

Fig 2.8 BLASTn results of initial GS20 pyrosequencing data indicating 16S rRNA region of *T. scotoductus* SA-01.

2.3.3 Assembly and Mapping of GS20/FLX data using the Newbler Assembly software

Results of the comparison of the different assembly runs indicated that there were minimal differences in the contig assembly. The differences could be due to assembly problems with repeat regions. Assembly by mapping was also attempted using the reference strain *Thermus thermophilus* HB27 which resulted in about 5% alignment with the contig assembly.

The latest FLX version of the Newbler Assembly software was then obtained and used on the SFF files generated from the first set of GS20 pyrosequencing data. The results showed a dramatic improvement in the assembly of data (any order of SFF files) as compared to previous results with the GS20 version of the software as shown in Table 2.5. From the results of the first GS20 pyrosequencing run, the initial number of contigs after assembly was 371. However, the latest assembly version reported 219 contigs. The number of large contigs dropped from 275 to 193. The largest contig size also increase from 60 069 bp to 107 670 bp and the average contig size increased from 8 420 bp to 12 008 bp.

Table 2.5 Assembly analysis of GS20 pyrosequencing data using the latest version of the Newbler assembly software.

Assembly Runs						
	Newbler Assembly Software			New FLX Software		
	Any order SFF files	Ascending Order SFF files	Descending Order SFF files	Any order SFF files	Ascending Order SFF files	Descending Order SFF files
Large Contig Metrics						
Number Of Contigs	275	273	276	193	192	194
Number of Bases	2315717	2315219	2317594	2317625	2315020	2317315
Avg Contig Size	8420	8480	8397	12008	12057	11944
N50 Contig Size	14955	14184	13838	22971	22972	22971
Largest Contig Size	60069	92552	69600	107670	107673	107666
Q40 Plus Bases	2301198, 99.4%	2300781, 99.4%	2302104, 99.3%	2257435, 97.40%	2254814, 97.40%	2257117, 97.40%
Q39 Minus Bases	14519, 0.6%	14438, 0.6%	15490, 0.7%	60190, 2.60%	60206, 2.60%	60198, 2.60%
All Contig Metrics						
Number Of Contigs	371	373	369	219	226	222
Number Of Bases	2334415	2334075	2334575	2323151	2322990	2323301

The second FLX pyrosequencing run of 50 Mbp resulted in 47.1 Mbp of sequence data.

It was then decided to use the latest Newbler assembly version on the FLX pyrosequencing data as well as combined data (GS20 and FLX) and compared these results with the original GS20 data. All assemblies were done in ascending, descending and any order of SFF files as the input data.

The results showed similar results for all assemblies when comparing the average numbers of the contig metrics for each individual pyrosequencing run and the combined results. The FLX run resulted in 42 and 40 large contigs for the ascending and descending assembly run, respectively. The assembly could not be done in any order as only two SFF files were obtained from this 47.1 Mbp run. When the combined GS20 and FLX data was assembled, 34 contigs were obtained for all 3 assemblies done. This was a dramatic decrease from an initial 192 large contigs from the original GS20 pyrosequencing run only (Table 2.6).

In order to determine if all reads were being used in the assembly of each run, data was compiled in a table with all the results from the assembly metrics obtained (Table 2.7). From the results there are a large number of reads that are not being used in the assemblies. The reads that are not used are because they have partial alignment, are singletons, repeats, duplicates or are outliers. These results indicate that the FLX Newbler Assembly program worked efficiently in obtaining the maximum assembly consensus data from the combined data set.

Table 2.6 Assembly analysis of GS20, FLX and combined pyrosequencing data using the latest version of the Newbler assembly software.

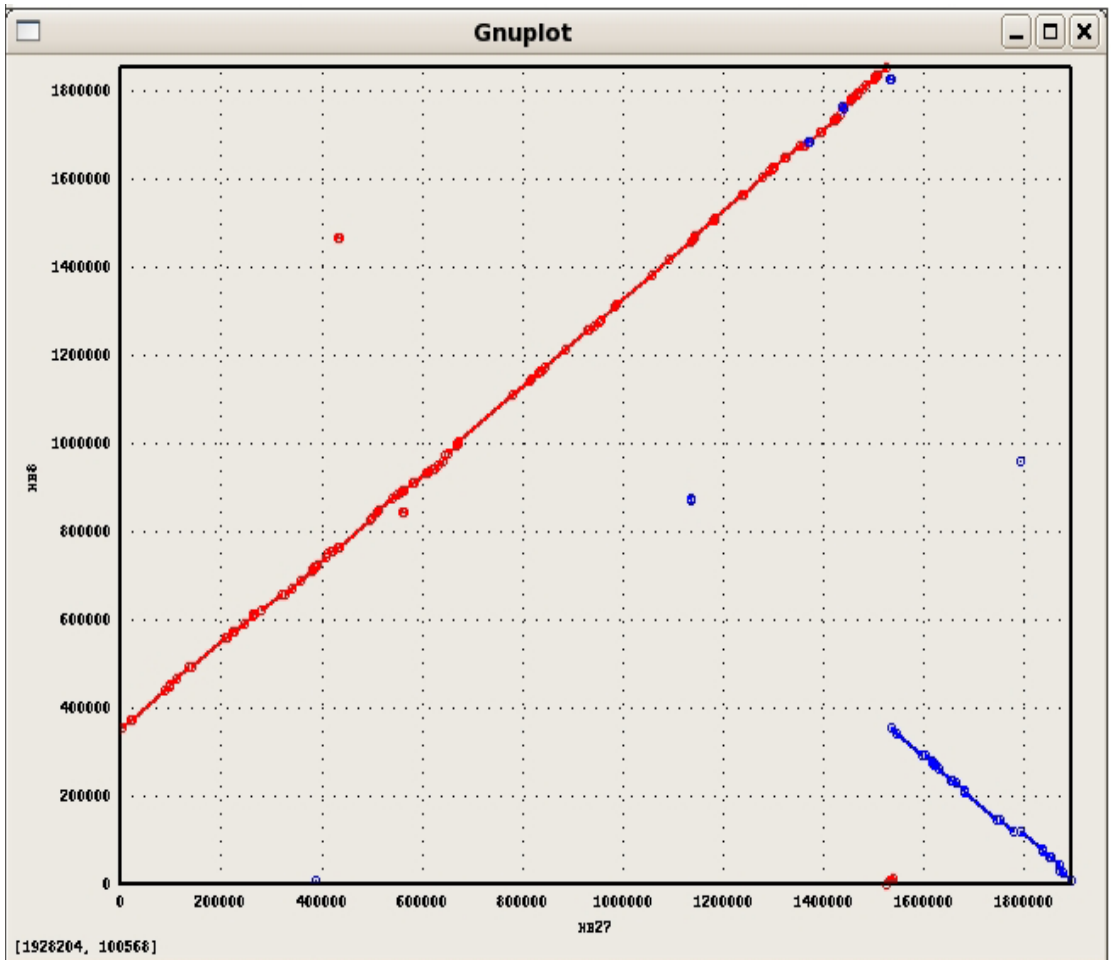
Assembly Runs								
	GS20			FLX		GS20/FLX Combined		
	Ascending	Descending	Any order	Ascending	Descending	Ascending	Descending	Any order
Large Contig Metrics								
Number Of Contigs	192	194	193	42	40	34	34	34
Number of Bases	2315020	2317315	2317625	2327773	2329447	2330128	2329901	2329997
Avg Contig Size	12057	11944	12008	55423	58236	68533	68526	68529
N50 Contig Size	22972	22971	22971	97586	115741	157753	157752	157753
Largest Contig Size	107673	107666	107670	283125	283125	283019	283144	283145
Q40 Plus Bases	2254814, 97.40%	2257117, 97.40%	2257435, 97.40%	2308666, 99.18%	2310305, 99.18%	2325573, 99.80%	2325829, 99.83%	2325952, 99.83%
Q39 Minus Bases	60206, 2.60%	60198, 2.60%	60190, 2.60%	19107, 0.82%	19142, 0.82%	4555, 0.20%	4072, 0.17%	4045, 0.17%
All Contig Metrics								
Number Of Contigs	226	222	219	57	50	54	52	53
Number Of Bases	2322990	2323301	2323151	2331332	2331807	2333296	2333024	2333009

Table 2.7 Reads used and used for the different assemblies done.

Assembly Runs (Reads Used and Unused for Assembly)								
	GS20			FLX		GS20/FLX Combined		
	Ascending	Descending	Any Order	Ascending	Descending	Ascending	Descending	Any Order
Run Metrics								
Total Number Of Reads	247 303	247 303	247 303	166 918	166 918	414 221	414 221	414 221
Total number of bases	27 718 392	27 718 392	27 718 392	41 283 319	41 283 319	69 001 711	69 001 711	69 001 711
Consensus Results : readStatus								
Number Assembled ***(Reads Used)***	211 568	211 646	211 580	147 083	147 037	361 093	361 447	361 152
Number Partial	10 472	10 422	10 499	1 976	1 946	10 530	10 491	10 474
Number Singleton	1 933	1 957	1 936	331	345	2 245	2 233	2 230
Number Repeat	731	691	685	356	436	822	529	876
Number Duplicate	21 790	21 758	21 786	16 998	16 987	38 678	38 700	38 658
Number Outlier	809	829	817	174	167	853	821	831
Read Difference ***(Reads Unused)***	35 735	35 657	35 723	19 835	19 881	53 128	52 774	53 069

2.3.4 MUMmer Analysis

By using MUMmer, the sequences that occur only once in each genome can be identified and using the complete alignment as a guide in closing the gaps between the aligned MUMs (maximal unique match) (Delcher *et al*., 1999). Here, MUMmer was used to perform a comparison on two strains of *T. thermophilus*. From the analysis, the 2 strains showed a high degree of synteny (Fig 2.9) with extremely slight differences noticed. According to Brüggemann and Chen (2006), the chromosomes are highly conserved: 94% of its genes (1860 genes) are shared in the two chromosomes, having an average amino acid identity of 97.6%. It was also found that the 2 strains create an X-shaped alignment (red and blue lines). This indicates a chromosome-scale inversion, which is a common evolutionary phenomenon, and the inversions are nearly always symmetric about the origin of replication (Delcher *et al*., 2002).

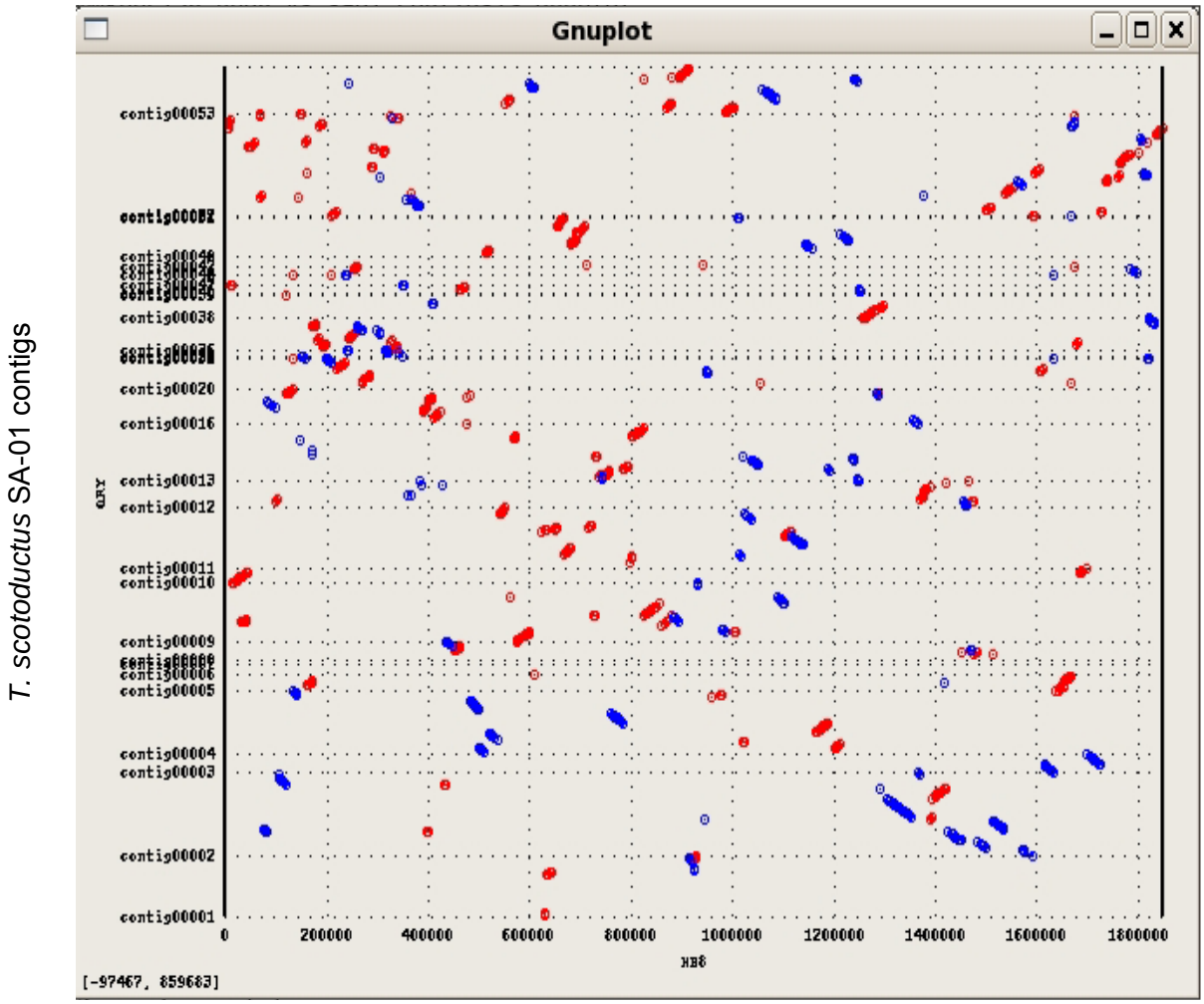


T. thermophilus HB8

T. thermophilus HB27

Fig 2.9 Genome comparison between *T. thermophilus* HB27 and *T. thermophilus* HB8 using MUMmer. Y-axis showing complete genome sequence of *T. thermophilus* HB8 and X-axis is complete genome sequence of *T. thermophilus* HB27.

The alignment of the complete genome sequence of *T. thermophilus* HB27 and the draft genome sequence of *T. scotoductus* SA-01 indicated the genome sequences are highly similar but have undergone significant or rather massive genome rearrangement (Fig 2.10). However, this very large-scale similarity, which contains many rearrangements, places additional demands on the closure of the genome. At this point, it indicated that genome closure cannot be done using a reference genome such as *T. thermophilus* and confirms why the assembly by mapping was not successful.



T. scotoductus SA-01 contigs

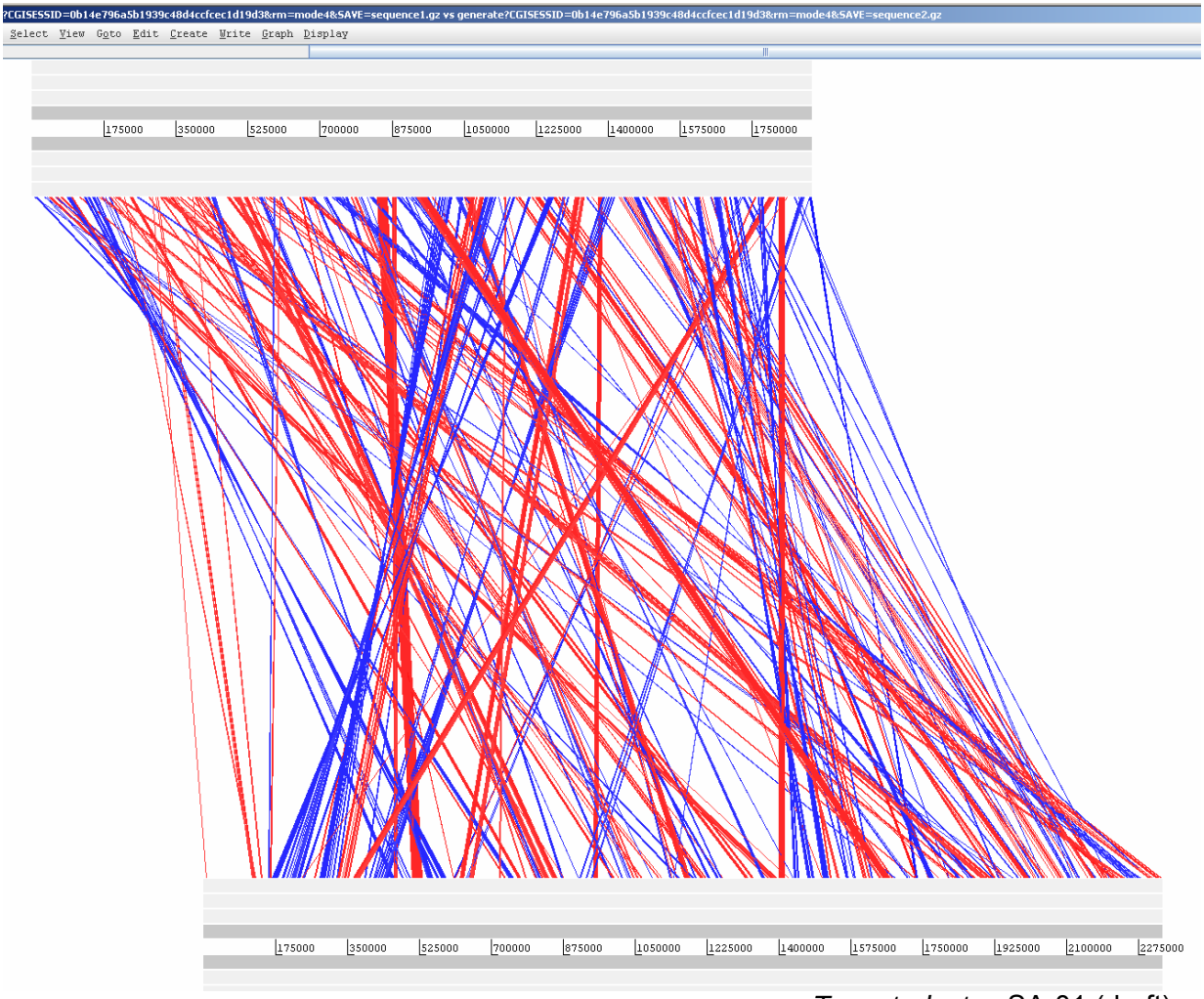
T. thermophilus HB27

Fig 2.10 Genome comparison between the complete genome sequence of *T. thermophilus* HB27 and the draft genome sequence of *T. scotoductus SA-01* using MUMmer. Y-axis showing all contigs from draft genome sequence of *T. scotoductus* SA-01 and X-axis is complete genome sequence of *T. thermophilus* HB27.

2.3.5 WebACT Mapping against *T. thermophilus* HB27

Once again the contigs were mapped with the genome sequence of *T. thermophilus* HB27 available at the NCBI database to check for regions of sequence similarity using the internet program called ACT: DNA Sequence Comparison Viewer. However, the alignment using this program showed no long regions of complete similarity, therefore mapping using a reference genome was not an option (Fig 2.11). These results once again, compared very well to the MUMmer results.

T. thermophilus HB27



T. scotoductus SA-01 (draft)

Fig 2.11 Mapping of linear DNA sequence comparison of *T. scotoductus* SA-01 contigs and *T. thermophilus* HB27 complete genome. Red blocks represent corresponding regions with a high similarity (98% or more). White spaces indicate no sequence alignment and blue indicates regions of sequences in reverse orientation.

2.3.6 Reverse-BLAST Analysis

Normal BLAST searches and annotated results were unable to pick up any plasmid associated genes present within the draft genome sequence of *T. scotoductus* SA-01. It was then decided to do a reverse-BLAST using the all the contigs from the draft genome against the complete genome sequence of *T. thermophilus* HB27 and HB8, as well as the megaplasmid (pTT27) and plasmid DNA (pTT8) sequences. Since the BLAST was narrowed down to only *T. thermophilus* sequences and not all the data present in public databases, the results revealed that there were indeed genes from the megaplasmid and plasmid of *T. thermophilus* present within the draft sequence of *T. scotoductus* SA-01. However, what was also apparent was that the plasmid associated genes were not found in a particular contig, as was thought, but are scattered among all the contigs, possibly indicating recombination of plasmid associated genes into the chromosome. Table 2.8 shows the size of the genome sequences of the two related *Thermus* strains. The approximate size of *T. scotoductus* SA-01 is 2.3 Mbp compared to the chromosome of *T. thermophilus* HB27 which is 1.8 Mbp.

Table 2.8 Comparison of the genome sizes of the completed genomes *T. thermophilus* HB27 and HB8 as well as draft genome sequence of *T. scotoductus* SA-01.

	T. thermophilus HB27		*T. thermophilus* HB8		*T. scotoductus* SA-01	
	Size	**Accession no.**	**Size**	**Accession no.**	**Size**	**Accession no.**
Draft sequence					~ 2.3 Mbp	-
Chromosome	1.8 Mbp	NC005835	1.8 Mbp	NC006461		
Megaplsmid (pTT27)	232 kbp	NC005838	256 kbp	NC006462		
Plasmid (pTT8)			9.3 kbp	NC006463		

2.3.7 Gap Closure using the Gap v4.11 Program

A successful fosmid library was obtained where >100 clones were obtained on each plate for both the *T. scotoductus* SA-01 gDNA and 1:100 dilution of the control DNA. This fosmid library was then used for the gap closure and finishing procedures. Four 96-well plates containing fosmid clones were sequenced with Sanger technology using universal forward and reverse primers, which directed sequencing from within the cloning vector into the inserts. The fosmid sequences were then added to the Contig List in Gap v4.11, which had already contained the existing assembled contigs (Fig 2.12).

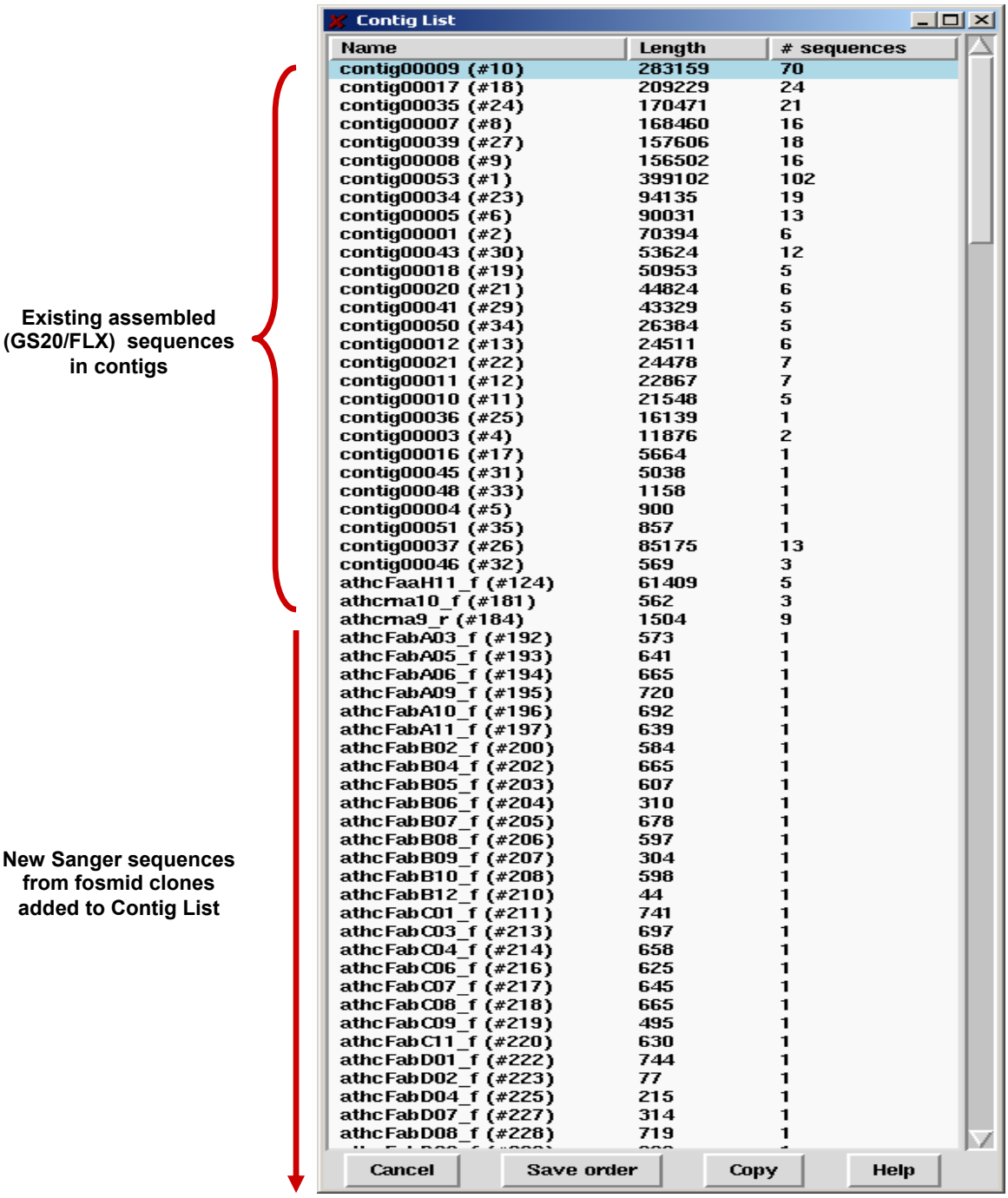


Fig 2.12 Contig list from the Gap4 software package showing all contigs and fosmid readings added into the database.

Subsequently, these sequence reads have been aligned and joined into the contigs by exploiting sequence overlaps as shown by dots in the diagram below (Fig 2.13).

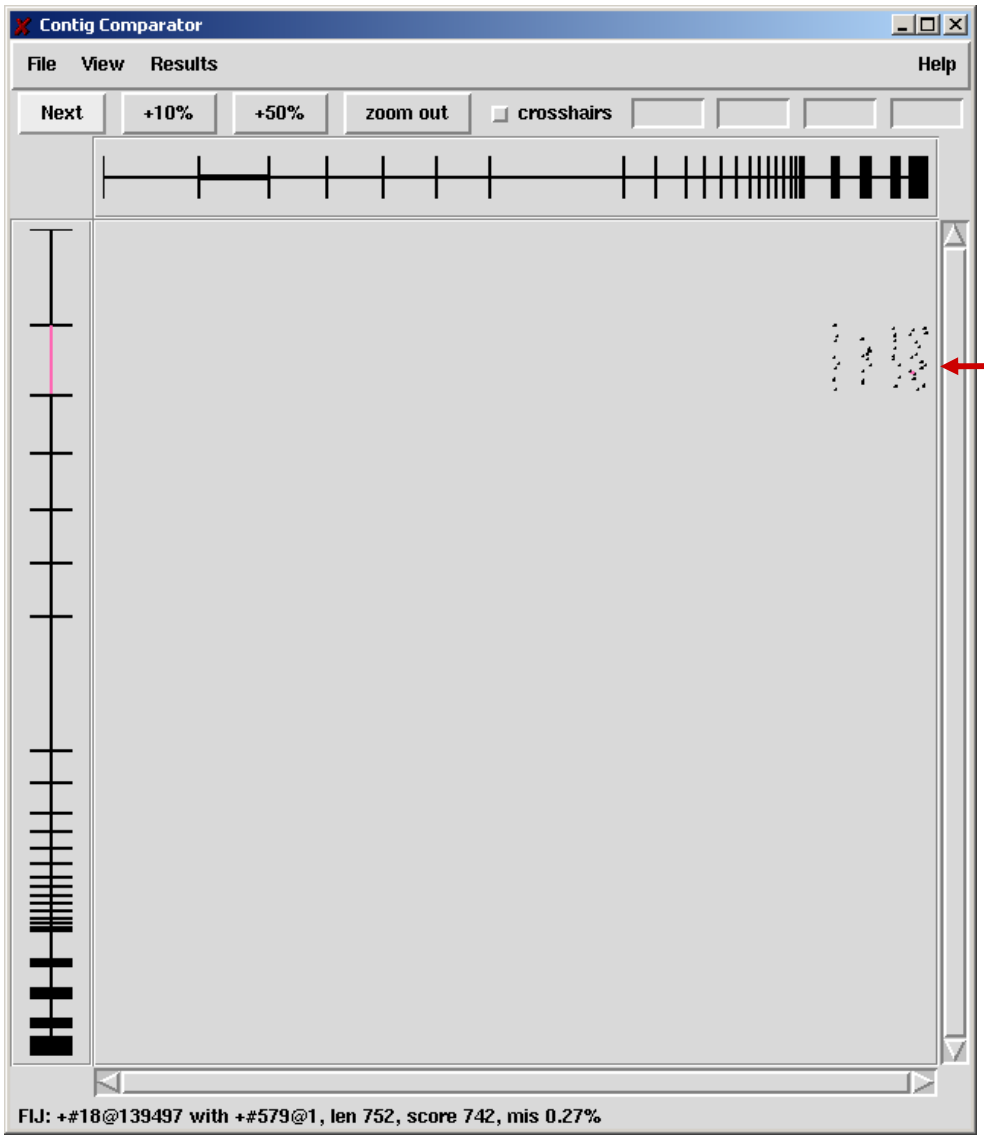


Fig 2.13 The Contig Comparator from the Gap4 software package showing all possible fosmid sequence joins to a particular existing assembled contig.

2.3.8 Joining of Fosmid Sequences

Fosmid sequences were joined to existing contigs one at a time. Initially the sequences would show many mismatches as shown in example below (Fig 2.14). However, once the program aligns the sequences completely using the Align tool, the sequence alignment can then be manually checked and edited (Fig 2.15).

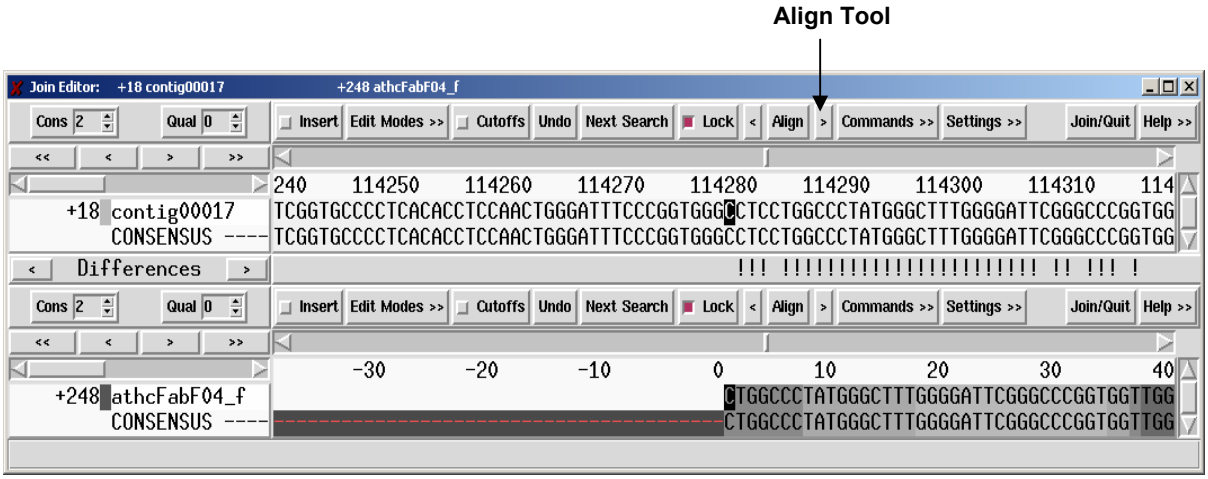


Fig 2.14 Fosmid sequences added to an existing contig before using the Align tool. Mismatches are seen by exclamation marks.

Fig 2.15 Fosmid sequences show very good alignment after using the Align tool and no exclamation marks are noticed.

2.3.9 Editing of Sequences

Prior to joining the fosmid sequences to existing contigs, some sequences needed to be edited if the quality of data is not good. The Gap4 Contig Editor is designed to allow rapid checking and editing of characters in assembled readings depending on the quality of the base call, indicated by darker shades of grey. Examples of both good and bad quality sequences are shown below. Such sequences seen in Fig 2.16 and Fig 2.17 below were first edited and then joined to contig, if sequence similarity was good. However, some poor quality sequences do occur and those were manually edited by checking the chromatogram (Fig 2.17) and then joined to sequences that show high similarity alignment.

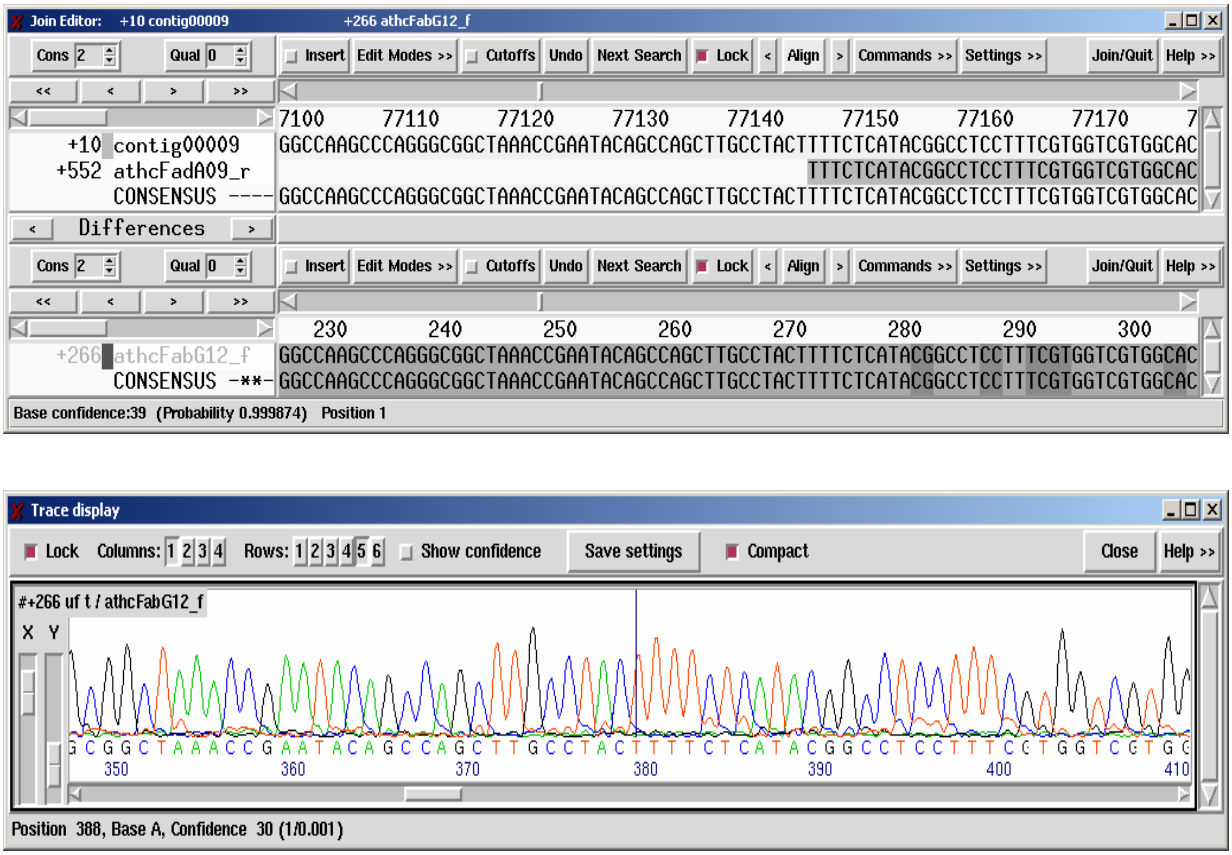


Fig 2.16 Chromatogram of sequences of fosmid clones being aligned to contigs with high quality base calling.

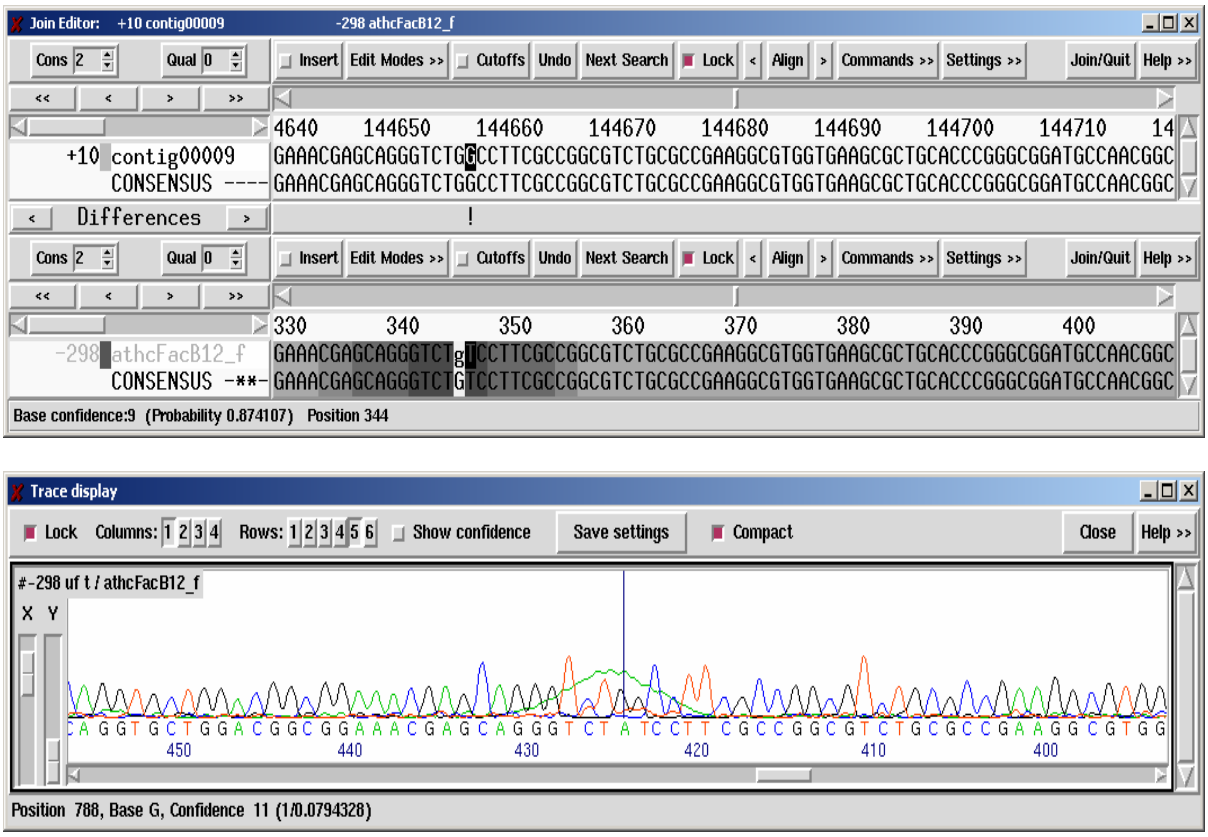


Fig 2.17 Chromatogram of sequences of fosmid clones being aligned to contigs with some errors during the sequencing reaction as well as low quality base calling indicated by darker shades of grey.

2.3.10 Gap Closure Strategies

Gap closure in sequencing projects is a major complicated, challenging and time consuming phase of the project especially in regions that contain repeat regions, high GC content and other factors that make the gaps difficult to close.

Generally, two types of gaps exist after a fosmid library is added. One gap type is spanned by one or more fosmid read pairs (at least one spanning the gap) and the second type is an unspanned gap. In the case of spanned gaps, the order and orientation of the contigs are established easily. However, unspanned gaps give no information about adjacent contigs or about the DNA spanning the gap. Some methods were chosen to close gaps i.e. using fosmid sequences that would give a clear indication of the contig order, using PCR to close the gaps and primer walking.

2.3.10.1 Gap Closure by BLASTn Analysis

In this strategy we attempted to BLASTn the ends of each contig and try to find the same hit at the end of another contig. However, this was found to be a very unreliable method of determining which contigs follow each other as BLASTn results were not accurate. Also repeat regions seemed to be problematic as 16S, 23S, 5S rRNA genes and tufB gene for Elongation factor, Tu were found in many smaller contigs (~500-3000 bp).

2.3.10.2 Gap Closure using Fosmid Library Sequences

The fosmid sequences obtained not only helped in determining the proper contig order but also helped in obtaining clones that were able to span the gap. In some cases, the sequence reads also helped in closing a few gaps between 2 contigs as shown in the Fig 2.18 below. In this particular case, there was no size gap between Contig00053 and Contig00006, however, the small letters in blue indicate the low base calling confidence and therefore the end of the contig.

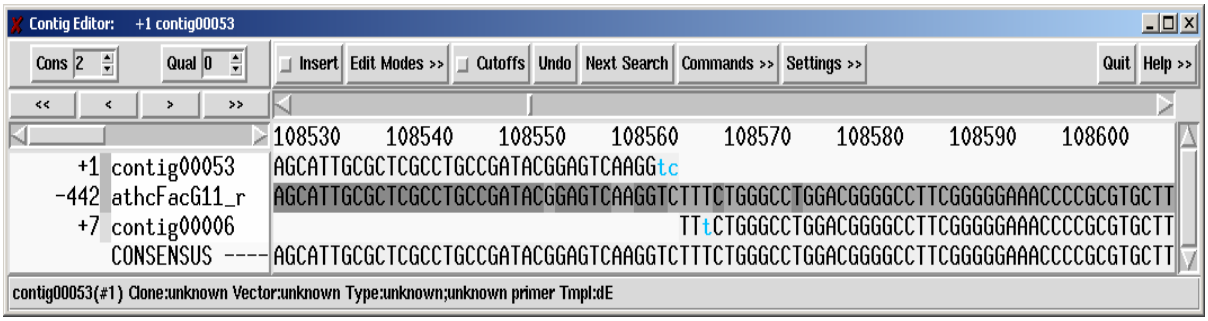


Fig 2.18 A sequence read from the end of a fosmid clone closing the 'gap' between these 2 contigs.

2.3.10.3 Gap Closure using Contig Order for PCR

The fosmid library was also able to produce linking clones that were able to produce a contig order. Fosmid clones with an approx. 40 kbp insert were sequenced at either end. Using the pairs of sequences from each clone and arranging them to point to each other, linking contigs could be determined. In this way, many supercontigs could also be obtained as shown in Fig 2.19 below. Once the order was obtained, the primers designed at the ends of all contigs were used to perform PCR amplifications and the resulting products sequenced using the ABI 3730xl DNA Analyzer (Applied Biosystems) (Fig 2.20a and 2.20b).

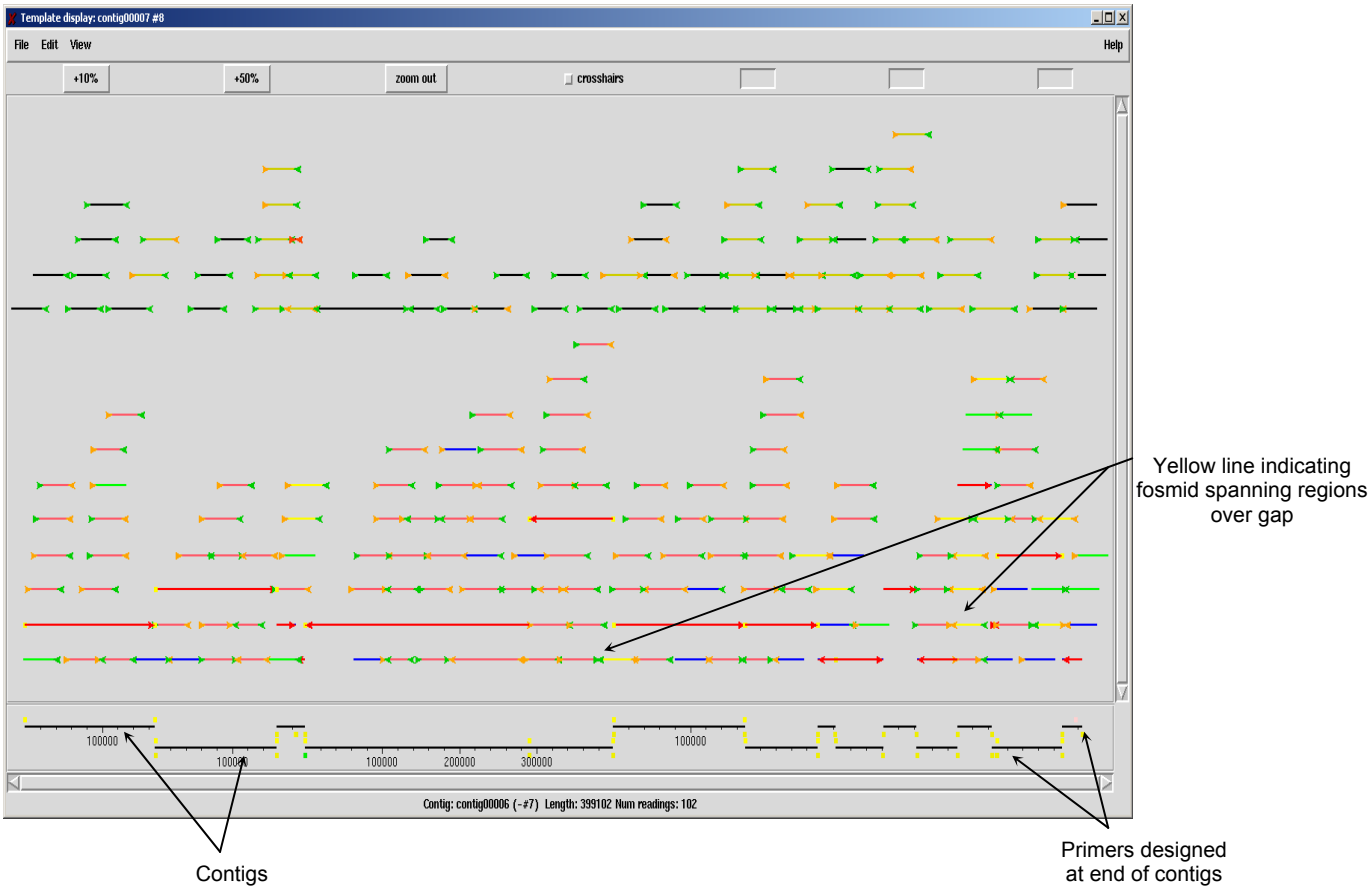


Fig 2.19 Contig order determined by fosmid spanning regions by creating a supercontig. Fosmid spanning gaps are shown by yellow lines. Primers designed are shown by yellow squares on consensus sequence.

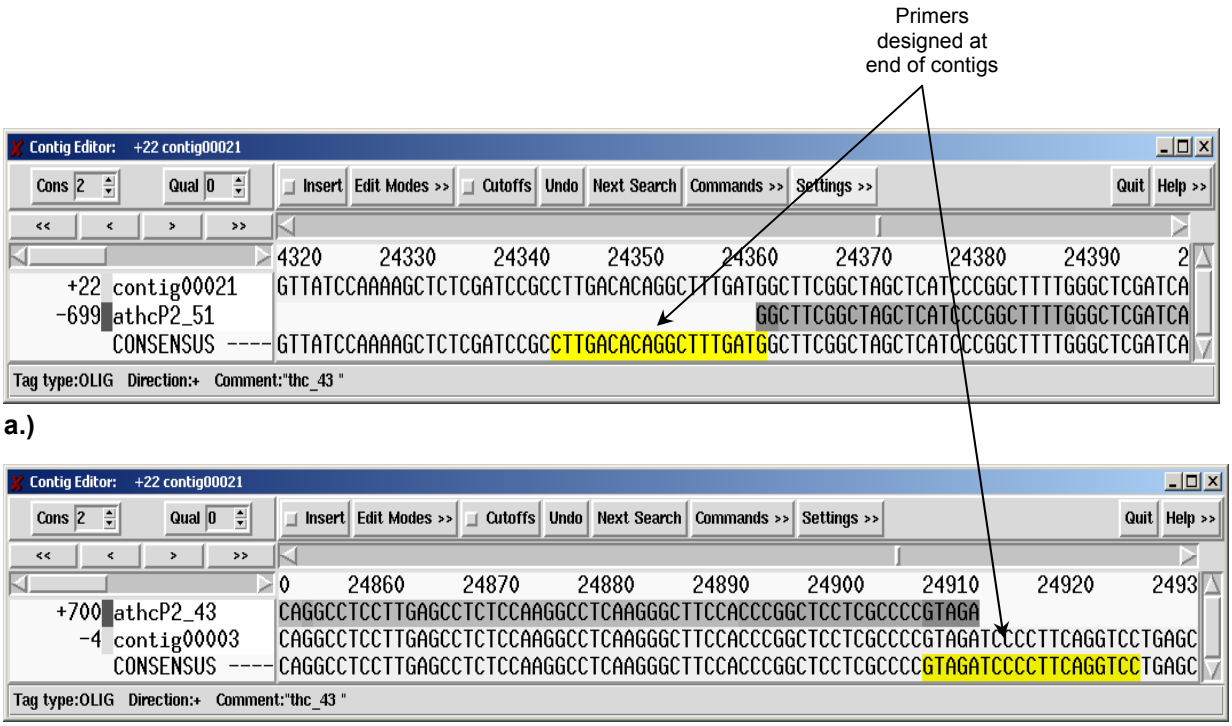


Fig 2.20 Gap closure using a sequenced PCR product obtained by using primers (highlighted in yellow) from the ends of 2 contigs that follow each other in order determined by checking fosmid that span gaps. a) PCR product sequence starting at primer from contig00021 and b) PCR product sequence beginning at primer of contig00003.

2.3.10.4 Gap Closure by Primer Walking

This strategy has to be used when the gap size is too big and cannot be closed by a single sequenced PCR product. Therefore a new internal primer is designed in order to 'walk' along the DNA sequence and needs to be repeated in order to effectively close the gap.

2.3.11 Overlaps Missed by Newbler Assembly

During the gap closure phase using fosmid and PCR sequences, it was found that in one instance, an overlap of 11 bases was not recognised by the Newbler assembly software between contig00001 and contig00018. However, the possible reason for this is probably the low confidence values with the last few sequences of each contig. Also, quite interestingly, it was shown that the last base of each contig (indicated in blue) was in fact the wrong base calling, which was the first case noticed of a wrong pyrosequence base calling considering the sequence high coverage. This particular overlap was confirmed by sequences obtained from a PCR using primers designed from the ends of both contigs (athcP10_31 and athcP10_25). A fosmid sequence (athcFabF08_f) unfortunately ended one base before the end of contig00001 (Fig 2.21).

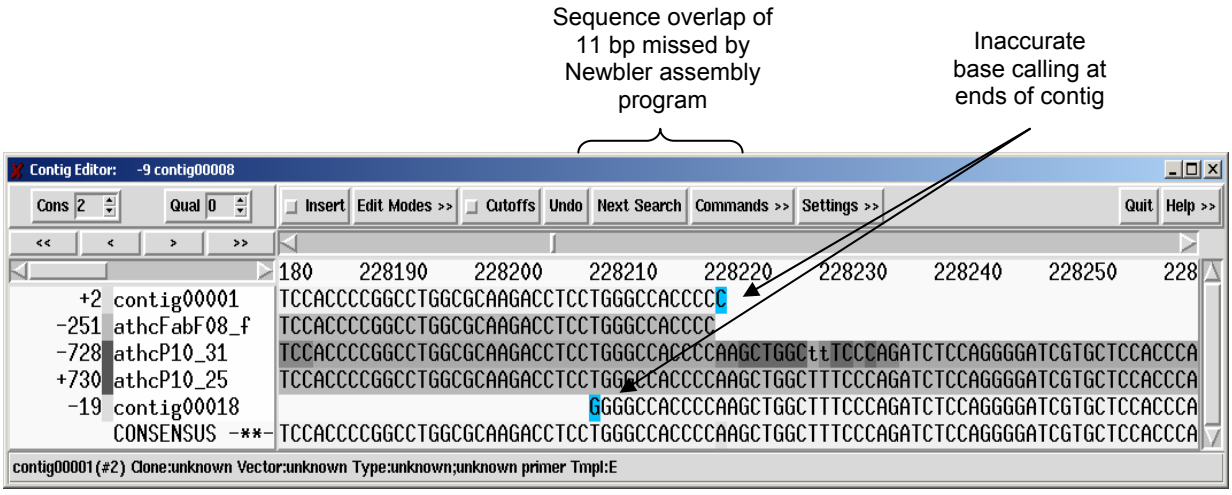


Fig 2.21 An overlap missed by the Newbler Assembly software program.

2.3.12 ORF Correction using Artemis

Possible ORFs were previously determined using automatic annotation by TIGR. In order to manually check each ORF for false positive ORFs, the data was loaded into the software program called Artemis with general features highlighted in Fig 2.22. With this software, ORFs were verified and edited manually using criteria such as the presence of a ribosome-binding site (consensus Shine-Delgarno sequence for prokaryotic organisms: 5'AGGAGG3') (Fig 2.23); codon usage analysis and checking the GC-frame plot analysis. In some cases, ORFs are added as they are missed by automated annotation (Fig 2.24). In these scenarios, the ORF sequence is BLASTed against public databases, checked for proper Shine-Delgarno sequences and if the potential ORF fits in the GC-frame plot.

Fig 2.22 Features of the Artemis program.

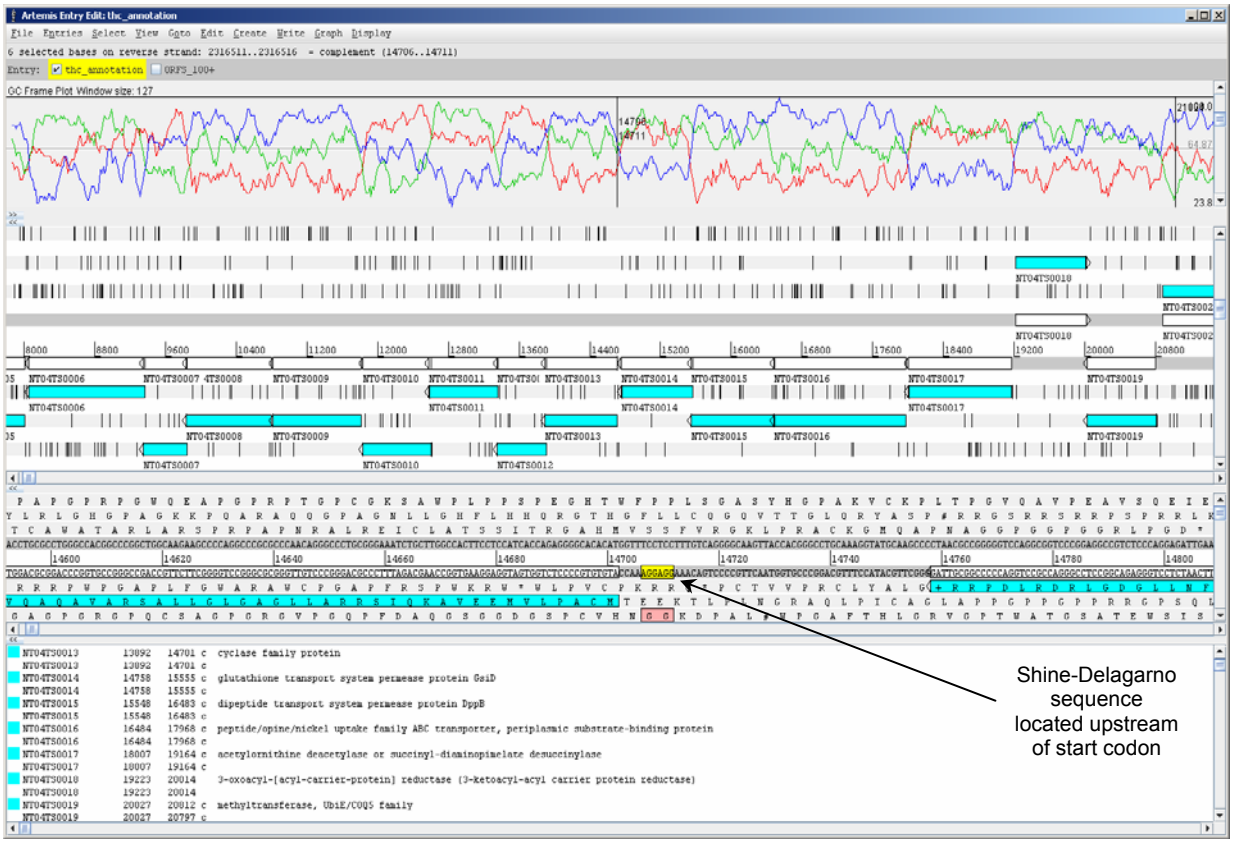


Fig 2.23 Showing the software Artemis used for ORF correction. ORFs indicated by blue boxes, Shine-Delgarno sequence highlighted in a yellow box. GC-Frame plots also used for correct start and end point of each ORF.

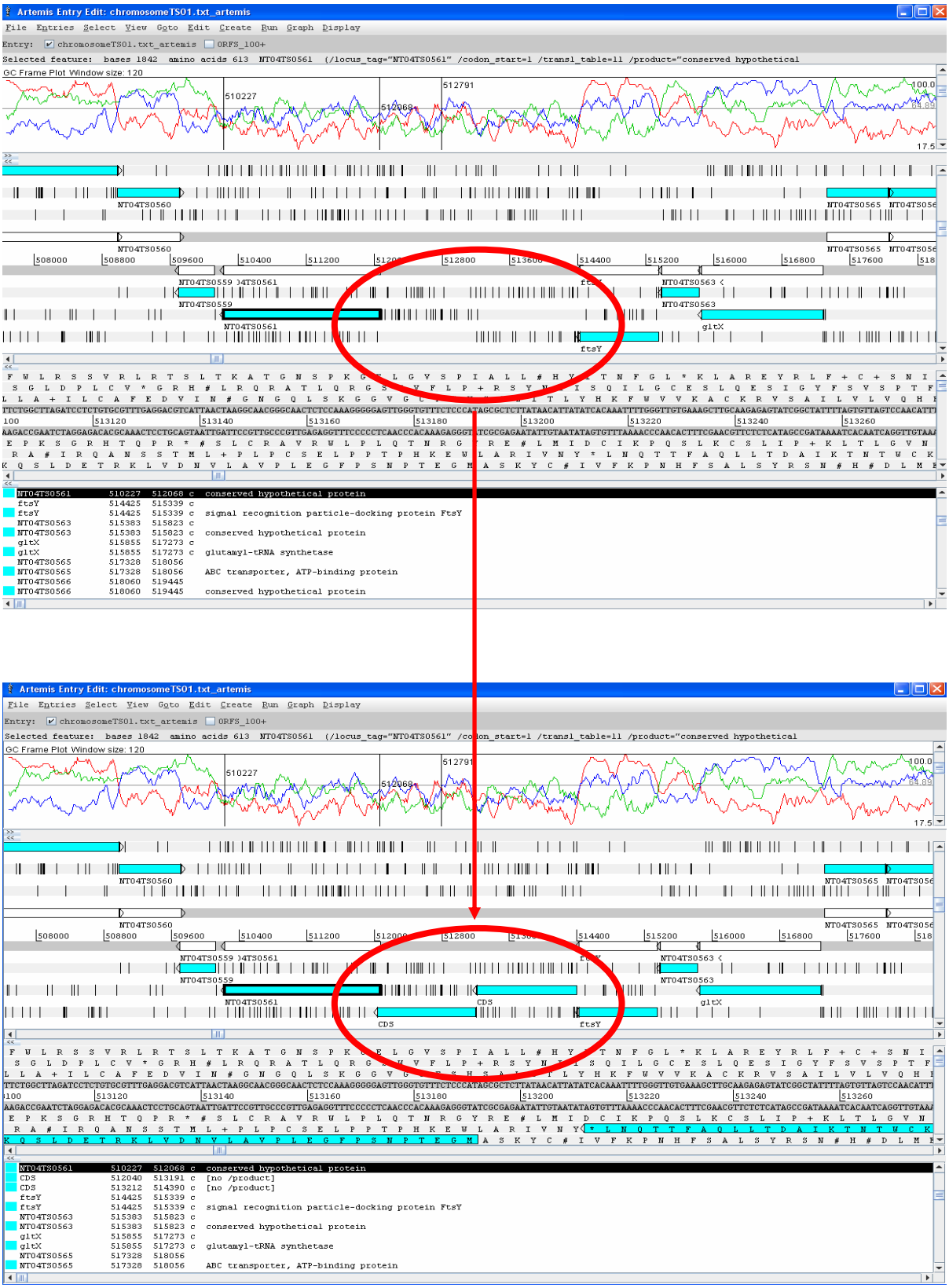


Fig 2.24 The addition of ORF's that are sometimes missed by automatic annotation.

2.3.13 Problems Working with GC-rich Organisms

The 454-pyrosequencing's lack of cloning bias and ability to sequence through regions of the genome that exhibit strong secondary structure ('hard stops') provide dramatic improvement over the quality of assemblies available from other sequencing technologies (Goldberg *et al*., 2006). These motifs are very often GC-rich sequences with high thermal and structural stability, presumably because the high duplex melting temperature permits stable secondary structures to form, thus preventing completion of a sequencing reaction or causing band compression in completed reactions (Keith *et al*., 2004). However, using Sanger sequencing alone to close gaps between contigs has proven difficult, which resulted creating 'hard-stops' in the sequence. This is probably due to the fact that *T. scotoductus* SA-01 is a high G+C organism (64.9%), and therefore there are some regions that may contain a homopolymeric G-stretch as can be seen in the figure below (Fig 2.25).

Some examples of reducing the stability of duplex DNA are the inclusion of denaturing chemicals, sulfones or dimethylsulfoxide (DMSO) and shearing of DNA into smaller pieces to disrupt the motif (Keith *et al*., 2004). It has been shown by Kieleczawa (2006), that the addition of 5% (w/v) betaine to the sequencing reaction might help in sequencing organisms containing G-stretches. In this case, different denaturation times were used for sequencing, and the addition of 5% betaine was tested. However, walking on fosmids, spanning this region, from the opposite direction eventually solved this problem.

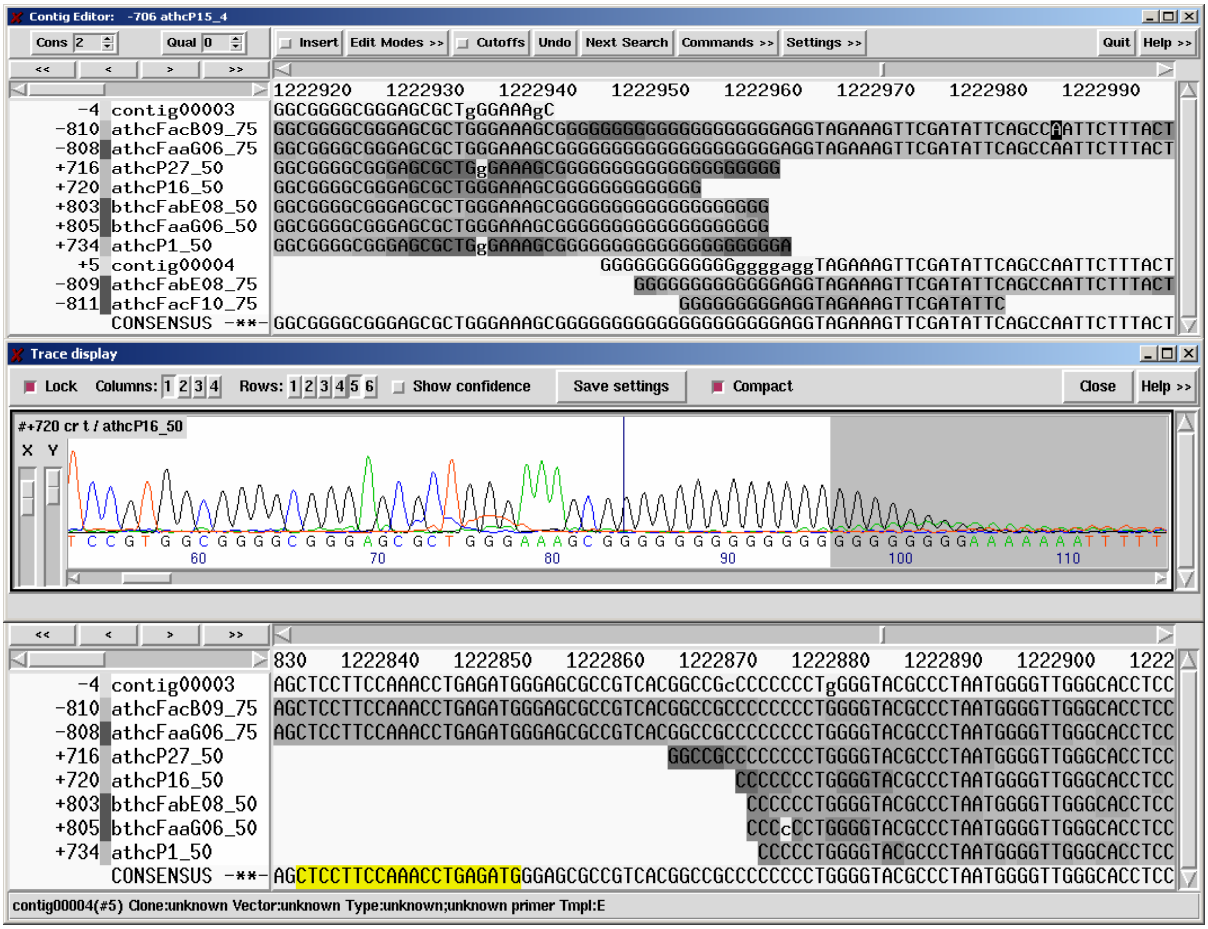


Fig 2.25 Contig editor showing sequence containing G-stretch of nucleotides.

2.3.14 16S rRNA Sequence Analysis

Sequences obtained from 10 clones of the 16S rRNA library were joined to contigs using Gap4. All sequences seemed to join a particular contig with minor base differences noticed. However, upon closer inspection, 3 sets of 16S rRNA sequences were determined and this was confirmed with other PCR sequences from the library construction as can be seen below (Fig 2.26). This compared well with the *Thermus thermophilus* strains which both contain 2 RNA clusters.

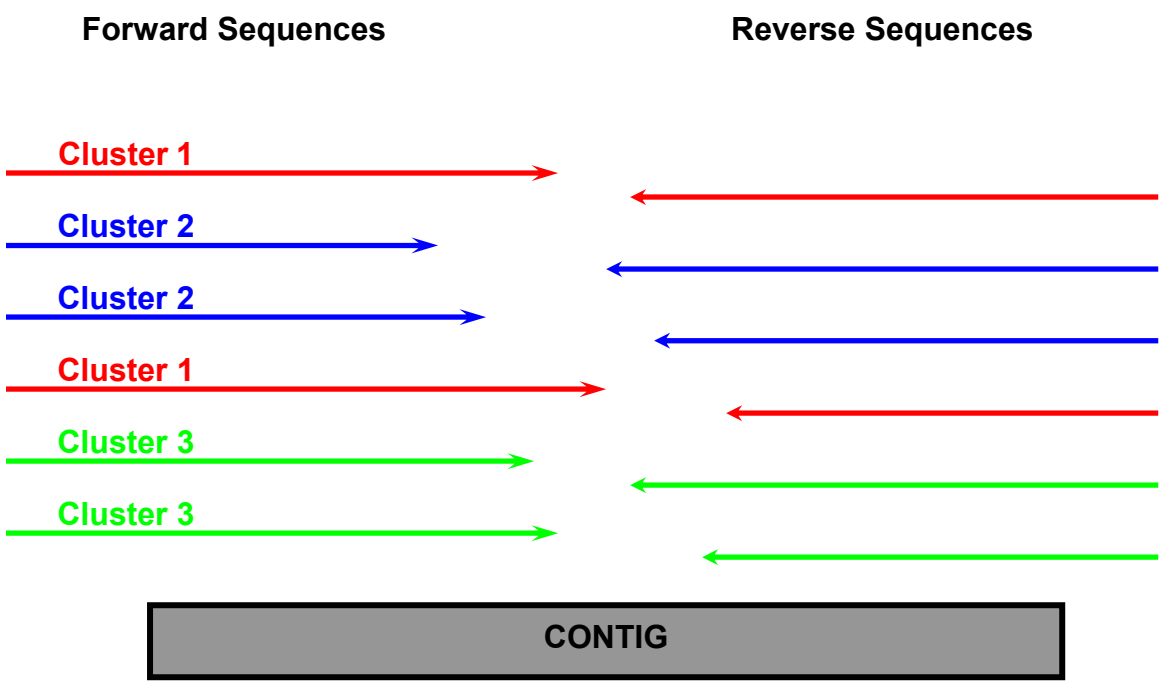


Fig 2.26 Schematic representation of the 16S rRNA sequences alignment with single base nucleotide differences. This indicates the possibility of 3 RNA clusters in the genome of *T. scotoductus* SA-01.

2.3.15 IS Search

Table 2.9 Results of IS search on genome sequence of *T. scotoductus* SA-01.

Sequences producing significant alignments	IS Family	Group	Origin	Score (bits)	E-value
ISH7B	ISNCY	-	pNRC100 of *Halobacterium* sp. NRC-1	42	0.012
ISH7A	ISNCY	-	pNRC100 of *Halobacterium* sp. NRC-1	42	0.012
ISSfl7	IS5	IS427	*Shigella flexneri*	38	0.19
ISNwi3	IS1595	ISNwi1	*Nitrobacter winogradskyi*	36	0.75
ISTth4	IS256		*Thermus thermophilus*	34	3.0
IS112	IS5	ISL2	*Streptomyces albus* G J1147	34	3.0

The results of the IS search indicated that no IS elements from any family were present in the genome sequence of *T. scotoductus* SA-01 (Table 2.9). The average length of the IS element is approximately 3 kb. The IS search however was only able to find matches with

sequences of ± 20 nucleotide sequences. However, the final genome sequence reveals the presence of transposases that are from the IS4 family which were not picked up by the IS search finder.

2.3.16 Polishing of Genome Sequence using Gap4 Confidence Value Graphs

By viewing the confidence value graphs in the Gap4 program, poor quality sequences can be checked. Lines below the 45 mark indicated regions of sequences of poor quality (Fig 2.27). These regions need to be resequenced in order to obtain a high quality base calling.

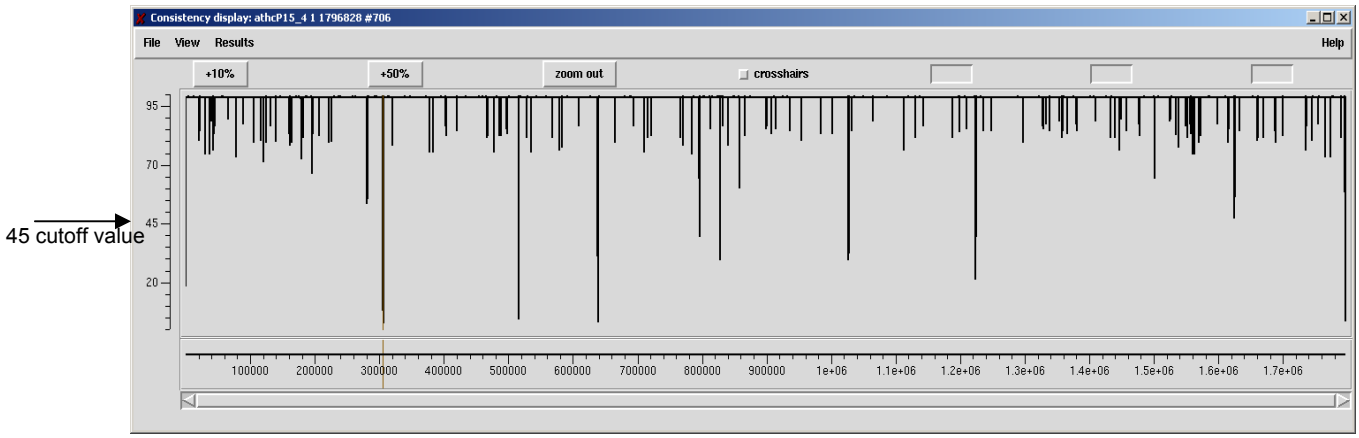


Fig 2.27 Confidence value graphs with few lines below the 45 mark, indicating regions of poor sequence quality.

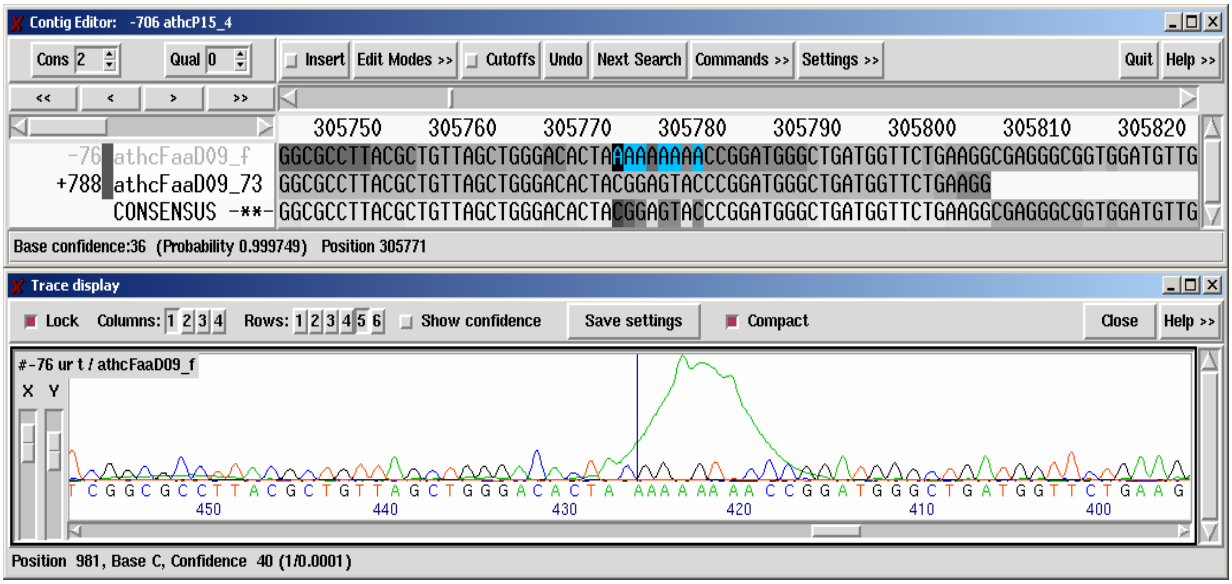


Fig 2.28 Region of poor quality that would require resequencing to improve quality.

Sequences with poor quality are indicated by a darker grey background. Mismatches are shown as blue background. In some cases as in the Fig 2.28, sequences can be manually corrected as the chromatogram shows good sequences, however, a single peak (A-stretch) causes wrong base calling. In many cases there was no need to repeat sequences, only in gap areas, as the 454-pyrosequencing had an overall good coverage as shown Fig 2.29 below.

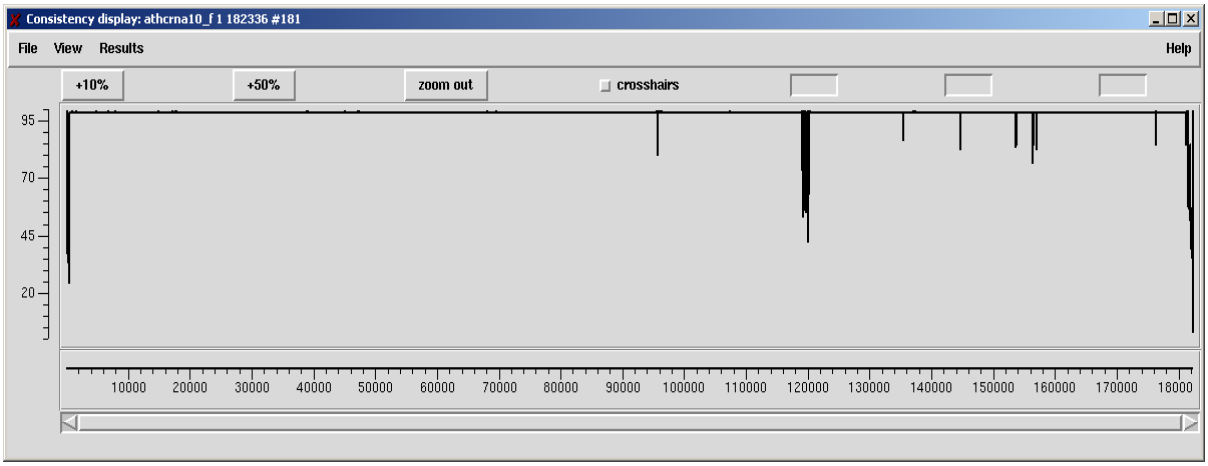


Fig 2.29 Large contig with relatively good quality sequences with little or no need for resequencing.

2.3.17 Automatic Annotation Results after GS20 and FLX Pyrosequencing

The results obtained from TIGR were viewed using the program Manatee and indicated a total of 2 458 ORFs, with not many frame shifts noticed after the 2 pyrosequencing runs. The number of ORFs decreased from 3 084 from the initial annotation results, proving that the 50 Mbp sequencing run was definitely required in order to remove frameshifts. A summary of the two annotation results after both sequencing reactions are shown Table 2.10 below.

Table 2.10 Summary of annotation results after the GS20 sequence run and after combining GS20 and FLX pyrosequencing data.

ORF Summary	ntts03 Data, GS20 (27 Mbp)		ntts04 Data, GS20+FLX (68.2 Mbp)	
Total ORFs:	**3084**	**100%**	**2458**	**100%**
Assigned function	2092	67.80%	1691	68.80%
Conserved hypothetical	453	14.70%	374	15.20%
Unknown function	271	8.80%	218	8.90%
Unclassified, no assigned role category	44	1.40%	34	1.40%
Hypothetical proteins	287	9.30%	183	7.40%

The distribution of genes in different role categories according to their assigned function is shown in Fig 2.30.

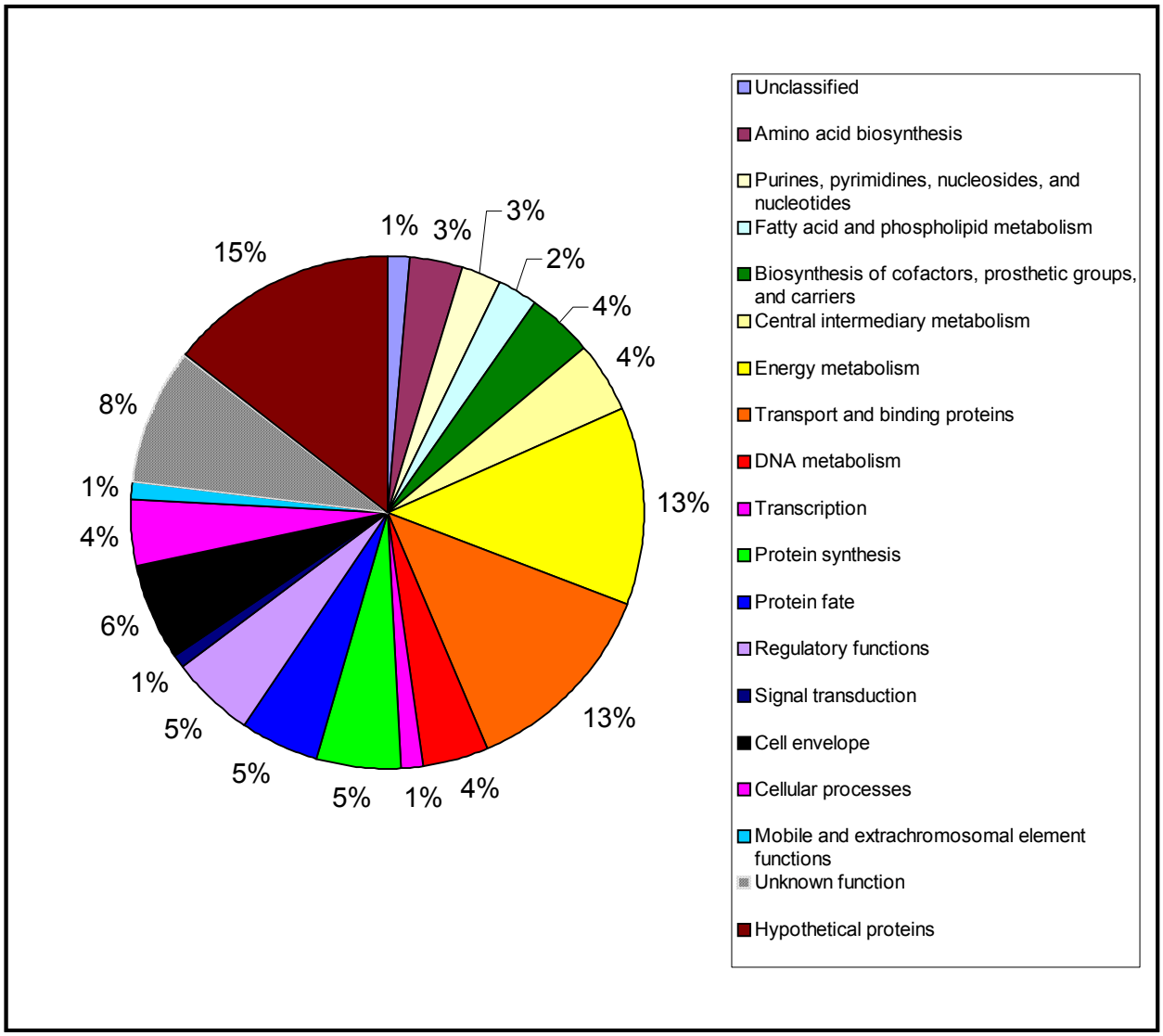


Fig 2.30 Relative percentage distribution of gene categories identified by the TIGR annotation engine after combining the GS20 and FLX sequence data.

The percentages of the various gene role categories seemed to differ slightly from previous annotation results indicating that many of the frame shifts have been eliminated with the addition of the 41.2 Mbp FLX sequence data (Table 2.11).

Table 2.11 Role category breakdown percentage differences between the GS20 and GS20+FLX pyrosequencing runs of *T. scotoductus* SA-01.

Role Breakdown	**Ntts03 Data (27 Mbp)**	**Ntts04 Data (68.2 Mbp)**
Unclassified	1.43%	1.38%
Amino acid biosynthesis	3.73%	3.58%
Purines, pyrimidines, nucleosides, and nucleotides	3.11%	2.73%
Fatty acid and phospholipid metabolism	2.69%	2.44%
Biosynthesis of cofactors, prosthetic groups, and carriers	4.31%	4.48%
Central intermediary metabolism	4.47%	4.52%
Energy metabolism	14.07%	13.18%
Transport and binding proteins	13.26%	13.30%
DNA metabolism	4.41%	4.39%
Transcription	1.56%	1.46%
Protein synthesis	5.16%	5.61%
Protein fate	4.70%	5.00%
Regulatory functions	5.25%	5.66%
Signal transduction	0.88%	0.90%
Cell envelope	6.06%	6.47%
Cellular processes	4.18%	4.15%
Mobile and extrachromosomal element functions	0.91%	1.26%
Unknown function	8.79%	8.87%
Hypothetical proteins	14.69%	15.22%
Disrupted reading frame	0.00%	0.00%
Glimmer rejects	0.00%	0.00%

From the automatic annotation results, the *T. scotoductus* SA-01 genome sequence thus far (GS20 and FLX pyrosequencing) possibly contains 2 458 protein coding genes identified and annotated by TIGR. The genome has an average G+C content of 64.9 %, which is slightly lower than the 69% G+C content of both its closest sequenced relatives, *Thermus thermophilus* HB27 and *Thermus thermophilus* HB8.

2.3.18 Manual Annotation

Each of the corrected ORFs were then manually annotated in order to check if the automatically predicted function was indeed correct (Fig 2.31). In this way, the entire genome sequence will provide large amounts of information on the character of the organism with respect to structure, function and process. This was accomplished using the ERGO Tool database containing the draft genome sequence of *T. scotoductus* SA-01, containing all the information derived from the automatic annotation results (Fig 2.32).

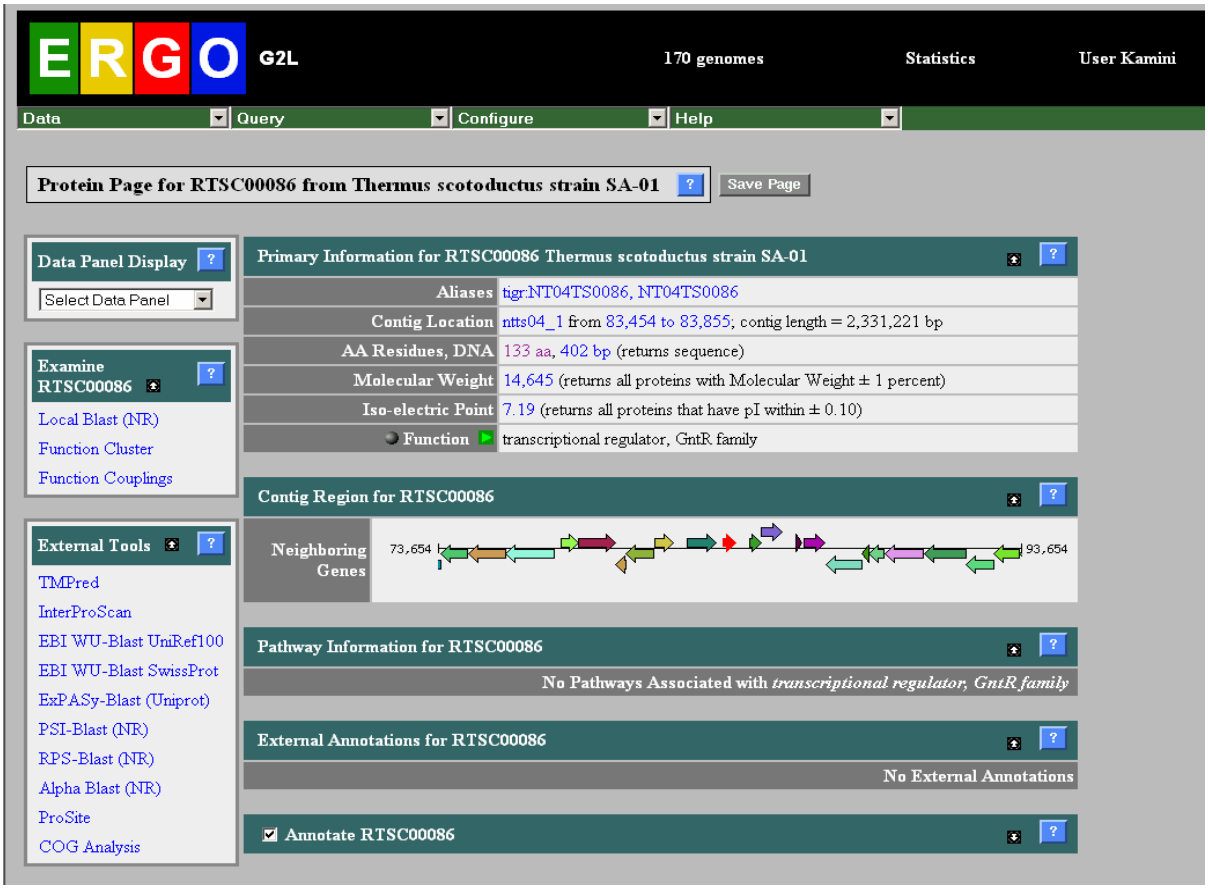


Fig 2.31 ERGO Tool database containing the automatically annotated information for each ORF.

Fig 2.32 The ERGO Tool showing the arrangement of the predicted ORFs (blue arrows) in the draft genome sequence as well as the RNA regions (red arrows).

In ERGO, for each of the automatically annotated ORFs there is the list of evidence from BLAST results for the predicted gene function as can be seen in the Fig 2.33 below.

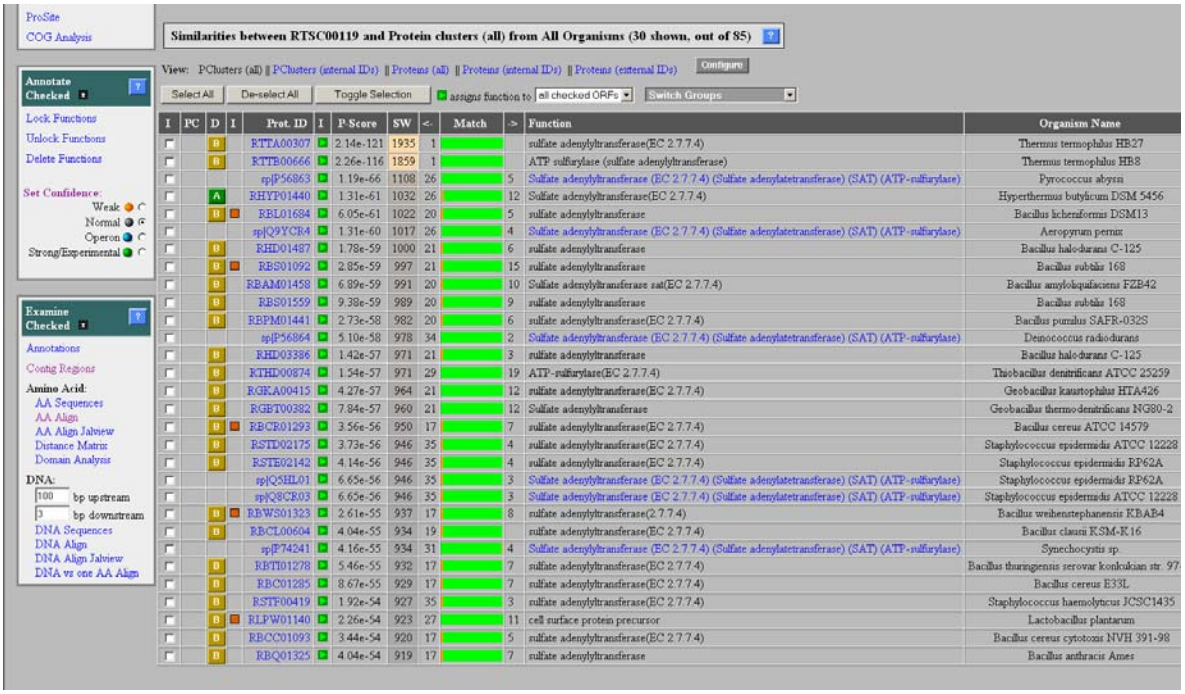


Fig 2.33 List of results from protein homology searches done using a wide variety of public databases on the individual ORF sequences.

Using the data available from the automatic annotation as well the manual protein BLAST results done, individual ORFs were annotated based on homology findings. For each ORF there were three possible results:

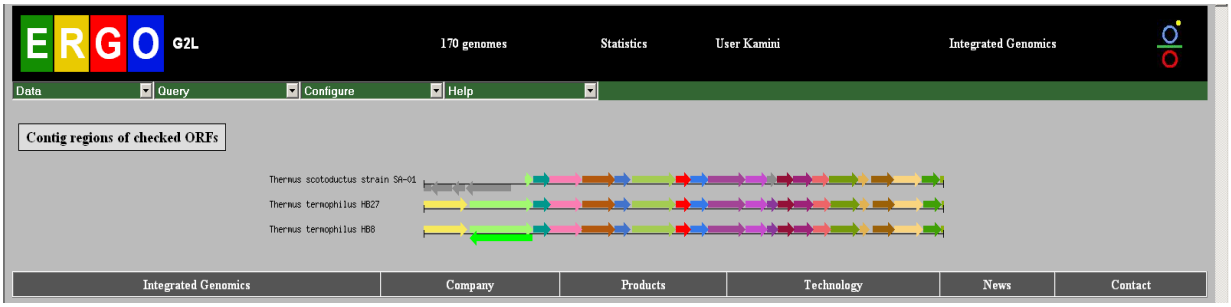
- clear sequence homology indicating function
- blocks of homology to defined functional motifs/domains (these should be confirmed experimentally)
- no significant homology or homology to proteins of unknown function

However, many ORFs could be assigned to a specific function based on homology. These ORFs were divided into two categories:

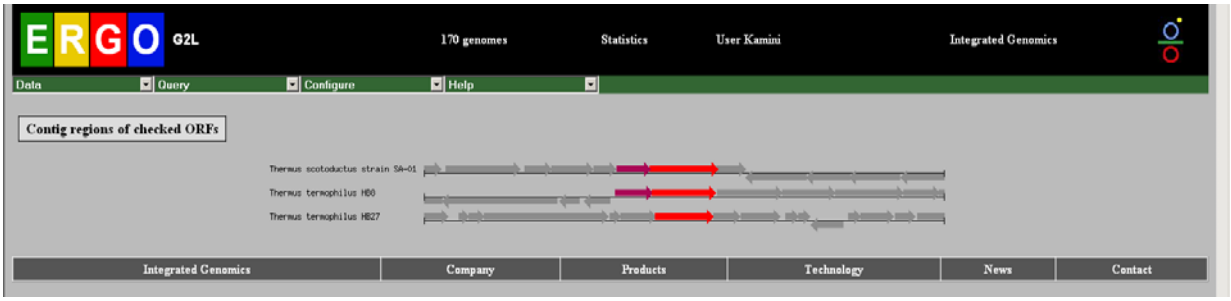
- **conserved hypothetical proteins** – ORFs with no homology to proteins of known function but with significant homology to unidentified ORFs of other species (these ORFs are therefore functionally conserved across numerous species and may represent components of central metabolism and others that have not yet been identified. The more universal the distribution of these ORFs the more likely they have a fundamental role in cellular processes).

- **hypothetical proteins**– ORFs without any homologues – these are ORFs that have no homology to any known sequences – these may represent genes related to more specific traits of the organism.

In many cases, conserved hypothetical proteins could be checked by using an alignment function on the ERGO Tool database for possible conserved regions within related organisms (Fig 2.34).



i)



ii)

Fig 2.34 Alignment of predicted ORFs to determine arrangement of ORFs when compared to other related organisms to check for conserved protein regions. i) Figure shows a highly conserved region of sequences with the *Thermus* species as compared to ii) sequences containing a genome area of a very low conservation of genes.

2.3.19 The *T. scotoductus* SA-01 complete chromosome sequence

2.3.19.1 General Features

The genome sequence of *T. scotoductus* SA-01 was completed and could be compared to other completely sequenced genomes. The genome was sequenced with an approx. 20-fold coverage. The 35 contigs obtained from a combination of GS20/FLX pyrosequence runs were assembled with additional Sanger sequencing to close the gaps between the contigs. The genome of *T. scotoductus* SA-01 is composed of a 2 346 803 bp chromosome (Fig 2.35) and a draft plasmid sequence is currently at 8 383 bp and the average G+C content is 65.9%, which was in good agreement with that of the *T. scotoductus* SA-01 chromosome of 64.9%. The general properties of the sequenced *T. scotoductus* SA-01 are shown in Table 2.12.

Table 2.12 General features of the *Thermus scotoductus* SA-01 genome.

Feature	Chromosome (complete)	Plasmid (draft)
Genome size	2 346 803 bp	8383 bp
G+C Content	64.9%	65.9%
Predicted protein coding genes (CDS or ORF)	2464 (+26, -8)	12
Average CDS length, bp	894	646
Longest CDS length, bp	8085	1140
Percent of genome protein coding (%)	93.90	92.50
Transfer RNAs	48	-
Ribosomal RNA	4	-

Based on GC skew analysis, the origin of replication was identified which was the location of the characteristic replication protein DnaA. Automatic annotation predicted 2 464 genes for the chromosome. These ORFs have been manually checked and corrected using Artemis. From the correction, 8 ORFs were deleted based on completely overlapping genes and extremely short ORF size. All 8 ORFs deleted were determined to be hypothetical proteins, as no substantial similarity was found to entries in the public databases. In addition, 26 ORFs were added to the genome, by checking the predicted GC-frame plot and filled in ORFs within the coding sequences.

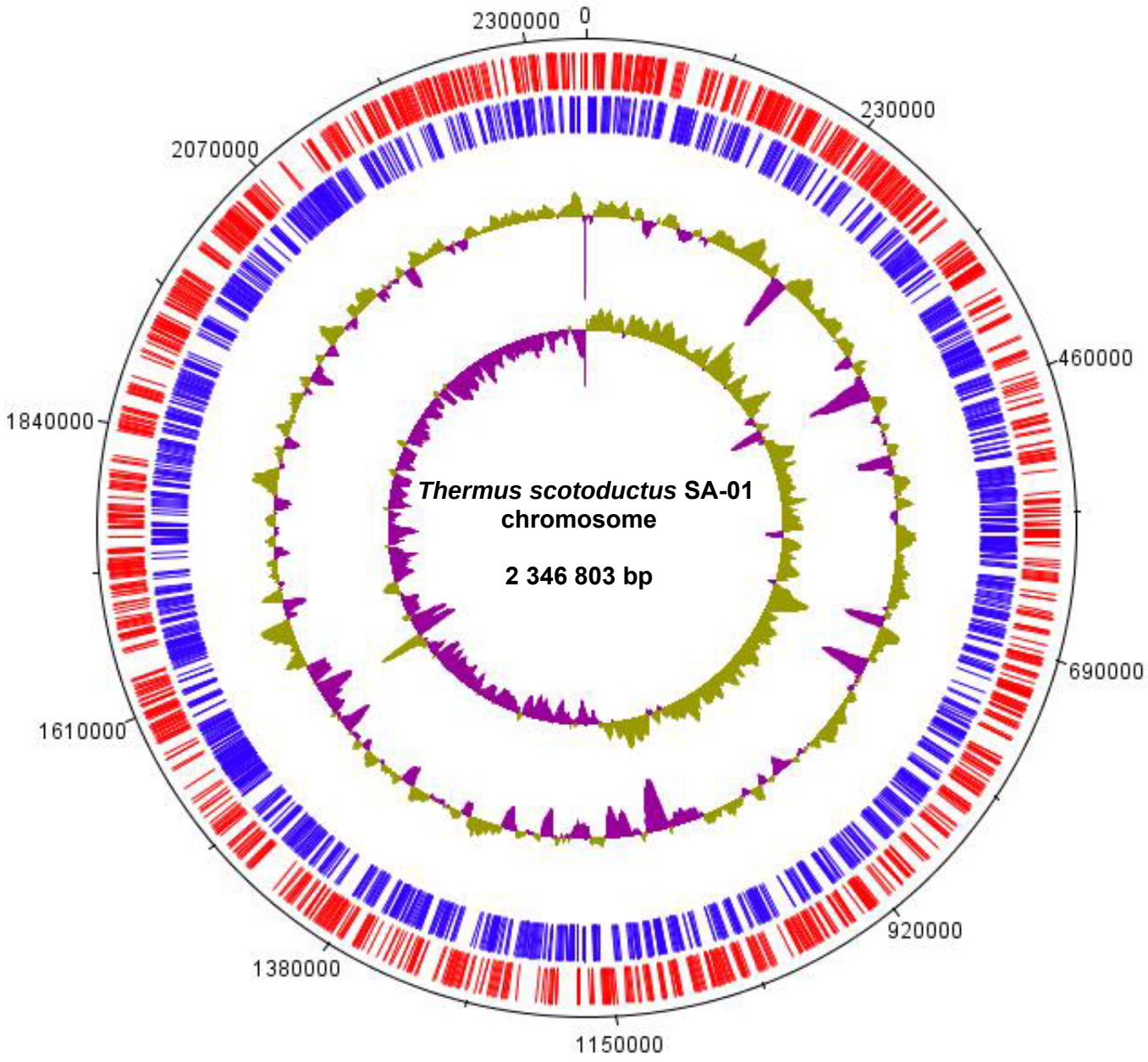


Fig 2.35 Map of the *T. scotoductus* SA-01 chromosome. Circle drawn using DNAPlotter (Carver *et al*., 2009). The protein coding sequence of the chromosome is shown in red and blue, depending on the strand orientation. The outermost circle represents the scale in bp, the 1st inner circle shows the G+C content variation and the 2nd innermost circle represents the GC skew analysis.

2.3.20 Automatic Annotation of Chromosome

The final results obtained from TIGR (viewed using the Manatee program), revealed the respective percentage of genes present in the different annotated categories i.e. the genes that a had an assigned function (the highest at 68.5%), proteins of unknown function at 9.3%, the conserved hypothetical and hypothetical proteins accounting for 15.5% and 7.1 %, respectively and 1% of all proteins from the genome of *T. scotoductus* SA-01 could not be assigned to any gene role category (Fig 2.36). According to Lioliou *et al* (2004), examination of all sequenced genomes reveals that almost 40% of each genome remains as hypothetical proteins. However, within the automatic annotation of the *T. scotoductus* SA-01 genome, the hypothetical proteins account for only 7%. This indicates that, over time, studies have revealed the function and submitted to public databases the many proteins that were previously identified as having an unknown function.

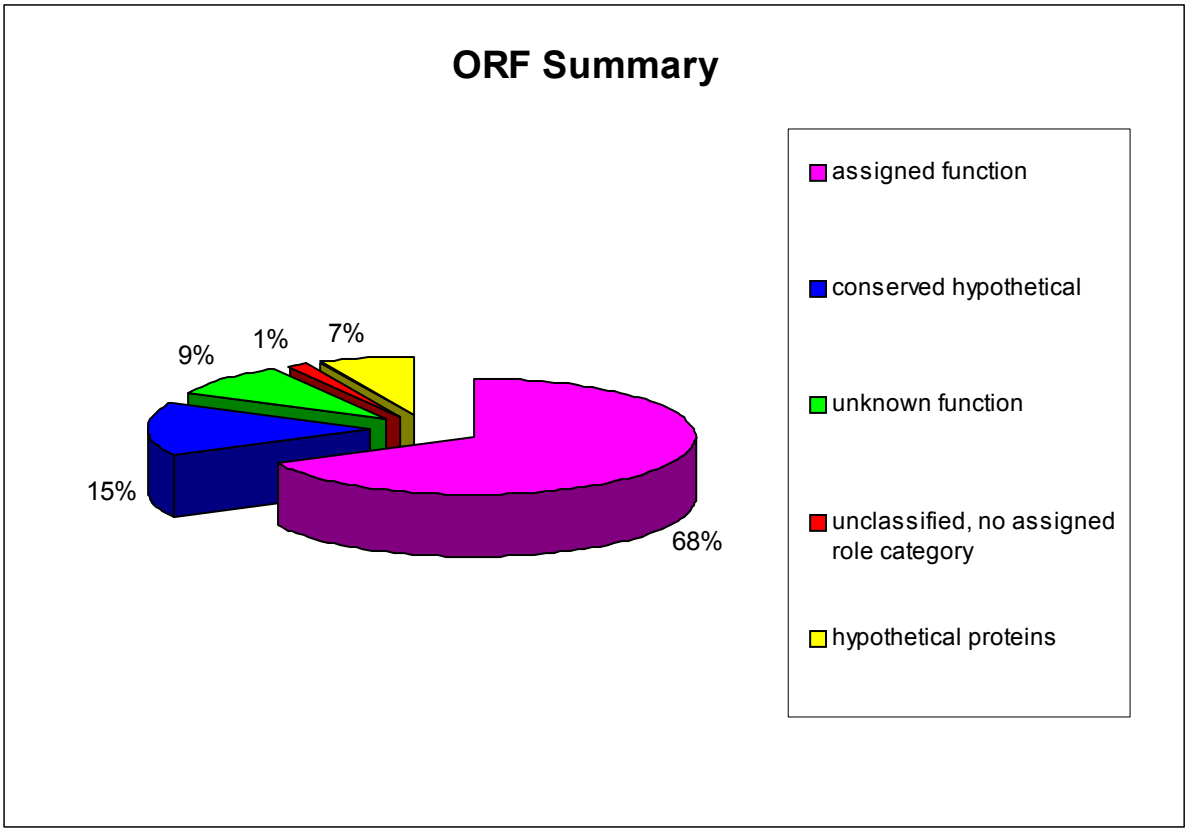


Fig 2.36 Functional classification of the complete *T. scotoductus* SA-01 chromosome ORFs.

2.3.21 Draft Plasmid Sequence (pTS01)

Unfortunately, the complete plasmid sequence was not obtained. There is still one gap left which, after several attempts to close the gap, remained. Numerous PCRs were performed, walking on fosmid sequences and subsequent sequencing did not result in a sequence, which could be used to close the gap or continue walking.

However, the existing draft plasmid sequence is 8 383 bp and 12 ORFs within the plasmid sequence were successfully identified (Fig 2.37). In comparison to *T. thermophilus* HB8, the pTT8 is 9.3 kbp in size. If the size of the plasmid in *T. scotoductus* SA-01 is approximately the same size, then the gap missing is ~ 1 kbp. Interestingly, the plasmid sequence also contains the chromate reductase gene identified by Opperman and van Heerden (2008). However, due to automatic and manual annotations, this gene has been referred to as the NADPH dehydrogenase (xenobiotic reductase) in the plasmid sequence. In a later publication, the authors (Opperman *et al*., 2008) revealed the relatedness of the enzyme to the Old Yellow enzyme family, in particular the xenobiotic reductases, which are involved in the oxidative stress response.

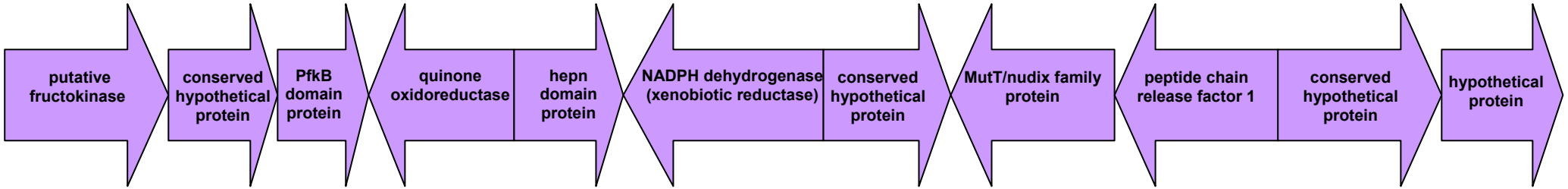


Fig 2.37 Linear representation of the ORFs present on the pTS01 draft sequence.

Table 2.13 BLAST results of plasmid sequence (pTS01) against complete chromosome sequence.

Gene present on plasmid	BLAST hit	% Similarity	E value	ORF size	ORF number on chromosome
putative fructokinase	putative fructokinase	100	e-131	272	NT04TS0050
conserved hypothetical protein	conserved hypothetical protein	100	2e-99	184	NT04TS1995
	conserved hypothetical protein	100	2e-99	184	NT04TS0820
	conserved hypothetical protein	100	2e-99	184	NT04TS0210
	conserved hypothetical protein	98	4e-98	184	NT04TS1692
	conserved hypothetical protein	98	4e-98	184	NT04TS1427
	conserved hypothetical protein	98	4e-98	184	NT04TS0520
	conserved hypothetical protein	96	5e-96	184	NT04TS0302
conserved hypothetical protein	conserved hypothetical protein	100	3e-95	171	NT04TS1690
	conserved hypothetical protein	100	3e-95	171	NT04TS0518
	conserved hypothetical protein	99	8e-95	171	NT04TS1428
	conserved hypothetical protein	99	8e-95	171	NT04TS0819
	conserved hypothetical protein	99	8e-95	171	NT04TS0211
	conserved hypothetical protein	99	8e-95	171	NT04TS0052
	conserved hypothetical protein	99	3e-94	171	NT04TS1996
	conserved hypothetical protein	95	1e-91	171	NT04TS0301
conserved hypothetical protein	conserved hypothetical protein	100	7e-43	87	NT04TS0051

The BLASTp results of the draft plasmid sequence (pTS01) against the complete chromosome sequence indicated that four ORFs present on the draft plasmid are also present as an identical copy (one or more than one copy) on the *T. scotoductus* SA-01 chromosome, providing evidence of genetic exchange between the chromosome and the extrachromosomal element. The four genes are the putative fructokinase and 3 conserved hypothetical proteins. The protein BLAST results are shown in Table 2.13. The BLASTp also shows that two conserved hypothetical proteins from the draft plasmid have seven and eight copies respectively, with ORFs on the chromosome, of very high similarity. Interestingly, the corresponding ORF BLASTp results of the plasmid found on the chromosome show that most of the ORFs that the plasmid genes hit with are found adjacent to each other on the chromosome. If the ORF numbers on chromosome of the corresponding BLASTp hit are put in numerical order, we can see the order of genes on the chromosome (Fig 2.38). This indicates that genes were mobilized on the plasmid from sets/clusters of genes found on the chromosome.

Fig 2.38 Representation of sets of ORFs found on the chromosome mobilised randomly into draft plasmid sequence. Each set indicates ORFs found adjacent to each other on the chromosome.

2.3.22 Complete genome comparisons

MUMmer plots and WebACT were used once again to compare the now completed chromosome sequence of *T. scotoductus* SA-01 with the reference *T. thermophilus* strains. The WebACT alignments (Fig 2.39) and MUMmer plots [Fig 2.40 (a) and (b)] agreed that the re-arrangements noticed previously with the draft genome sequence of *T. scotoductus* SA-01 (contigs data) against *T. thermophilus* HB27 no longer are apparent. However, now that the correct sequence order in terms of the chromosome has been elucidated, genes inversions are more apparent with the genome comparisons. The X-alignment noticed is a common evolutionary feature when comparing phylogenetically related organisms. However, from this later comparison, it has become more apparent now that genome comparisons are more accurate when complete sequence data is used.

In addition, with the MUMmer plots, a nucleotide (nucmer) and protein (promer) BLAST was performed against *T. thermophilus* HB27. Analysing both data sets, specific differences can be noticed with the apparent loss of certain genic regions from the protein BLAST comparison.

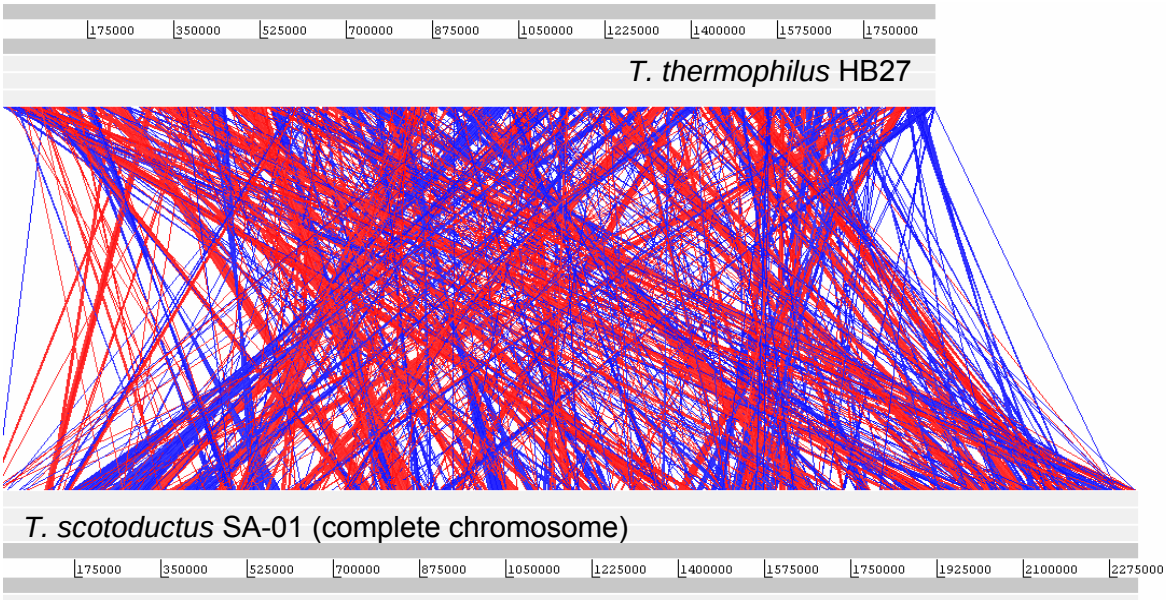


Fig 2.39 Alignment of the complete chromosome sequence of *T. scotoductus* SA-01 against *T. thermophilus* HB27 using the WebACT program.

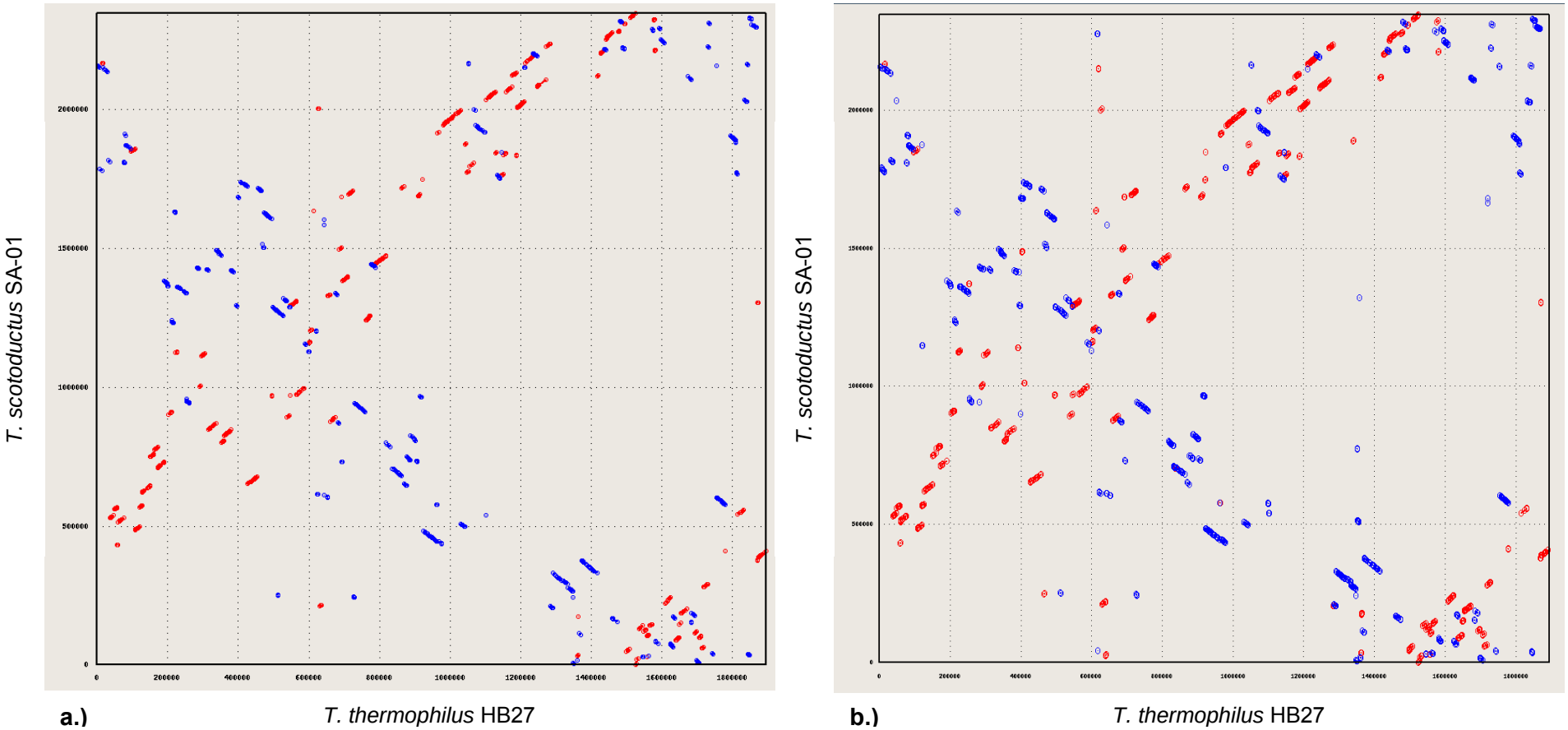
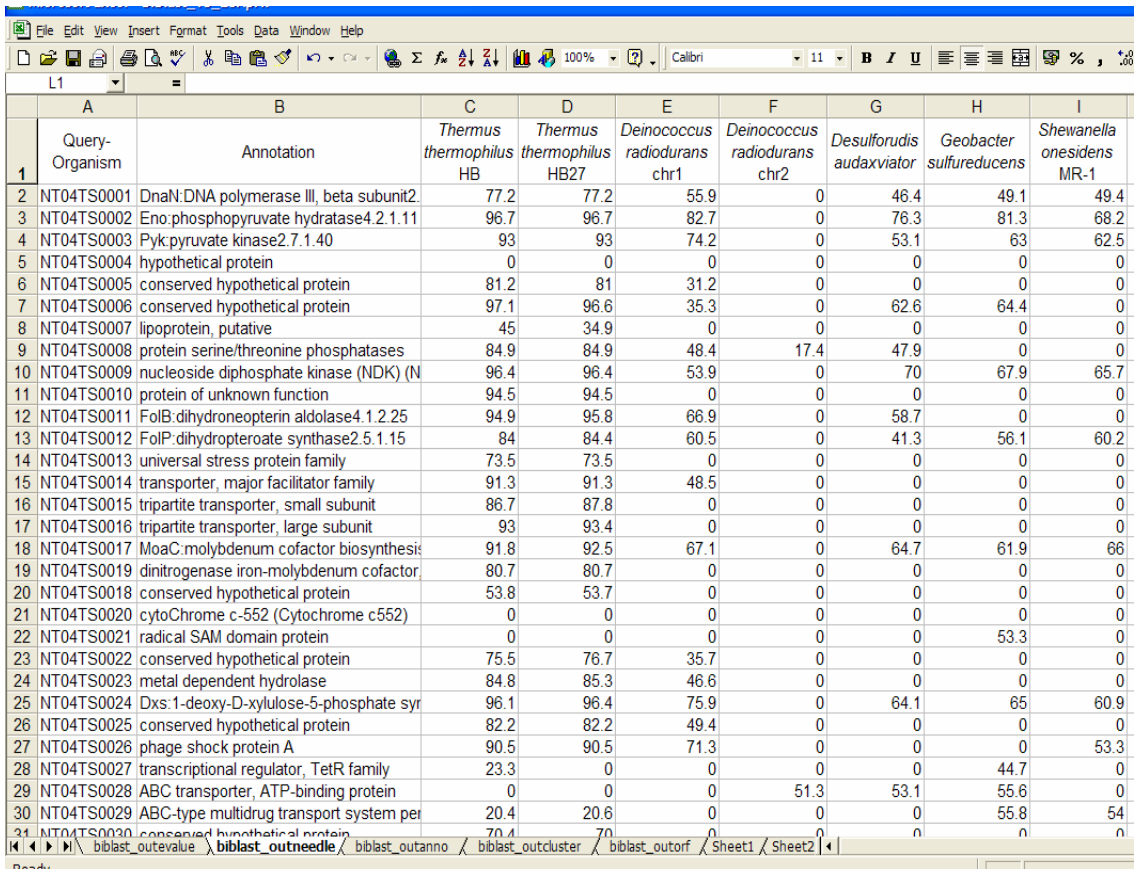


Fig 2.40 Genome comparison between *T. scotoductus* SA-01 and *T. thermophilus* HB27 using MUMmer. Y-axis showing complete genome sequence of *T. scotoductus* SA-01 and X-axis is complete genome sequence of *T. thermophilus* HB27. (a.) Alignment performed using the Nucmer and (b.) Promer BLAST.

2.3.23 Bi-directional BLAST

The tab-separated tables containing the e-values (Fig 2.41), represent the Needleman-Wunsch similarities generated for *T. scotoductus* SA-01 compared to other complete genomes of choice, together with the colour coded similarities (Fig 2.42). The first column represents the chromosomal localisation of query organism genes and each following column belongs to one organism and contains the values to the corresponding best bi-BLAST hits.



	A	B	C	D	E	F	G	H	I
1	Query-Organism	Annotation	*Thermus thermophilus* HB	*Thermus thermophilus* HB27	*Deinococcus radiodurans* chr1	*Deinococcus radiodurans* chr2	*Desulforudis audaxviator*	*Geobacter sulfureducens*	*Shewanella onesidens* MR-1
2	NT04TS0001	DnaN:DNA polymerase III, beta subunit2.	77.2	77.2	55.9	0	46.4	49.1	49.4
3	NT04TS0002	Eno:phosphopyruvate hydratase4.2.1.11	96.7	96.7	82.7	0	76.3	81.3	68.2
4	NT04TS0003	Pyk:pyruvate kinase2.7.1.40	93	93	74.2	0	53.1	63	62.5
5	NT04TS0004	hypothetical protein	0	0	0	0	0	0	0
6	NT04TS0005	conserved hypothetical protein	81.2	81	31.2	0	0	0	0
7	NT04TS0006	conserved hypothetical protein	97.1	96.6	35.3	0	62.6	64.4	0
8	NT04TS0007	lipoprotein, putative	45	34.9	0	0	0	0	0
9	NT04TS0008	protein serine/threonine phosphatases	84.9	84.9	48.4	17.4	47.9	0	0
10	NT04TS0009	nucleoside diphosphate kinase (NDK) (N	96.4	96.4	53.9	0	70	67.9	65.7
11	NT04TS0010	protein of unknown function	94.5	94.5	0	0	0	0	0
12	NT04TS0011	FolB:dihydroneopterin aldolase4.1.2.25	94.9	95.8	66.9	0	58.7	0	0
13	NT04TS0012	FolP:dihydropteroate synthase2.5.1.15	84	84.4	60.5	0	41.3	56.1	60.2
14	NT04TS0013	universal stress protein family	73.5	73.5	0	0	0	0	0
15	NT04TS0014	transporter, major facilitator family	91.3	91.3	48.5	0	0	0	0
16	NT04TS0015	tripartite transporter, small subunit	86.7	87.8	0	0	0	0	0
17	NT04TS0016	tripartite transporter, large subunit	93	93.4	0	0	0	0	0
18	NT04TS0017	MoaC:molybdenum cofactor biosynthesis	91.8	92.5	67.1	0	64.7	61.9	66
19	NT04TS0019	dinitrogenase iron-molybdenum cofactor	80.7	80.7	0	0	0	0	0
20	NT04TS0018	conserved hypothetical protein	53.8	53.7	0	0	0	0	0
21	NT04TS0020	cytoChrome c-552 (Cytochrome c552)	0	0	0	0	0	0	0
22	NT04TS0021	radical SAM domain protein	0	0	0	0	0	53.3	0
23	NT04TS0022	conserved hypothetical protein	75.5	76.7	35.7	0	0	0	0
24	NT04TS0023	metal dependent hydrolase	84.8	85.3	46.6	0	0	0	0
25	NT04TS0024	Dxs:1-deoxy-D-xylulose-5-phosphate syr	96.1	96.4	75.9	0	64.1	65	60.9
26	NT04TS0025	conserved hypothetical protein	82.2	82.2	49.4	0	0	0	0
27	NT04TS0026	phage shock protein A	90.5	90.5	71.3	0	0	0	53.3
28	NT04TS0027	transcriptional regulator, TetR family	23.3	0	0	0	0	44.7	0
29	NT04TS0028	ABC transporter, ATP-binding protein	0	0	0	51.3	53.1	55.6	0
30	NT04TS0029	ABC-type multidrug transport system per	20.4	20.6	0	0	0	55.8	54
31	NT04TS0030	conserved hypothetical protein	70.4	70	0	0	0	0	0

Fig 2.41 Excel sheet showing part of the results of a bi-BLAST containing the e-value representing the Needleman-Wunsch similarities generated of *T. scotoductus* SA-01 against *Thermus thermophilus* HB27, *Thermus thermophilus* HB8, *Deinococcus radiodurans, Desulforidis auduxviator*, *Shewanella oneidensis* MR-1 and *Geobacter sulfurreducens* PCA.

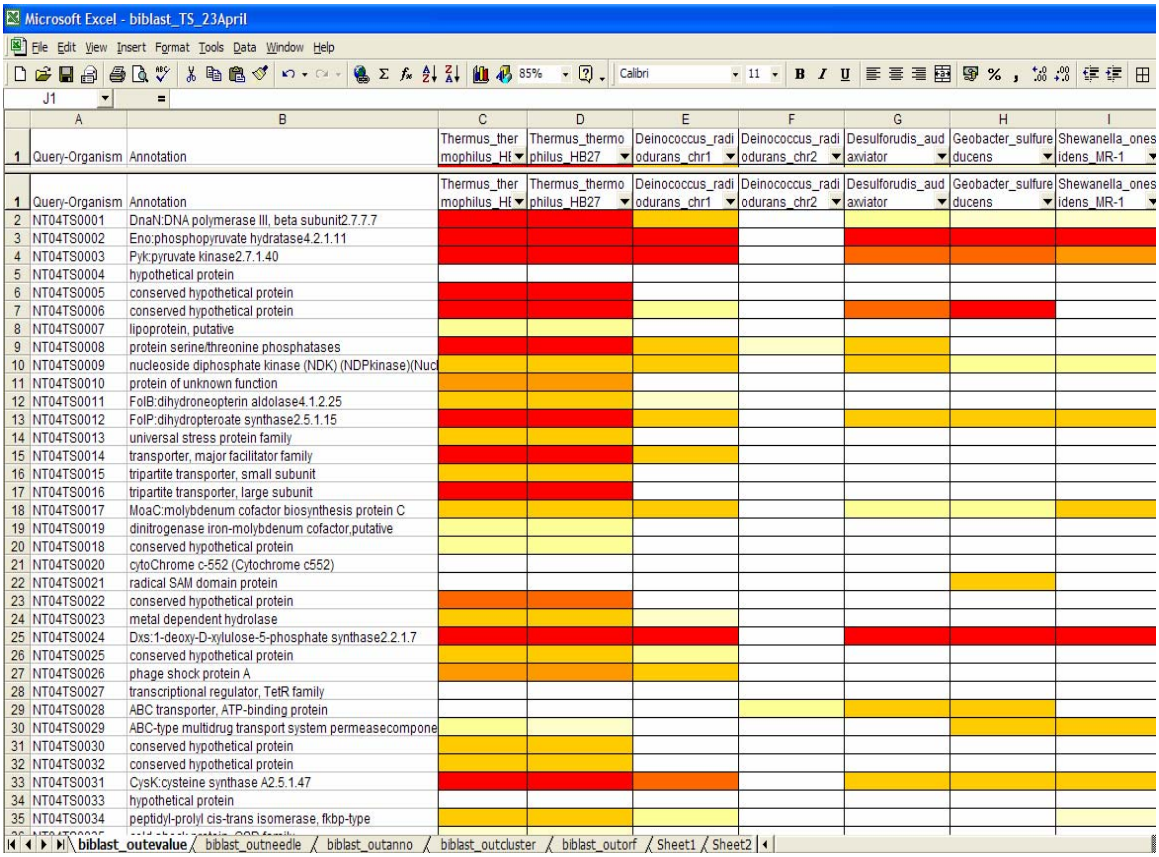


Fig 2.42 Excel sheet showing part of the result of a bi-BLAST of *T. scotoductus* SA-01 against *Thermus thermophilus* HB27, *Thermus thermophilus* HB8, *Deinococcus radiodurans, Desulforidis auduxviator, Shewanella oneidensis* MR-1 and *Geobacter sulfurreducens* PCA. Red coloured cells represent high similarity whereas lighter colours correlate with lower similarities. White cells imply no bi-directional best BLAST hit.

2.3.24 Bi-directional BLAST genome comparison

Bi-directional BLAST results were plotted against *T. scotoductus* SA-01 using the DNAPlotter (Artemis) program available from the Sanger website (Fig 2.43).

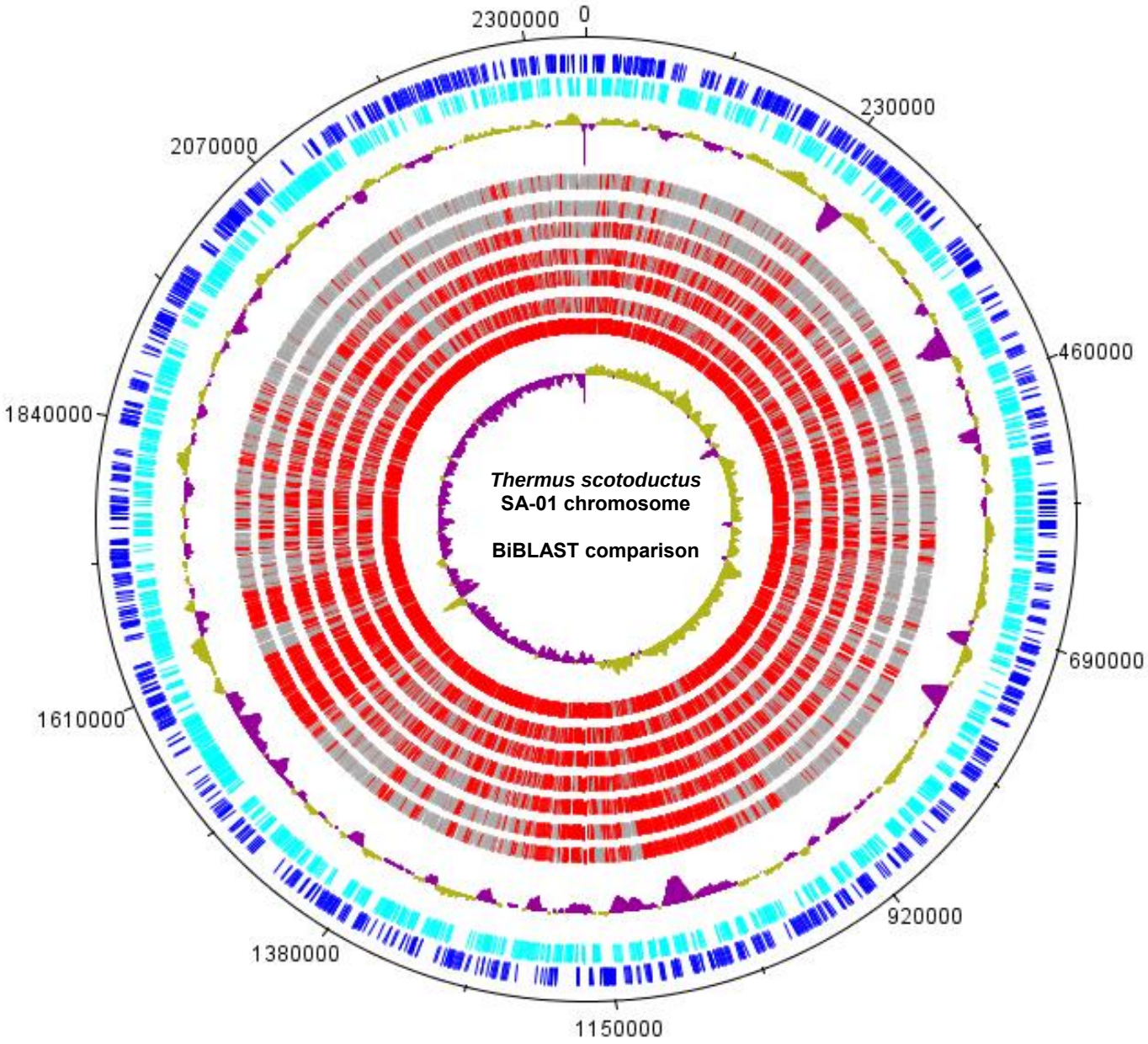


Fig 2.43 Six-way comparison of genomes of choice used for the Bi-BLAST analysis. The innermost ring represents the GC skew, the first red ring represents all putative genes of the genome of *T. scotoductus* SA-01, the third to eighth ring shows all ORFs orthologous to *T. scotoductus* SA-01 in the following order: (*Thermus thermophilus* HB27, *Thermus thermophilus* HB8, *Deinococcus radiodurans, Desulforidis audaxviator, Geobacter sulfurreducens and Shewanella oneidensis*). Red lines indicate high homology whereas grey lines represent low homology the ninth ring represents the G+C variation, the two blue rings represent the ORFs from *T. scotoductus* SA-01 in their respective orientations and the outermost circle represents the scale of the genome.

Automatic genome comparison of the annotated ORFs present in *T. scotoductus* SA-01 was performed at the protein level and compared to other genomes of choice. The results are summarized in Table 2.14.

Table 2.14 Six-genome bi-directional BLAST comparison with *T. scotoductus* SA-01.

	Thermus thermophilus HB27	***Thermus thermophilus HB8***	***Deinococcus radiodurans chr1***	***Deinococcus radiodurans chr2***	***Desulforudis auduxviator***	***Geobacter sulfurreducens PCA***	***Shewanella oneidensis MR-1***
Whole genome amount of genes	2 035	2 026	142	369	2 295	3 523	4 561
Orthologous hits to *T. scotoductus* SA-01	807	789	208	17	82	95	56
Genes without hits	779	813	1 533	2 387	1 871	1 814	1 872
Chromosome length (bp)	1,894,877	1,849,742	177,466	412,348	2,349,476	3,814,139	4,969,803

There are 807 and 789 ORFs from *T. thermophilus* HB27 and *T. thermophilus* HB8 respectively, that are orthologous to *T. scotoductus* SA-01. However, the genome comparison reveals much lower ORF similarities that are orthologous to *T. scotoductus* SA-01 from *D. radiodurans* (chromosome 1 and chromosome 2), *D. auduxviator*, *G. sulfurreducens* and *S. oneidensis*.

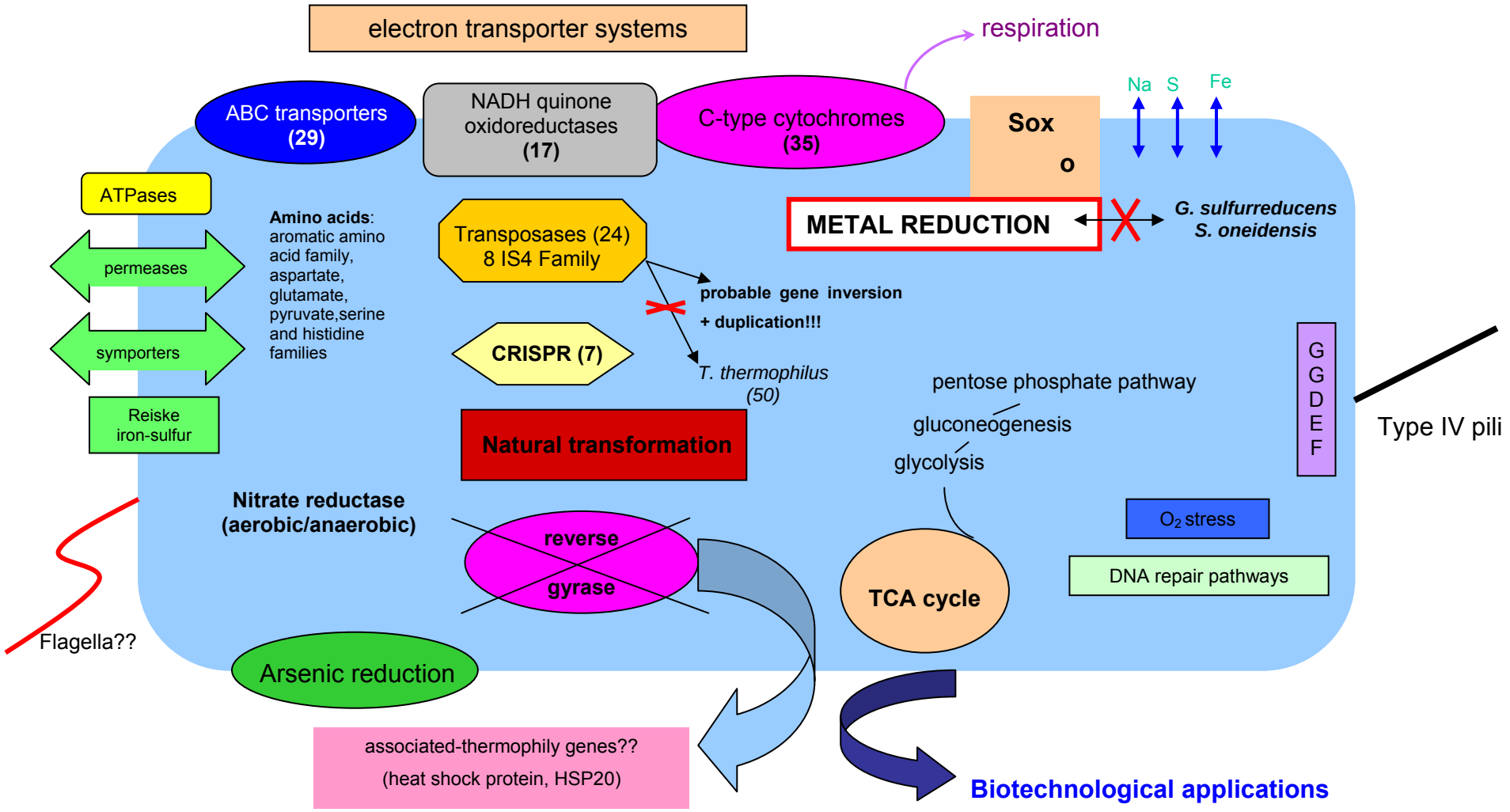


Fig 2.44 Predicted metabolic pathways systems occurring in *T. scotoductus* SA-01.

Using the annotated data obtained as well as the genome comparison results from the Bi-BLAST, the genome features of *T. scotoductus* SA-01 were identified and a metabolic pathways system could be predicted (Fig 2.44). Features of the genome were analysed in their specific gene role categories and compared to organisms of choice.

- Biosynthetic pathways for the following amino acid are present in *T. scotoductus* SA-01: aromatic amino acid family, aspartate, glutamate, pyruvate, serine and histidine families.

- *T. scotoductus* SA-01 has been shown to be metabolically versatile, able to grow aerobically and anaerobically. This is probably due to the presence of the nitrate reductase gene. Three nitrate reductase subunits (alpha, beta and gamma) were determined after annotation of the complete chromosome of *T. scotoductus* SA-01. Comparative analysis indicates that the nitrate reductase is encoded on the plasmid of *T. thermophilus* HB8 as was also indicated by Ramírez-Arcos *et al*., 1998.

- Metabolic pathway features: The predicted metabolic pathways for *T. scotoductus* SA-01 seem to retain pathways for glycolysis, gluconeogenesis, pentose phosphate pathway, pyruvate dehydrogenase and the tricarboxylic acid cycle (TCA).

  The TCA cycle is vital as it provides the substrates of many biosynthetic pathways (Deckert *et al*., 1998). *T. scotoductus* SA-01 appears to contain a complete set of genes for the TCA cycle. The genome contains the genes that encode malate dehydrogenase, fumarate hydratase, fumarate reductase, succinate CoA ligase, ferredoxin oxidoreductase, isocitrate dehydrogenase, aconitase, and citrate synthase which together could constitute the TCA pathway.

  A complete set of genes involved in glycolysis and the pentose phosphate pathway are present. The genome encodes the phosphoenolpyruvate carboxylase, which fulfils the sole reaction by the irreversible carboxylation of phosphoenolpyruvate to oxaloacetate, however it seems that gluconeogenesis is accomplished by phosphoenolpyruvate carboxykinase. The key enzyme in gluconeogenesis is fructose-1,6-bisphosphatase, which

catalyses the irreversible dephosphorylation of fructose-1,6-phosphate to fructose-6-phosphate is also present on the genome.

- Electron transport: The electron transport system seems quite extensive in *T. scotoductus* SA-01 containing a large number of NADH quinone oxidoreductases (17 ORFs) and 35 cytochromes C that probably feed electrons into the transport chains, in addition to various transporters for the uptake of substrates and ions. We have been able to identify 29 ABC transporters, 1 sodium-alanine symporter and other permeases (iron/zinc, etc).

- Balkwill *et al* (1999), characterized *T. scotoductus* SA-01 as a facultative anaerobe capable of coupling the oxidation of organic substrates to reduction of a wide range of electron acceptors, including nitrate, Fe(III), Mn(IV) and Sº. It is therefore probable that the cytochrome C oxidase subunit cluster, electron transport protein SCO1/SenC and iron-sulfur proteins aid in the organisms growth in anaerobic conditions. In addition 3 rieske iron-sulfur proteins were identified as well in the *T. scotoductus* SA-01 genome. Rieske proteins are essential subunits of the cytochrome *bc*-complexes, which are often of crucial importance for the energy metabolism of the cells (Schmidt, 2004).

- Arsenite oxidase (large and small subunits), which detoxifies arsenic, was found on the chromosome of *T. scotoductus* SA-01. However, these genes are found on the plasmid TT8 of *T. thermophilus.* Previous reports indicate that investigations done at the hot spring in the Yellowstone National Park showed rapid arsenite oxidation by *Thermus* species, indicating an arsenic-rich thermal environment where the HB8 strain was originally isolated from (Brüggemann and Chen, 2006). However, no evidence has shown that arsenic was present in the thermal mine environment where *T. scotoductus* SA-01 was isolated so it highly possible that these genes are present in the genome due to horizontal gene transfer. However, this would need to be determined by further research.

- *T. scotoductus* SA-01 also contains proteins for heat shock HSP20 family proteins and for sporulation and germination that may provide heat resistance.

- When compared to the *D. audaxviator genome*, *T. scotoductus* SA-01 is able to cope with oxygen stress by possessing several mechanisms to protect the cell against oxidative damage with proteins such as a superoxide dismutase [Mn] (General stress protein 24), manganese catalase and a peroxiredoxin-like protein.

- *T. scotoductus* SA-01 has repair pathways that include DNA replication, recombination, and repair. Some important proteins of this category include the RecA protein, uracil-DNA glycosylase, DNA mismatch repair protein MutS domain as well as DNA polymerase I among other proteins.

- A cluster of Sox genes are also present in the genome of *T. scotoductus* SA-01, located downstream of a number of cytochrome c genes, which have a high similarity to the same proteins present in both *Thermus thermophilus* strains. The cluster of Sox genes present in *T. thermophilus* is homologous to the Sox genes present in sulfur-oxidizing organisms (Henne *et al*., 2004). According to Omelchenko *et al*., 2005, the Sox operon present in *T. thermophilus* might have been horizontally transferred from *Aquifex aeolicus* with some local rearrangements. The presence of this Sox operon suggests that *T. scotoductus* SA-01 can reduce sulfur compounds as a source of energy and sulfur.

- Thermophily genes: Comparative genomics is a useful approach for extracting candidate genes associated with thermophily. Studies have shown that one particular gene that features prominently in hyperthermophiles is the reverse gyrase. This gene is thought to help DNA to function at high temperatures by increasing topological links between the two DNA strands. However, Takami *et al* (2004) noticed that the reverse gyrase is noticeably absent from the *T. thermophilus* genome. After inspection of the *T. scotoductus* SA-01, we also noticed that the reverse gyrase is absent. Forterre in 2002, looked at all proteins present in hyperthermophiles but absent from mesophile genomes, and found only one hyperthermophile-

specific protein: reverse gyrase. This indicates that this gene is important to organisms that have broken the 80ºC growth temperature barrier in their adaptation of life to very high temperatures.

- Transposable elements/mobile elements: The *T. scotoductus* SA-01 genome possesses 24 genes encoding transposases of which 8 seem to be carried by IS4 family protein. Although the number of transposases was high in *T. scotoductus* SA-01, only one transposase was similar to *T. thermophilus* which contains 50 IS elements (Henne *et al*., 2004).

  The genome re-arrangement observed in *T. scotoductus* SA-01 is probably mediated by the transposons present and can play a role in the genome plasticity. This was also seen with the presence of some genes duplicated on chromosome and plasmid as shown by the BLAST results of plasmid against chromosome. We have been able to identify four ORFs found on the plasmid present in one or more copies on the chromosome. The study of *Burkholderia mallei* indicated that the large number of IS elements could be the cause of most of the synteny break points when compared to the genome of *B. pseudomallei* (Fraser-Liggett, 2005). Nierman *et al* (2004) also indicated that the idea of genome rearrangement can play a large role in genome structural alteration in certain species as is also seen in *T. scotoductus* SA-01.

- The annotated results indicate that *T. scotoductus* contains genes for chemotaxis and motility, such as twitching mobility protein and PilT domain protein. Type IV pili and homologues of PilT have been found to be essential for the social gliding motility exhibited by other microorganisms (Okamoto and Ohmori, 2002). This gene also plays a functional role in attachment, surface colonisation as well as natural transformation (Henne *et al*., 2004). In addition, *T. scotoductus* SA-01 possesses 3 GGDEF domain proteins, which may also be involved in environmental signal transduction.

- Metal reduction: In order to determine the presence of metal reducing genes, the genome of *T. scotoductus* SA-01 was compared with the genomes of the metal reducing organisms *Shewanella oneidensis* and *G. sulfurreducens* (shown in the Bi-BLAST comparison). The complete genome sequence of the dissimilatory metal ion-reducing bacterium *S. oneidensis* offered a starting

point to define *T. scotoductus* SA-01 organisms complex electron transport systems and metal ion-reducing capabilities (Heidelberg *et al*., 2002). In addition it was also noticed in 2006 by Roh *et al*, that this strain is able to reduce metals such as Fe(III), Co(III), Cr(VI), Mn(VI) and U(VI). Similarly, in 1999, Kieft *et al*., described the isolation of the facultatively anaerobic *Thermus* strain that is capable of dissimilatory iron reduction and is capable of growing using nitrate, Fe(III), Mn(IV) or S^o as terminal electron acceptors and to also be capable of reducing Cr(VI), U(VI), Co(III) and the quinone-containing compound anthraquinone-2,6-disulfonate (Balkwill *et al*., 2004).

Genome analysis shows that *S. oneidensis* has more c-type cytochromes than any other organism sequenced to date (*S. oneidensis* (39), *V. cholerae* (12), *E. coli* (7), *P. aeroginosa* (32)), including 14 c-type cytochromes with four or more heme-binding sites not described before in *S. oneidensis*. In addition, 80% of membrane-bound c-type heme is localized to the outer membrane suggesting a direct role for c-type cytochromes in metal reduction (Heidelberg *et al*, 2002). However, in 2005, Fredrickson and Romine, determined that after the reannotation of the *S. oneidensis* genome (Daraselia *et al*., 2003), the previously predicted number of c-type cytochromes increased from 39 to 42. Remarkably, within the smaller size of the complete genome of the metal reducing microorganism *G. sulfurreducens* (Methe *et al*., 2003), 90 newly reported predicted c-type cytochromes were present. This high number of c-type cytochromes in these organisms is thought to reflect their highly branched electron transfer transport systems that convey extensive versatility in terms of electron acceptor utilization (Fredrickson and Romine, 2005). In addition, Marshall *et al* (2006), showed experimentally that the c-type cytochromes of *S. oneidensis* MR-1 are essential for the reduction of U(VI) and formation of extracellular UO_2 nanoparticles.

Annotation results of the *T. scotoductus* SA-01 genome data indicate the presence of 35 c-type cytochromes. Bi-BLAST analysis however, indicate that none of the c-type cytochromes present in *T. scotoductus* SA-01 are similar to the c-type cytochromes present in either models of metal-reducing microorganisms: *S. oneidensis* and *G. sulfurreducens*. According to our data (annotated and Bi-BLAST results), we postulate that the c-type cytochromes possibly play a role in the metal reducing capabilities of *T. scotoductus* SA-01,

however to deduce the complex networks and pathways involved in the mechanisms of these enzymes, further experimental data is required.

- DNA transformation: Natural competence (the ability to take up and process exogenous DNA in specific growth conditions) has been observed in the *T. thermophilus* strains. High frequencies of natural transformation has also been displayed by this organism (Friedrich *et al*., 2003) and Schwarzenlander and Averhoff (2006) also proposed that the extraordinary broad substrate specificity of the highly efficient *T. thermophilus* HB27 DNA uptake system may contribute significantly to thermoadaptation of the organism to interdomain DNA transfer in hot environments. Several genes associated with transformation have been determined in *T. scotoductus* SA-01 after annotation. These genes include competence/damage-inducible protein, competence proteins DprA, ComF, PilM, PilN, PilO, competence factor ComEA and DNA internalization-related competence protein ComEC/Rec2. Friedrich *et al* (2001), obtained clear evidence that the comEA and comEC competence genes are essential for natural transformation in *T. thermophilus* HB27. Although the genome data of *T. scotoductus* SA-01 encodes of the competence functions, DNA transformation using this strain has yet to be tested experimentally.

- Biotechnological applications: Pantazaki *et al*., (2002), discussed the biotechnological relevant genes from *Thermus thermophilus*. We have found several genes of biotechnological applications that are present in *T. scotoductus* SA-01 such as β-glucosidase, β-galactosidases, zinc protease, ATP-dependant serine protease, L-asparaginase, phosphatases, pyrophosphatases, antioxide enzymes such as superoxide dismutase, DNA polymerases, DNA ligase (NAD dependent), Holliday junction DNA helicase RuvA, endonucleases (Type I and TypeIII present) and exonucleases, RNA helicases and ribonucleases, all of which have interesting applications.

- CRISPR regions: Clustered regularly interspaced short palindromic repeats (CRISPR) are a distinctive feature of the genomes of most Bacteria and Archaea and are thought to be involved in resistance to bacteriophages **(**Barrangou *et al*., 2007). These regions are usually between 24-48 bp long and are separated by spacers of similar length.

Initially, using a CRISPRFinder program online program (http://crispr.u-psud.fr/Server/CRISPRfinder.php/) on the draft genome sequence, we were able to determine that our assembled contigs contain 3 confirmed CRISPR regions and 1 questionable CRISPR region. This result is almost similar to the *T. thermophilus* HB27 and *T. thermophilus* HB8 genome, both of which contain 2 CRISPR regions. The completed *T. scotoductus* SA-01 chromosome revealed a cluster of the following 7 CRISPR genes: crispr-associated protein Cas2, Cas1, Cse3 family, Cas5, *E. coli* subtype, Cse4 family, Cse2 family, Cse1 family and crispr-associated helicase Cas3. Interestingly, the genome comparison results indicated that these genes are not common to the other genomes used in the Bi-BLAST. The cas3 gene showed motifs characteristics for helicases of the superfamily 2 and the cas4 gene showed motifs of the recB family of exonucleases, suggesting that these genes are involved in DNA metabolism or gene expression (Jansen *et al*., 2002). According to Barrangou *et al* (2007), CRISPR, together with the *cas* genes, may provide acquired resistance to bacteriophages, with a resistance specificity determined by similarity between the spacers and phage sequences. The apparent presence of an intact CRISPR system, could possible also imply that the ancestral genome may have suffered an invasion of exogenous genetic components, although only a few traces of phages are present in the genome remain (Hongou *et al*., 2008).

2.4 Conclusion

In contrast to a report that suggests that high-density pyrosequencing is unable to replace Sanger sequencing for *de novo* microbial genome projects (Goldberg *et al*., 2006), this study has shown that the 454 sequencing technique is able to provide an efficient coverage of a genome sequence even if the sequence has a high GC content or a repetitive genome is targeted. However, if a finished genome is required the use of an appropriate strategy employing Sanger sequencing is imperative. We also found that an initial assembly of ~ 20 fold coverage pyrosequencing data is sufficient to determine the genome size and to build a working draft that can be used for genome analysis with automatic annotation.

The final assembly with both the GS20 and FLX data resulted in a low number of contigs (35) which is remarkable, considering that repetitive sequences such as IS elements and RNA regions are the main cause of gaps in pyrosequencing projects due to the short length of the sequencing reads. BLASTn results performed at the end of each contig indicated that gaps were caused due to repeat regions and problematic sequence areas in the genome (e.g. high GC content, secondary structures etc.). However, with the construction of a fosmid library and an appropriate contig order, many of the gaps could be closed with the use of Sanger sequencing. Sanger reads are much longer than the GS20 and FLX read length thereby providing the required overlap required to close a gap area, especially repeat regions. In most cases, repeats were poorly covered by pyrosequencing and were pooled and assembled into a consensus contig due to the 100 and 200 bp read lengths.

Using the MUMmer software package, comparative genome analysis of *T. scotoductus* SA-01 was carried out with *T. thermophilus*. Only very short regions of synteny were found. The presence of extensive rearrangements present in the genome organization of *T. scotoductus* SA-01 when compared to *T. thermophilus* is quite remarkable since there are moderate rearrangements in the other two *Thermus* genome architecture. However, we could attribute that the extent of rearrangements could reflect the lifestyle of *T. scotoductus* SA-01.

A comparative genomic approach was employed to determine the metabolic capabilities of *T. scotoductus* SA-01 as well as determine its metal reducing

capabilities. However, identifying the genes that are generally associated with metabolism and thermophily through comparative genomics is generally not easy.

Metal reducing organisms used for BLAST comparison are not phylogenetically related to the *Thermus* species and genes involved in metal reduction are not similar to genes present in *T. scotoductus* SA-01. The reason for difficulty in determining genes for thermophily is because phylogenetically related thermophiles share many genes that are not directly associated with thermophily. According to Takami *et al* (2004), phylogenetically distant thermophiles may have different mechanisms for thermoadaptation and it is possible that genes responsible for thermophily may be among those genes whose function is yet unknown.

A complete genome sequence is indeed important. It is able to provide not only an invaluable tool for ongoing biological research but also are capable of generating new hypotheses for future research work. The complete chromosome and draft sequence of *T. scotoductus* SA-01 not only provides valuable basic data in terms of the organisms lifestyle and capabilities but may also pose many questions warranting several new lines of research with regards to biological processes, activities and potential of this species that had not been evident before. The genome analysis also yielded many functional predictions that can be tested experimentally.

From the increasing amount of data obtained from various genome sequencing projects, growing databases of sequences as well as comparative genomic approaches, we can significantly improve our current annotation data. This would further improve our interpretation of the *T. scotoductus* SA-01 unique metabolic versatility.

Chapter 3

Cloning and Expression of the DNA polymerase I (DNA PolI) and single-stranded DNA-binding (SSB) protein from *T. scotoductus* SA-01 to enhance the efficiency of PCR.

3.1 Introduction

The extremely thermophilic bacterial species of the genus *Thermus* have an optimal growth temperature above 65ºC. Therefore, the genus has attracted considerable attention as a source of robust, thermostable enzymes, utilized in various biotechnological applications (Brüggemann and Chen, 2006). A number of highly thermostable enzymes, which have proven to be useful in high-temperature systems, have been isolated from *Thermus* bacteria (Park *et al*., 1993). Enzymes from thermophiles are not only more resistant to temperature than their mesophilic counterparts, but they also generally exhibit greater tolerance to pH, exposure to solvents and exposure to pressure. Thus, thermostable enzymes possess qualities that make them more robust and better suited for use in industrial processes. Several products from thermophiles have already been commercialised and there is a high level of interest in identifying new enzymes from thermophiles (Park *et al*., 2004). In particular, DNA polymerases from extreme thermophiles have drawn interest because of the application to gene amplification (Park *et al*., 1993).

The Polymerase Chain Reaction (PCR), which uses a thermostable DNA polymerase, is one of the most important developments in protein and genetic research and is currently used in a broad array of biological applications (Kim *et al*., 2007) and accordingly thermostable DNA polymerases have become an indispensable tool for DNA experiments in molecular biology (Park *et al*., 1993). Early PCR experiments used the thermolabile Klenow fragment, which had to be added to every cycle. The introduction of a thermostable DNA polymerase allowed the automation of the process (Choi *et al*., 2001). Thermostable DNA polymerase was first isolated and purified from the thermophilic bacterium *T. aquaticus* YT-1 (Chang *et al*., 2001). The high temperature optimum, 75ºC, affords unique advantages when comparing Taq Pol I to *Escherichia coli* DNA polymerase I. Also, *E. coli* DNA polymerase I is inactivated at 93-95ºC, the temperature range required to

denature the duplex DNA (Lawyer *et al*., 1989). Thus far, more than 50 DNA polymerases have been cloned from various organisms, such as thermophiles and archaea (Kim *et al*., 2007). DNA polymerases that have been purified from the *Thermus* species include *T. thermophilus* (Moreno *et al*., 2005), *T. ruber*, *T. flavus*, *T. cauldophilus* GK24 (Park *et al*., 1993, Kwon *et al*., 1997), *Thermococcus* sp. (Kim *et al*., 2007), and *T. scotoductus* (online patent). Other DNA polymerases have been isolated from thermophiles such as *Aquifex pyrophilus* (Choi and Kwon, 2004) and a Pfu DNA polymerase from *Pyrococcus furiosus*, which contained an integrated 3´-5´ exonuclease activity that corrects errors introduced during the polymerisation (Lu and Erickson, 1997). Taq polymerase belongs, like the *E. coli* DNA polymerase I (*E. coli* DNA Pol I) and the *Thermotoga neapolitana* DNA polymerase (*Tne* polymerase), to the group of Pol I-like DNA polymerases (Villbrandt *et al*., 2000).

There are an increasing number of studies which report the usefulness of single-stranded DNA-binding proteins (SSBs) for PCR (Kur *et al*., 2005). SSBs are indispensable elements in cells of all living organisms. These proteins interact with ssDNA in sequence in an independent manner, preventing them from forming secondary structures and protecting them from degradation by nucleases. In such a manner, SSB-binding proteins can participate in all processes involving ssDNA, such as replication, repair and recombination (Dabrowski *et al*., 2002). Most SSBs bind non-specifically to single-stranded DNA (ssDNA), conferring a regular structure upon it, which is recognized and exploited by a variety of enzymes (Perales *et al*., 2003). Reports have shown that the DNA-binding proteins, gene 32 protein from bacteriophage T4 and the native or His_6-tagged *Eco*SSB proteins were successfully used for enhancement of amplification efficiency for large and small fragments. The SSB-like proteins of the thermophilic bacteria *Thermus thermophilus* HB-8 and *Thermus aquaticus* have also proven to be generally applicable in improving the PCR efficiency (Dabrowski *et al*., 2002). SSBs of thermophilic origin would be ideal candidates for such an application due to their high thermostability. The unique ability of SSB to bind single-stranded DNA (ssDNA) but not double-stranded DNA (dsDNA) allows efficient separation of three types of DNA molecules in the PCR reaction mixture: primers, products (amplified templates) and by products, which originate from non-specific DNA hybridisation (Kur *et al*., 2005). Perales *et al*., 2003 overexpressed and purified the native form and two His-tagged fusions of the SSB from *T. thermophilus* (TthSSB). The three proteins bound in vitro to ssDNA specifically over a temperature range of 4-80ºC and the wild-type protein could withstand incubation at 94ºC for 2 mins. Also, the addition of TthSSB to PCR halved

the elongation time required for the DNA polymerase of *T. thermophilus* and *Pyrococcus furiosus* (Pfu) to synthesis DNA fragments in PCRs.

Recently, the publication of an online patent application (Lee *et al*., 2007) reported the successful fusion of a thermostable SSB from *Sulfolobus solfataricus* and the thermostable DNA polymerase from *Thermococcus zilligi*. The authors isolated and purified a SSB-nucleic acid polymerase fusion protein which resulted in increasing the yield of PCR on a target DNA by contacting the target DNA with a primer, which specifically hybridises thereto. In addition in 2004, Wang *et al.,* showed that by using Sso7d from *Sulfolobus solfataricus* as the DNA binding protein, the processivity of both family A and B polymerases can be significantly enhanced. However, this was done by introducing point mutations in Sso7d which was found to be essential for the enhancement.

Whole genome sequencing has been done on the extremophile *T. scotoductus* SA-01 and the preliminary BLAST and annotation results revealed the presence of a DNA polymerase I as well as a single-stranded DNA-binding (SSB) protein.

This chapter describes the cloning, expression in the heterologous host *E. coli*, and purification of the DNA polymerase and single-stranded DNA binding protein obtained from the draft genome sequence of *T. scotoductus* SA-01.

3.2 Materials And Methods

3.2.1 Bacterial strains, plasmids and growth conditions

All bacterial strains and plasmids used in this study are listed in Table 3.1. *Thermus scotoductus* was cultured in TYG medium. *Escherichia coli* strains TOP10 (Invitrogen), BL21 (DE3) (Lucigen) and BL21 (pLysS) were used as hosts for genetic manipulation and expression of proteins respectively. *E. coli* strains were grown in Luria- Bertani (LB) medium at 37°C with shaking (200 rpm). Kanamycin (30 $\mu g.ml^{-1}$) was added when required. Plasmid pET28b(+) (Novagen) was used for expression of the proteins in *E. coli* BL21 (DE3) and BL21 (pLysS).

Table 3.1 Bacterial strains and plasmids used in this study.

Strain or plasmid	Description	Reference
Thermus scotoductus SA-01		ATCC 700910
Escherichia coli TOP10	One Shot TOP10 chemically competent cells F^- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lacZ*ΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (StrPRP) *end*A1 *nup*G	Invitrogen
Escherichia coli BL21 (DE3)	E.cloni EXPRESS BL21(DE3) chemically competent cells F^- *ompT hsdS*BBB (*r*BB-B *m*BBPB$^-$) *gal dcm* (DE3)	Lucigen
Escherichia coli BL21 (DE3) pLysS	F–, *omp*T, *hsd*S_B (r_B–, m_B–), *dcm*, *gal*, λ(DE3), pLysS, Cm^r.	Promega
pGEM®-T Easy	Amp^rP, T7 and SP6 promoter, *Lac*Z, ori	Promega
pET28(b)+	Kan^rP, T7 promoter, *Lac*I, N-terminal His-Tag and Thrombin configuration, ori	Novagen

3.2.2 Cloning of the *T. scotoductus* SA-01 DNA Polymerase I and SSB genes

The complete DNA polymerase gene and single-stranded DNA binding protein (SSB) genes were amplified by PCR from genomic DNA using the Expand High Fidelity PCR System (Roche). PCR reactions were performed in a total reaction volume of 50 µl using a Thermal Cycler (PxE 0.2, Thermo Electron Corporation). Reaction mixtures consisted of 10 X Expand High Fidelity Buffer with 15 mM $MgCl_2$ (5 µl), dNTP's (0.8 mM), Expand High Fidelity Enzyme mix (0.75 µl), 50 ng of gDNA and 0.2 µM of both the forward and reverse primers. Primer sets for each product are given in Table 3.2.

Reaction conditions for the DNA polymerase I gene consisted of an initial denaturing step at 95°C for 2 min, followed by 30 cycles of denaturing at 95°C (30 sec), annealing at 58°C (30 sec) and elongation at 72°C (3 min). A final elongation step of 10 min at 72°C was added to ensure complete elongation of amplified products. The reaction conditions for the SSB gene were the same as above with one exception; the elongation step was done at 72°C for 1 minute.

Table 3.2 Primer sequences used for PCR amplification of the selected genes from *T. scotoductus* SA-01.

Primer	Sequence
TS_PolIF	5'- CAT ATG AGG GCG ATG CTG CCC CTC TTT -3'
TS_PolIR	5'- AAG CTT CTA GGC CTT GGC GGA AAG CCA GTC -3'
TscSSB_F	5' -GCCATATGGCAAGAGGCCTGAA-3'
TscSSB_R	5'-GCAAGCTTTCAAAACGGCAAAT-3'

Underlined sequences indicate introduced restriction sites for *Nde*I and *Hin*dIII.

The PCR with primers TS_PolIF and TS_PolIR led to the amplification of a ~2 500 bp and the primers TscSSB_F and TscSSB_R amplified the 800 bp respectively, coding sequence within the *T. scotoductus* SA-01 genomic DNA. PCR products were cut from agarose gels and purified using the Zymoclean™ Gel DNA Recovery Kit (Zymo Research).

3.2.3 Constructs for Expression in *E. coli*

The purified PCR products were ligated into pGEM®-T Easy vector overnight at 4°C according to the manufacturer's instructions and proliferated in One Shot TOP10 *E. coli* competent cells (Invitrogen). Plasmids were isolated using the GeneJet MiniPrep kit (Fermentas). Plasmids containing inserts were double digested with the restriction enzymes *Nde*I (0.5 $U.\mu l^{-1}$, Fermentas) and *Hin*dIII (0.5 $U.\mu l^{-1}$, Fermentas) at 37°C (Buffer R, 3 hr) for ligation into the pET28b(+) similarly digested vector. Inserts and digested pET28b(+) vectors (Fig 3.1) were cleaned from an agarose gel using the GFX PCR DNA and Gel Band Purification Kit (Amersham Biosciences). Cohesive end ligations were performed on 50 ng of vector. Ligations were performed in 20 µl reaction volumes overnight at 16°C with 1.5 Weiss $U.\mu l^{-1}$ T4 DNA ligase (New England Biolabs). Ligation mixtures were again transformed into TOP10 *E. coli* and positive clones were identified through plasmid isolation and restriction digestion as described above. The resulting recombinant plasmids designated pETpolI and pETSSB were sequenced and used for further expression studies.

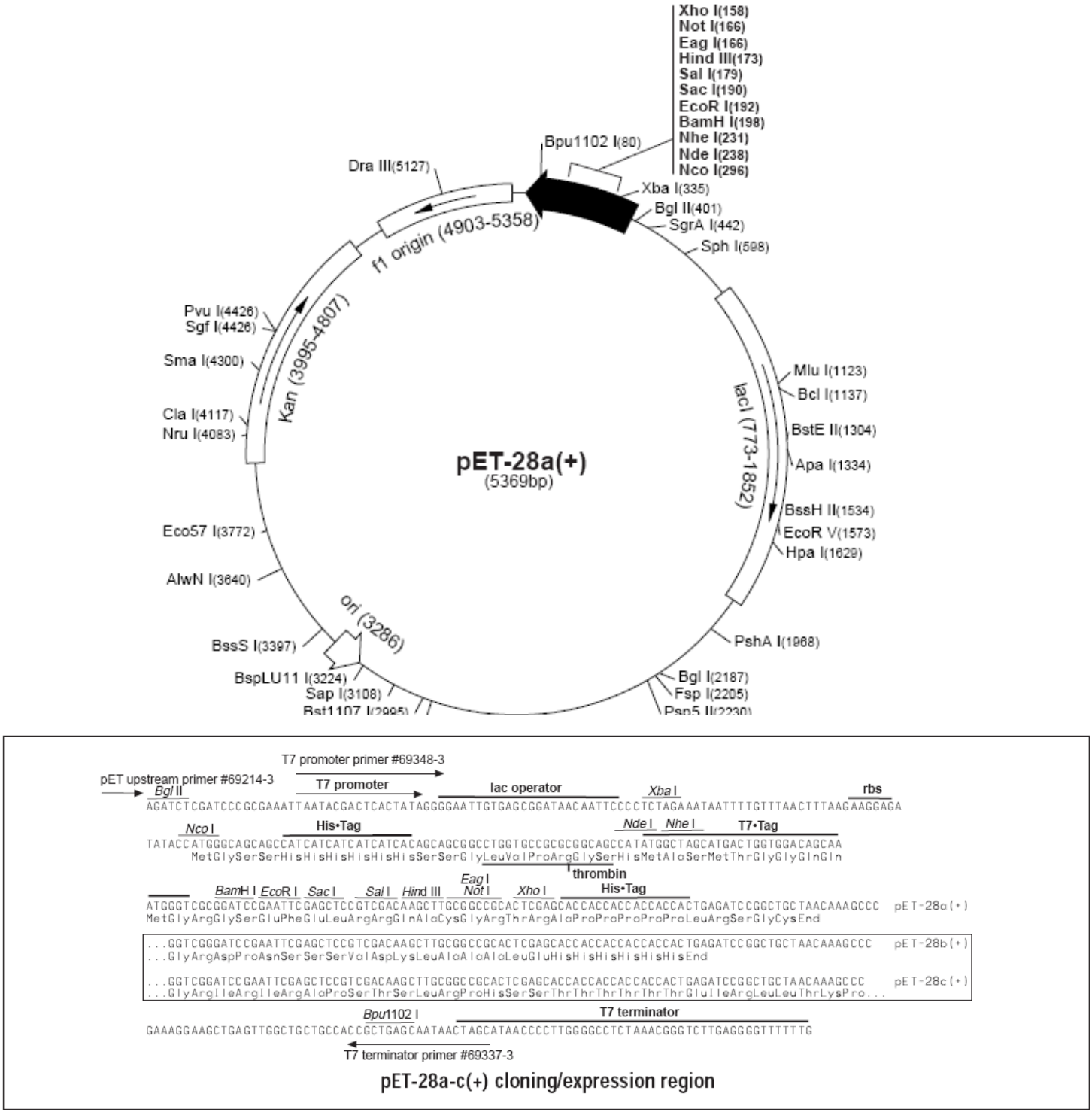


Fig 3.1 Vector map of pET-28b(+) indicating the kanamycin resistance gene, ColE1 origin of plasmid replication, *lacI* coding sequence and the multiple cloning site under the T7 promoter. Sequence of the pET-28b(+) cloning region showing the ribosome binding site and configuration for the N-terminal His-Tag and thrombin cleavage site fusion (Taken from Novagen Vector Manual).

3.2.4 DNA Sequencing and Analysis

The recombinant clones were sequenced using an ABI 3130xl genetic analyser (Applied Biosystems, Foster City, CA), incorporating the ABI Big Dye Terminator Cycle Sequencing kit version 3.1 (Applied Biosystems, Foster City, CA) using the universal T7 terminator and T7 promoter primers.

Sequencing was performed by Inqaba Biotechnical Industries Pty. Ltd., South Africa. Electropherograms of the sequences generated were inspected with FinchTV software (Geospiza) and Vector NTI (Invitrogen). Translated open reading frames (ORFs) were also compared to known sequences deposited in the non-redundant protein databases of the National Centre for Biotechnology Information (NCBI, USA) using standard protein-protein BLAST (BLASTp) (Altschul *et al*., 1997). Sequence alignments were performed using the DNAssist program (Patterton and Graves, 2000).

3.2.5 Protein Sequence Analysis of the pETpolI and pETSSB clones

The translated amino acid sequences of the thermostable DNA polymerase I and SSB-like protein were analyzed using the BLASTp program. Standard protein-protein BLAST search was used to compare the sequences of the proteins against other DNA polymerase and SSB proteins whose sequences are deposited in the database of NCBI. Multiple sequence alignments were generated using the ClustalX program.

3.2.6 Over-expression of the DNA Polymerase

The pETpolI and pETSSB constructs were transformed into *E. coli* BL21 (DE3) and *E. coli* BL21 pLysS competent cells (Lucigen) for expression. Positive clones were identified through selection on LB-plates containing 30 $\mu g.ml^{-1}$ kanamycin and inoculated into LB-medium also containing the appropriate antibiotic.

Cells were incubated in a 37°C shaker incubator (200 rpm) and cell growth monitored until an optical density reading (OD_{600nm}) of approximately 0.8 – 1 was reached. Enzyme production was induced by the addition of IPTG to a final concentration of 1 mM and cells grown for an additional 4 hr. Cells were harvested by centrifugation (8 000 x *g*, 10 min) and the cells washed three times using 50 mM Tris (pH 7).

3.2.7 Purification of Recombinant DNA polymerase I and SSB protein

Harvested cells were resuspended in 20 mM MOPS (pH 7.4) containing 50mM imidazole and 0.5 M NaCl [approximately 1 g cells (wet weight) in 10 ml]. Cells were broken by ultrasonic treatment for 5 mins (100 W), where after unbroken cells and debris were removed by centrifugation (8 000 x *g* for 10 min). The soluble fraction (cytoplasm) was separated from the insoluble fraction (membranes) by ultracentrifugation (100 000 x *g*, 90 mins).

3.2.8 Purification of the DNA polymerase I and SSB protein

All purification steps were carried out at room temperature. The recombinant DNA polymerase and SSB protein purifications entailed two chromatographic steps: metal affinity chromatography and size exclusion chromatography, using the ACTA Prime Purification System (Amersham Biosciences).

The recombinant proteins were purified by immobilized metal affinity chromatography (IMAC). The cytoplasmic fraction was loaded onto a HisTrap FF column (5 ml, Amersham Biosciences) and unbound proteins were eluted (5 ml.min $^{-1}$) using 20 mM MOPS (pH 7.4) containing 20 mM imidazole and 0.5 M NaCl. Bound proteins were then eluted in the same buffer using a linear gradient (100 ml) of imidazole up to 0.5 M. Fractions containing activity were pooled for subsequent purification steps. Collected fractions were pooled and concentrated on an Amicon stirred cell through a 10 kDa MWCO membrane (Osmonics Inc.).

3.2.9 Size-exclusion chromatography

The final purification step was size exclusion chromatography, whereby the native molecular weight (*M*r) of the proteins was also determined. The concentrate was loaded onto a Sephacryl S-100HR column (2.6 x 65 cm; Sigma-Aldrich), equilibrated with 20 mM MOPS−NaOH (pH 7) containing 50 mM NaCl. Proteins were eluted with the same buffer at a flow rate of 0.5 ml.min^{-1}.

3.2.10 SDS-PAGE

Electrophoresis in 10% polyacrylamide resolving and 4% stacking gels in the presence of the anionic detergent SDS was used to monitor the purification process, to assess the homogeneity of the purified fractions and to estimate the relative molecular mass of the enzymes by comparing the electrophoretic mobility with those of standard proteins of known molecular masses.

The SDS-PAGE was performed according to Laemmli (1970). Protein bands were detected with Coomassie Brilliant Blue R-250 stain.

3.2.11 Protein concentrations

Protein concentrations were determined using the bicinchoninic acid (BCA) method (Smith *et al*., 1985). BCA Protein Assay Kit from Pierce (Rockford, IL, USA) was used according to the manufacturer's instructions with bovine serum albumin (BSA) as standard (supplied with kit) to draw a standard curve (Fig 3.2).

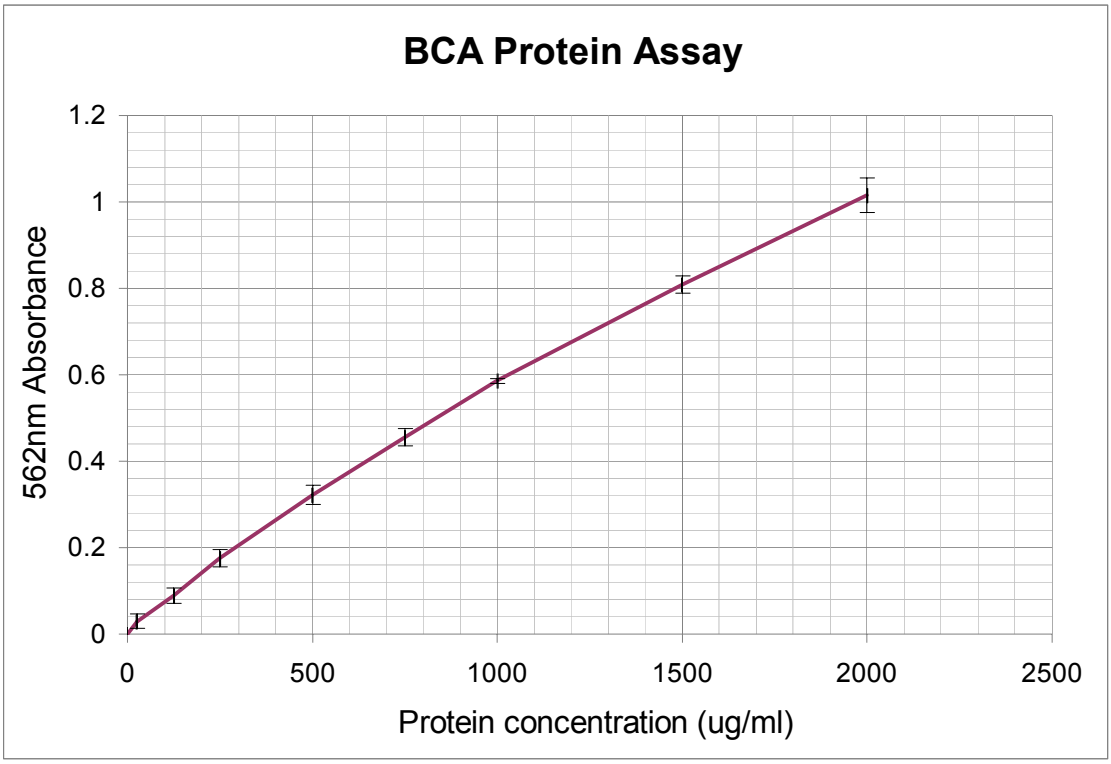


Fig 3.2 Standard curve for the BCA protein assay kit (Pierce) at 37°C using BSA as protein standard.

3.2.12 DNA Polymerase Activity Assay

Enzymatic activity of the thermostable DNA polymerase was analyzed by PCR titrations and compared to commercial Taq enzyme (New England Biolabs). The DNA polymerase protein was diluted 1:10, 1:100; 1:200; 1:400; 1:800 and 1:1600 in storage buffer (50 mM Tris-HCl, pH 8, 100 mM NaCl, 0.1 mM EDTA, 0.5 mM DTT, 1% Triton X-100 and 50% glycerol). In addition, different 10 X PCR reaction buffer systems were prepared and tested: Buffer A (commercial 10 X PCR reaction buffer, New England Biolabs), Buffer B (50 mM Tris-HCl, pH 8, 100 mM NaCl, 0.5 mM DTT, 1% Triton X-100 and 50% glycerol) (Desai and Pfaffle, 1995); Buffer C (50mM KCl, 10 mM Tris-HCl, pH 9, 0.1% Triton X-100) and Buffer D (750 mM Tris-HCl, pH 9, 500 mM KCl, 200 mM $(NH_4)_2SO4$, 20 mM $MgCl_2$) (Biotools Native DNA polymerase, Biotools B&M Labs, S.A.).

The internal 16S rRNA bacterial-specific primers 314F (5′-CCTACGGGAGGCAGCAG-3′) and the 517R (5′-ATTACCGCGGCTGCTGG -3′) primer were used to amplify the 200 bp product internal to the 16S rRNA genes of bacteria. The PCR amplification was carried out in a PxE Thermal Cycler (Thermo Electron) with a total reaction mixture volume of 50 μl.

The PCR reaction mixture contained 1.0 μl of template DNA (*Geobacillus kaustophilus* HTA426), 1 μl of the 314F primer (10 μM), 1 μl of the 517R primer (10 μM), 2.0 μl of 20 mM deoxynucleoside triphosphates (dNTPs), 1 μl of each diluted purified DNA polymerase protein, 5.0 μl of 10 X buffer, 3 μl of 25 mM $MgCl_2$, 1 μl of 10% BSA and 35.75 μl of sterile distilled water.

The reaction mixture was incubated at 95°C for 5 mins to denature the DNA. This was followed by 25 cycles of amplification, each of which consisted of three steps in the following order: denaturation at 95°C for 45 sec, annealing at 55°C for 45 sec and extension of the primers at 72°C for 1 min. Final extension was at 72°C for 10 mins. Amplification products were visualised on an ethidium bromide containing 1% agarose gel using an UV transilluminator after electrophoresis at 100 V for 90 mins.

3.3 Results And Discussion

3.3.1 DNA Polymerase I and SSB PCR

The DNA polymerase I and SSB gene of *T. scotoductus* SA-01 was successfully amplified to produce a 2 500 bp and 800 bp band respectively (Figure 3.3 and 3.4). Ligation of the amplified gene into pGEM®-T Easy vector followed by subsequent subcloning into the expression vector pET28(b)+ resulted in the construct denoted pETpolI and pETSSB. Clones containing inserts were determined by performing colony PCR using specific primers and restriction analysis (Fig 3.5). Both genes were subcloned into pET28(b)+, such that the recombinant proteins contain a 6X His-tag at the N-terminal domain.

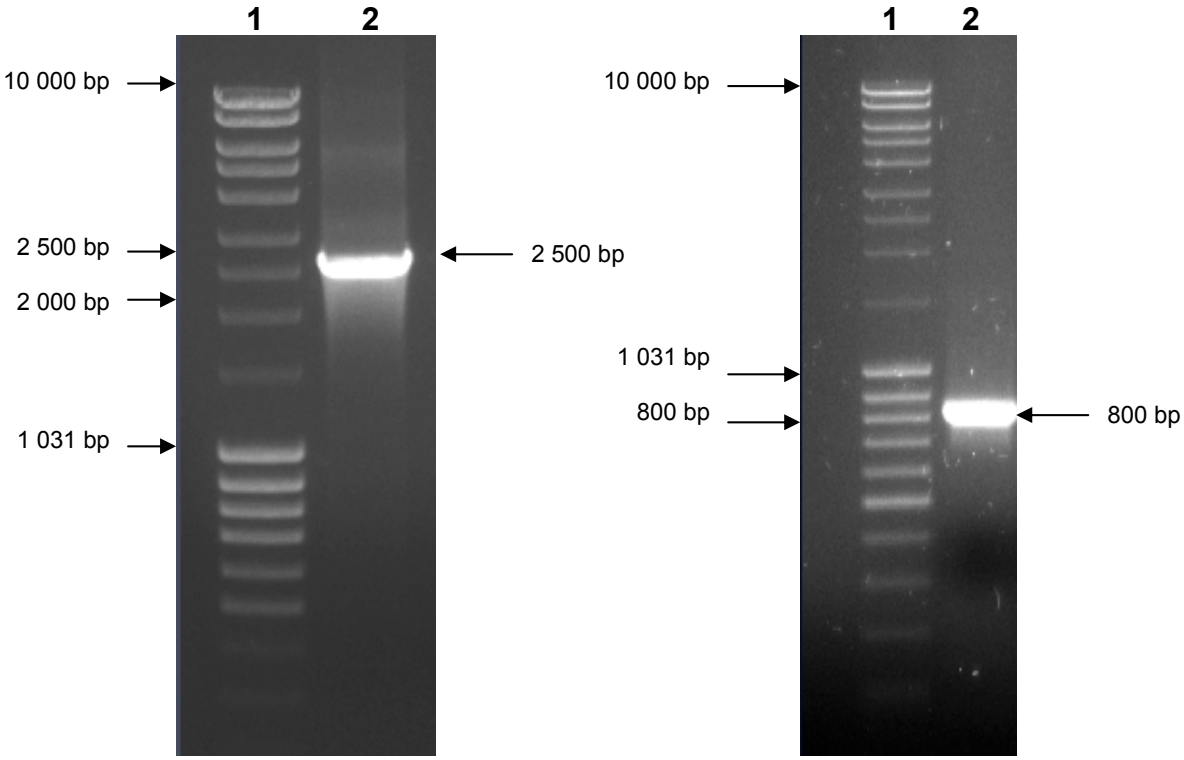


Fig 3.3 Agarose gel electrophoresis of PCR amplified 2 500bp coding sequence for *T. scotoductus* SA-01 DNA polymerase gene (lane 2). Lane 1: Molecular weight marker: MassRuler (Fermentas).

Fig 3.4 Agarose gel electrophoresis of PCR amplified 800 bp coding sequence for *T. scotoductus* SA-01 single-stranded DNA binding (SSB) protein (lane 2). Lane 1: Molecular weight marker: MassRuler (Fermentas).

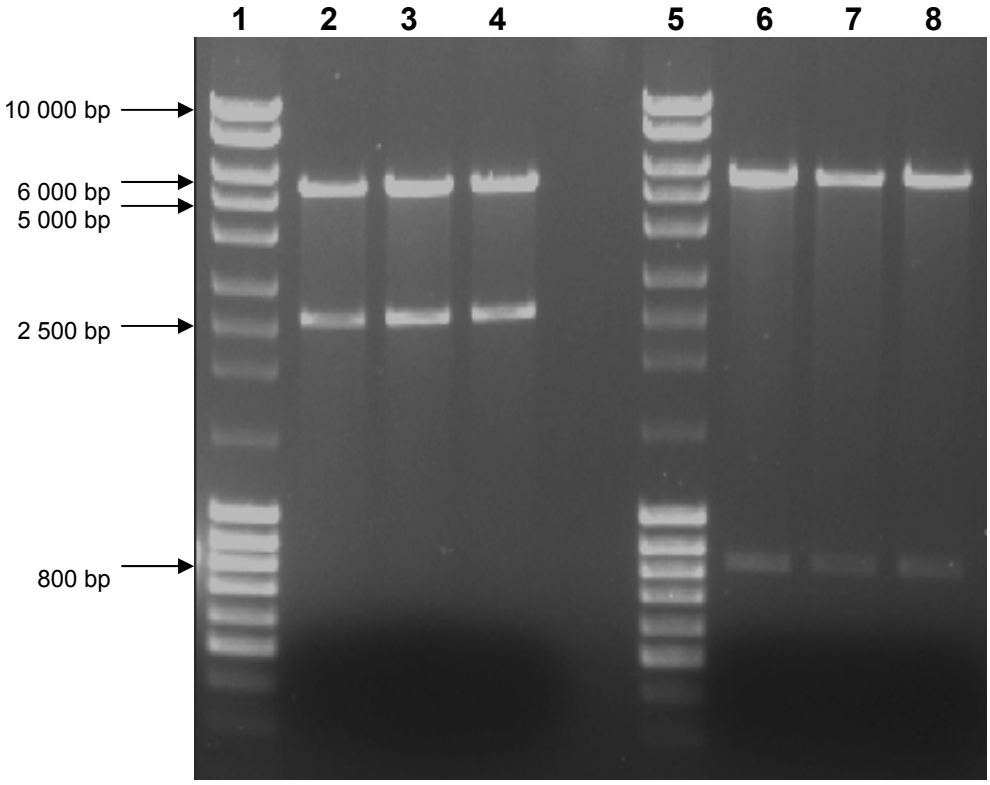


Fig 3.5 Agarose gel electrophoresis of restriction digest of pETpolI and pETSSB clones with enzymes *Hin*dIII and *Nde*I. Lane 1 and 5: MassRuler (Fermentas); lane 2-4: digested pETpolI clone and lane 6-8: digested pETSSB clone with *Hin*dIII and *Nde*I.

3.3.2 Sequence analysis of thermostable DNA polymerase I and SSB

The whole amino acid sequence of *T. scotoductus* SA-01 DNA polymerase I (ORF00918) showed high similarity with those of the *T. thermophilus* HB8 (GenBank accession no. D28878.1), *T. thermophilus* HB27 (GenBank accession no. AE017221.1), *T. aquaticus* (GenBank accession no. J04639.1), *Thermus* sp. NMX2 A1 (FJ358543.1) (Fig 3.6). The alignments of the protein and nucleotide deduced sequence of the DNA polymerase gene bore the highest overall 99% similarity and 99% identity to *Thermus* sp NMX2.A1. The alignment of the DNA polymerase I showed a high similarity of between 85-87% and an identity of between 81-98% when compared to the DNA polymerase I in related species. In 1990, Delarue *et al* compared the C-terminal polymerase domains amongst distantly related members of the PolI family of enzymes. Following the alignments, five conserved regions

(numbered 1-5) within the diverse DNA PolI's were delineated. In addition, a sixth conserved region was observed (region 6). Three of these regions (3, 4 and 5) resembled the most highly conserved domains of the mammalian DNA α in family B (termed motif A, B and C). Alignments of the *T. scotoductus* SA-01 DNA polymerase I, were shown to contain all six conserved regions when compared to related DNA polymerase sequences. According to Patel *et al*., 2001, analysis of the high-resolution crystal structures of family A polymerases in complex with DNA and an incoming nucleotide suggests each of these six regions have an important role during DNA synthesis.

The DNA polymerase I gene from *T. scotoductus* SA-01 consists of 2 490 bp with a GC content of 64.5%, coding for a protein with 830 amino acid residues. The molecular mass of the protein derived from this amino acid sequence was 91 300 Da.

The SSB alignments showed a high similarity between proteins. The amino acid sequence alignment of thermostable SSB-like proteins from *T. thermophilus* HB8 (GenBank accession no. AF079160), *T. thermophilus* VK-1 (GenBank accession no. AF146075), *T. aquaticus* YT-1 ((GenBank accession no. AF276705), *T. scotoductus* SA-01 and pETSSB (ORF00211) is presented in Fig 3.7. According to Dabrowski *et al*., 2002, the C-terminal region of the protein (SSB proteins) contains the four conserved amino acids (DLPF) that are responsible for the binding properties of this protein. We were able to identify the conserved amino acids within our sequence, thereby indicating that the SSB protein from *T. scotoductus* SA-01 could function as a DNA-binding protein. The SSB gene from *T. scotoductus* SA-01 consists of 810 bp coding for a protein with 270 amino acid residues. The molecular mass of the protein derived from this amino acid sequence was 29 700 Da and has a GC content of 67% which is slightly higher than GC content of the genome sequence of 64.9%.

The protein sequence predicted for *T. scotoductus* SA-01 SSB protein shares 80% and 88% similarity and 81% and 86% identity to *T. thermophilus* VK-1 and *T. aquaticus* YT-1, respectively. Alignments of the sequence pETSSB construct and the *T. scotoductus* SA-01 ORF00211 showed a 100% similarity and identity thus indicating no substitutions occurring while cloning.

One distinguishing DNA-metabolic feature of *Deinococcus-Thermus* bacteria is that their SSB proteins are homodimeric, with each SSB monomer encoding two oligonucleotides/oligosaccharide-binding (OB) folds linked by a conserved spacer

sequence (Kur *et al*., 2005). Although the length and the sequence of the C-terminal regions (the region extending past the OB fold is variable across bacterial species, the last 10 amino acids at the C-terminus are highly acidic and well conserved (Eggington *et al*., 2004). The *T. scotoductus* SA-01 SSB gene contains no translational frameshifts and contains two predicted OB folds and its sequence is closely related to the *Thermus* sp. SSB proteins (Fig 3.8). In order to determine if the *T. scotoductus* SA-01 SSB contains two OB folds, the sequences were divided into N- and C-terminal fragments and then aligned (Fig 3.9).

```
T. thermophilus HB8      1   MEAMLPLFEPKGRVLLVDGHHLAYRTFFALKGLTTSRGEPVQAVYGFAKSLLKALKEDGYKAVFVVFDAKAPS  73
T. thermophilus HB27     1   MEAMLPLFESKGRVLLVDGHHLAYRTFFALKGLTTSRGEPVQAVYGFAKSLLKALKEDGYKSVFVVFDAKAPS  73
T. aquaticus             1   MRGMLPLFEPKGRVLLVDGHHLAYRTFHALKGLTTSRGEPVQAVYGFAKSLLKALKEDG-DAVIVVFDAKAPS  72
Thermus sp. NMX2.A1      1      MLPLFEPKGRVLLVDGHHLAYRTFFALKGLTTSRGEPVQAVYGFAKSLLKALREDG-DVVIVVFDAKAPS  69
T. scotoductus SA-01     1      MLPLFEPKGRVLLVDGHHLAYRTFFALKGLTTSRGEPVQAVYGFAKSLLKALREDG-DVVIVVFDAKAPS  69

T. thermophilus HB8      74  FRHEAYEAYKAGRAPTPEDFPRQLALIKELVDLLGFTRLEVPGYEADDVLATLAKKAEKEGYEVRILTADRDL  146
T. thermophilus HB27     74  FRHEAYEAYKAGRAPTPEDFPRQLALIKELVDLLGFTRLEVPGYEADDVLATLAKKAEKEGYEVRILTADRDL  146
T. aquaticus             73  FRHEAYGGYKAGRAPTPEDFPRQLALIKELVDLLGLARLEVPGYEADDVLASLAKKAEKEGYEVRILTADKDL  145
Thermus sp. NMX2.A1      70  FRHQTYEAYKAGRAPTPEDFPRQLALIKEMVDLLGLERLEVPGFEADDVLATLAKKAEKEGYEVRILTADRDL  142
T. scotoductus SA-01     70  FRHQTYEAYKAGRAPTPEDFPRQLALIKEMVDLLGLERLEVPGFEADDVLATLAKKAEKEGYEVRILTADRDL  142

T. thermophilus HB8      147 YQLVSDRVAVLHPEGHLITPEWLWEKYGLRPEQWVDFRALVGDPSDNLPGVKGIGEKTALKLLKEWGSLENLL 219
T. thermophilus HB27     147 YQLVSDRVAVLHPEGHLITPEWLWEKYGLRPEQWVDFRALVGDPSDNLPGVKGIGEKTALKLLKEWGSLESLL 219
T. aquaticus             146 YQLLSDRIHVLHPEGYLITPAWLWEKYGLRPDQWADYRALTGDESDNLPGVKGIGEKTARKLLEEWGSLEALL 218
Thermus sp. NMX2.A1      143 YQLLSERISILHPEGYLITPEWLWEKYGLKPSQWVDYRALAGDPSDNIPGVKGIGEKTAAKLIREWGSLENLL 215
T. scotoductus SA-01     143 YQLLSERISILHPEGYLITPEWLWEKYGLKPSQWVDYRALAGDPSDNIPGVKGIGEKTAAKLIREWGSLENLL 215

T. thermophilus HB8      220 KNLDRVKPENVREKIKAHLEDLRLSLELSRVRTDLPLEVDLAQGREPDREGLRAFLERLEFGSLLHEFGLLEA 292
T. thermophilus HB27     220 KNLDRVKPENVREKIKAHLEDLRLSLELSRVRADLPLEVDLAQGREPDREGLRAFLERLEFGSLLHEFGLLEA 292
T. aquaticus             219 KNLDRLKP-AIREKILAHMDDLKLSWDLAKVRTDLPLEVDFAKRREPDRERLRAFLERLEFGSLLHEFGLLES 290
Thermus sp. NMX2.A1      216 KHLEQVKPASVREKILSHMEDLKLSLELSRVRTDLPLQVDFARRREPDREGLKAFLERLEFGSLLHEFGLLES 288
T. scotoductus SA-01     216 KHLEQVKPASVREKILSHMEDLKLSLELSRVHTDLPLQVDFARRREPDREGLKAFLERLEFGSLLHEFGLLES 288

T. thermophilus HB8      293 PAPLEEAPWPPPEGAFVGFVLSRPEPMWAELKALAACRDGRVHRAADPLAGLKDLKEVRGLLAKDLAVLASRE 365
T. thermophilus HB27     293 PTPLEEAPWPPPEGAFVGFVLSRPEPMWAELKALAACRDGRVHRAEDPLAGLGDLEEVRGLLAKDLAVLALRE 365
T. aquaticus             291 PKALEEAPWPPPEGAFVGFVLSRKEPMWADLLALAAARGGRVHRAPEPYKALRDLKEARGLLAKDLSVLALRE 363
Thermus sp. NMX2.A1      289 PVAAEEAPWPPPEGAFVGYVLSRPEPMWAELNALAAAWEGRVYRAEDPLEALRGLGEVRGLLAKDLAVLALRE 361
T. scotoductus SA-01     289 PVAAEEAPWPPPEGAFVGYVLSRPEPMWAELNALAAAWEGRVYRAEDPLEALRGLGEVRGLLAKDLAVLALRE 361

T. thermophilus HB8      366 GLDLVPGDDPMLLAYLLDPSNTTPEGVARRYGGEWTEDAAHRALLSERLHRNLLKRLEGEEKLLWLYHEVEKP 438
T. thermophilus HB27     366 GLDLAPGDDPMLLAYLLDPSNTTPEGVARRYGGEWTEDAAHRALLSERLHRNLLKRLEGEEKLLWLYHEVEKP 438
T. aquaticus             364 GLGLPPGDDPMLLAYLLDPSNTTPEGVARRYGGEWTEEAGERAALSERLFANLWGRLEGEERLLWLYREVERP 436
Thermus sp. NMX2.A1      362 GIALAPGDDPMLLAYLLDPSNTAPEGVARRYGGEWTEEAGERALLSERLYAALLERLKGEERLLWLYEEVEKP 434
T. scotoductus SA-01     362 GIALAPGDDPMLLAYLLDPSNTAPEGVARRYGGEWTEEAGERALLSERLYAALLERLKGEERLLWLYEEVEKP 434

                                                                                  ▬▬▬▬▬▬▬▬
T. thermophilus HB8      439 LSRVLAHMEATGVRLDVAYLQALSLELAEEIRRLEEEVFRLAGHPFNLNSRDQLERVLFDELRLPALGKTQKT 511
T. thermophilus HB27     439 LSRVLAHMEATGVRLDVAYLQALSLELAEEIRRLEEEVFRLAGHPFNLNSRDQLERVLFDELRLPALGKTQKT 511
```

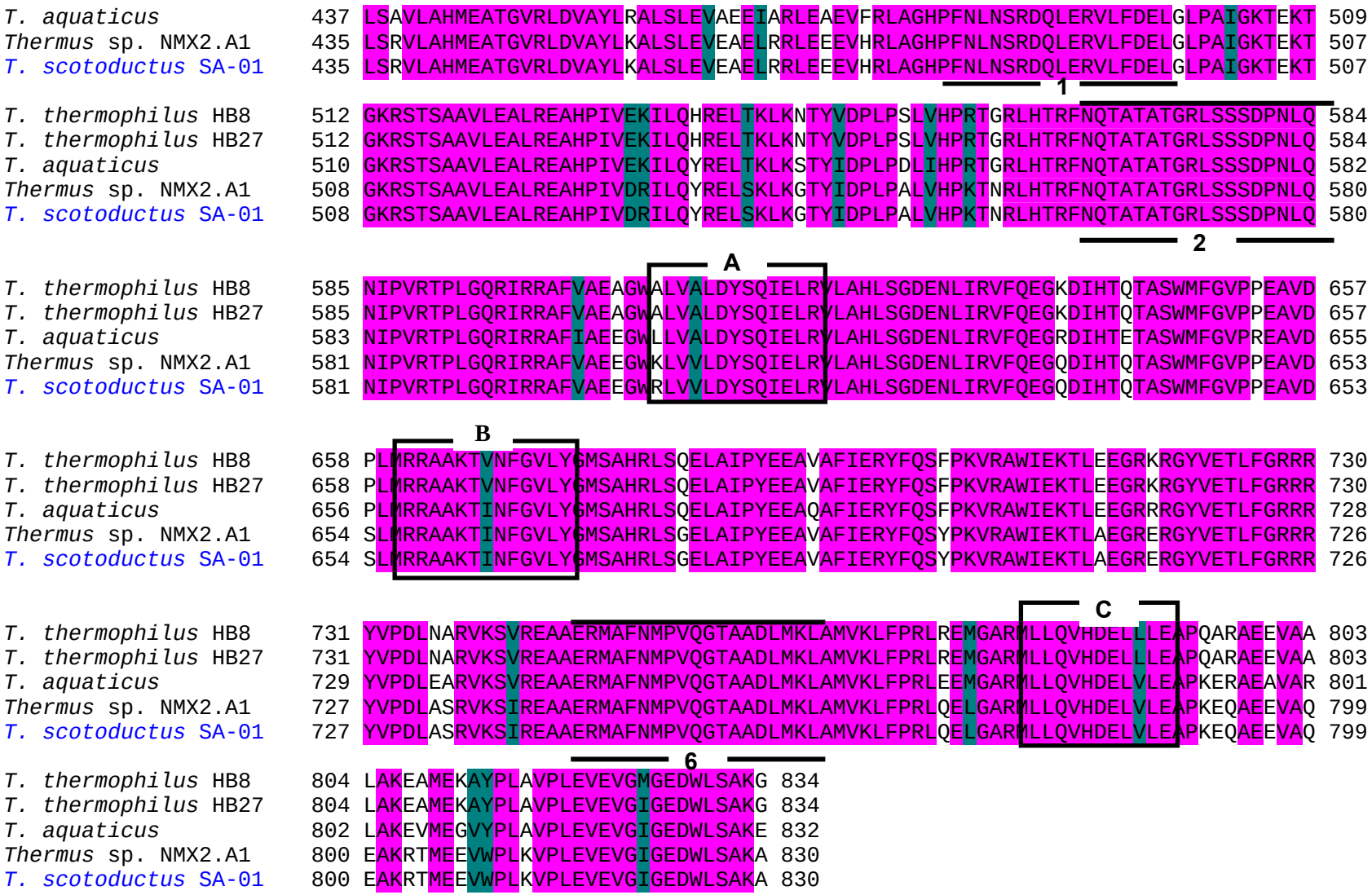


Fig 3.6 Multiple amino acid sequence alignments of thermostable DNA polymerase I protein with thermophilic bacteria. *T. scotoductus* SA-01 DNAPolI sequence obtained from draft genome annotation data. Other sequences used for alignments were obtained from GenBank and aligned using the DNAssist program. Description of similarity: Pink shaded blocks: 100% identity; green blocks: similarity under 80% and white blocks: similarity under 60%. Conserved amino acid regions are listed (1, 2 and 6) and motifs A, B and C are in highlighted in black boxes.

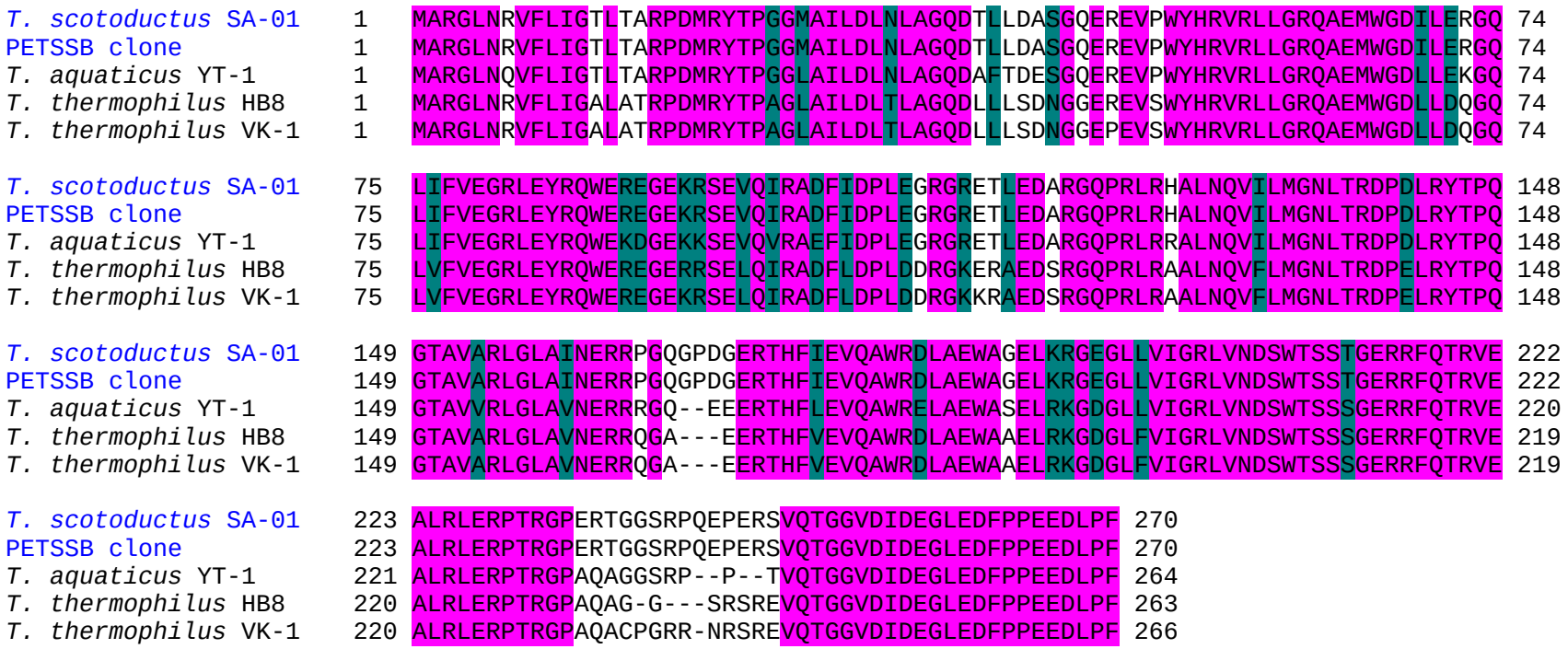


Fig 3.7 Multiple amino acid sequence alignments of thermostable SSB-like proteins with SSBs from thermophilic bacteria. *T. scotoductus* SA-01 SSB sequence obtained from draft genome annotation data and pETSSB sequence obtained from clone construct. Other sequences used for alignments were obtained from GenBank and aligned using the DNAssist program. Description of similarity: Pink shaded blocks: 100% identity; green blocks: similarity under 80% and white blocks: similarity under 60%.

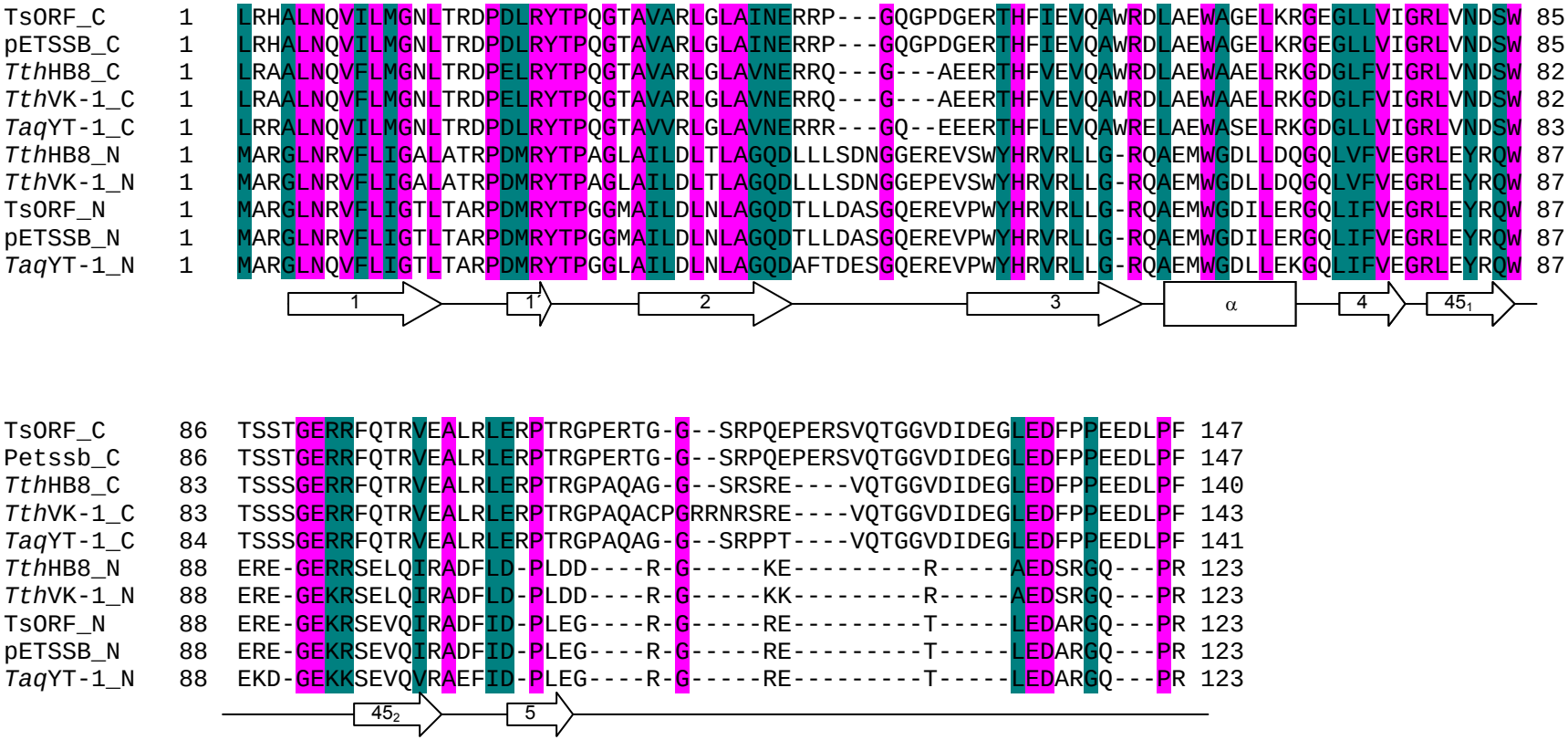


Fig 3.8 Multiple amino acid sequence alignment of thermostable SSB-like proteins with other SSBs showing the sequence similarity by dividing the N- and C-terminal fragments in order to highlight the OB fold regions. The *Taq*YT-1, *Tth*HB8, *Tth*VK-1 SSB proteins contain two OB folds each. The characteristic motifs that make up an OB fold are highlighted with open boxes/arrows and are numbered. The arrows, bar and lines show β-sheets, α-helix and loops, respectively identified in the structure of EcoSSB. The assignment of secondary structures is marked according to the OB fold rule (Murzin, 1993). Abbreviations:*Taq*YT-1 N or C: *T. aquaticus* YT-1, *Tth*HB8 N or C: *T. thermophilus* HB8, *Tth*VK-1 N or C: *T. thermophilus* VK-1, TsORF N or C: *T. scotoductus* SA-01 and pETSSB N or C: sequenced cloned SSB into pET28b.

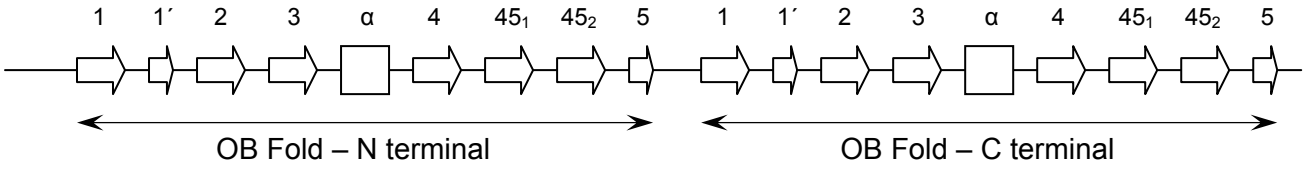


Fig 3.9 Schematic representation of the *T. scotoductus* SA-01 SSB protein highlighting the two OB fold regions present in the protein sequence.

3.3.3 Expression of the Recombinant pETpolI Protein

The constructs were used to transform *Escherichia coli* (JM109 DE3) host cells. The resultant transformants (carrying the correct recombinant plasmids) were then used for functional expression. The expression levels of the DNA polymerase did not result in any significant enzyme production using the *E. coli* (DE3) (data not shown). The expected ~100kDa band was not observable on a SDS-PAGE gel. This observation indicated that pETpolI possibly leaks toxic DNA polymerase I prior to IPTG induction (Dąbrowski *et al*., 2002). The level of expression was examined at different temperatures (42°C, 37°C and 25°C) and different IPTG concentrations (0.5 mM and 1 mM). However, no observable difference was noted. This was also observed by Lu and Erickson (1997), when expressing the *Pfu* DNA polymerase from *Pyrococcus furiosus*. The construct pETpfu was transformed into BL21(DE3), and many colonies were formed at 30°C but at 37°C no colonies were obtained. In addition, no expression was obtained from the colonies obtained at 30°C. This is surprising since the DNA polymerase is expected to be inactive at 37°C and has been easily expressed in *E. coli*. The authors suggest the use of a more stringent control of expression to permit the growth of the plasmid and expression of the DNA polymerase protein.

It was then decided to make use of the *E. coli* (DE3) pLysS cells for transformation and subsequent expression. The role of the pLysS or pLysE plasmid is to express the T7 lysozyme that strongly represses transcription from the T7 promoter both in non-induced and induced pET vectors; that stabilizes clones that leak toxic proteins (Dabrowski *et al*., 2002). The experiment was repeated but with the use of the *E. coli* BL21 (pLysS) cells as well as the additional 1% glucose in the media (as it is known to reduce background expression prior to induction with IPTG). Low level expression was observable after 4 hrs induction at 37°C with 1 mM IPTG on the SDS-PAGE gel with the use of the *E. coli* BL21 (pLysS) cells and

additional 1% glucose in the media (Fig 3.10). After induction, an approximately 97 kDa protein was produced in strain BL21(DE3) carrying the plasmid pLysS.

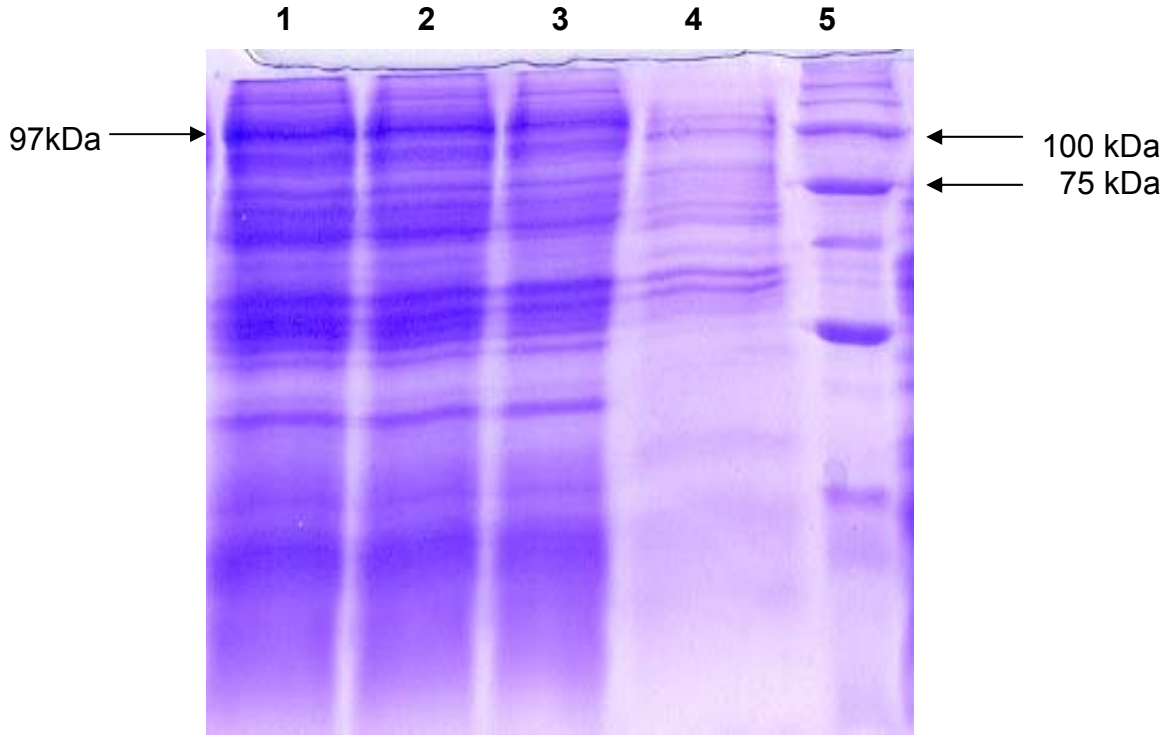


Fig 3.10 SDS-electrophoresis in 10% polyacrylamide gel of the *E. coli* cell extracts after expression of pETpolI constructs. Lanes 1-3: soluble protein cell extract from *E. coli* pLysS+pETDNAPolI clones; lanes 4: uninduced IPTG soluble protein cell extract from *E. coli* pLysS+pETpolI and lane 5: Precision Plus Protein Unstained Standard Marker (Biorad).

3.3.4 Recombinant DNA Polymerase I (His-Tag purification)

The recombinant protein was purified by taking advantage of the N-terminal poly(6)-histidine tag using Ni-affinity (Ni-NTA) chromatography. The DNA Polymerase I protein eluted as single activity peak (Fig 3.11) from the His-Trap FF (Amersham Biosciences) column through the use of linear imidazole concentration gradients. Fractions containing the DNA polymerase I protein, were then visualized by SDS-PAGE. Through the use of immobilised metal affinity chromatography (IMAC) and size-exclusion chromatography, the DNA polymerase I protein was purified from the soluble fraction from *E. coli* to near purity (Fig 3.12).

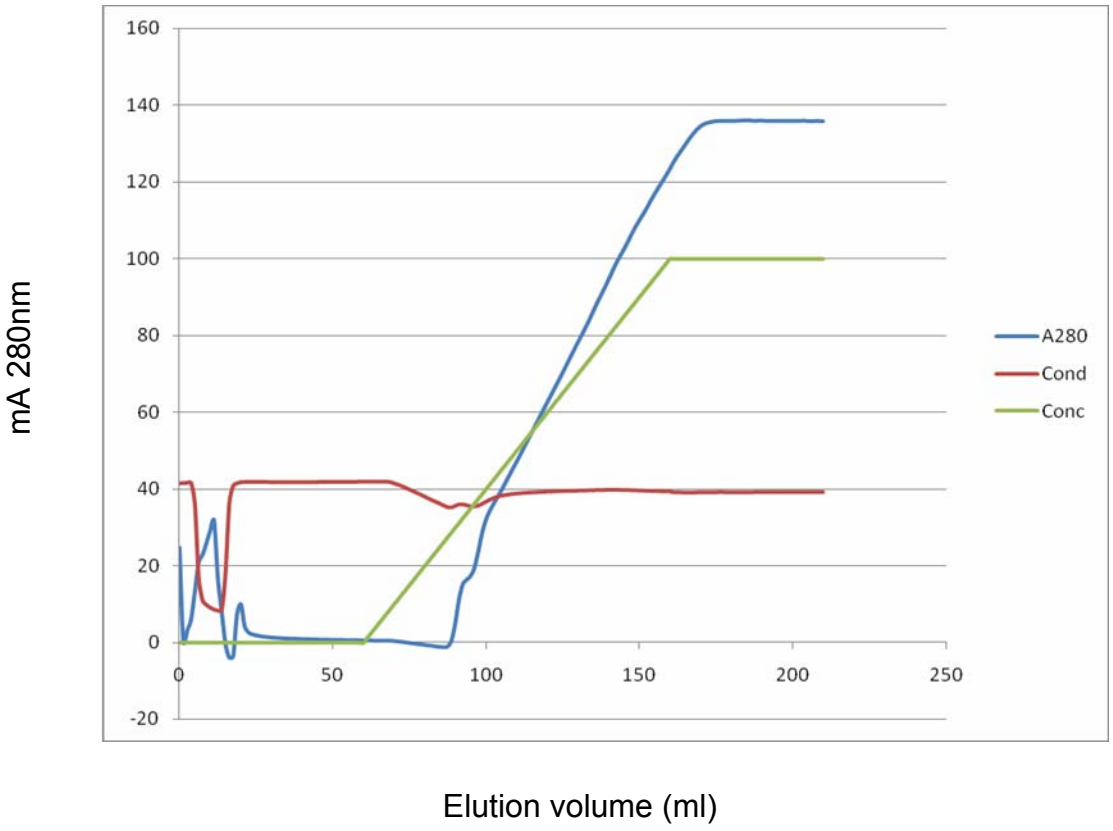


Fig 3.11 Purification of the recombinant soluble DNA polymerase I (DNAPolI) protein overproduced in *E. coli* through Ni-affinity chromatography.

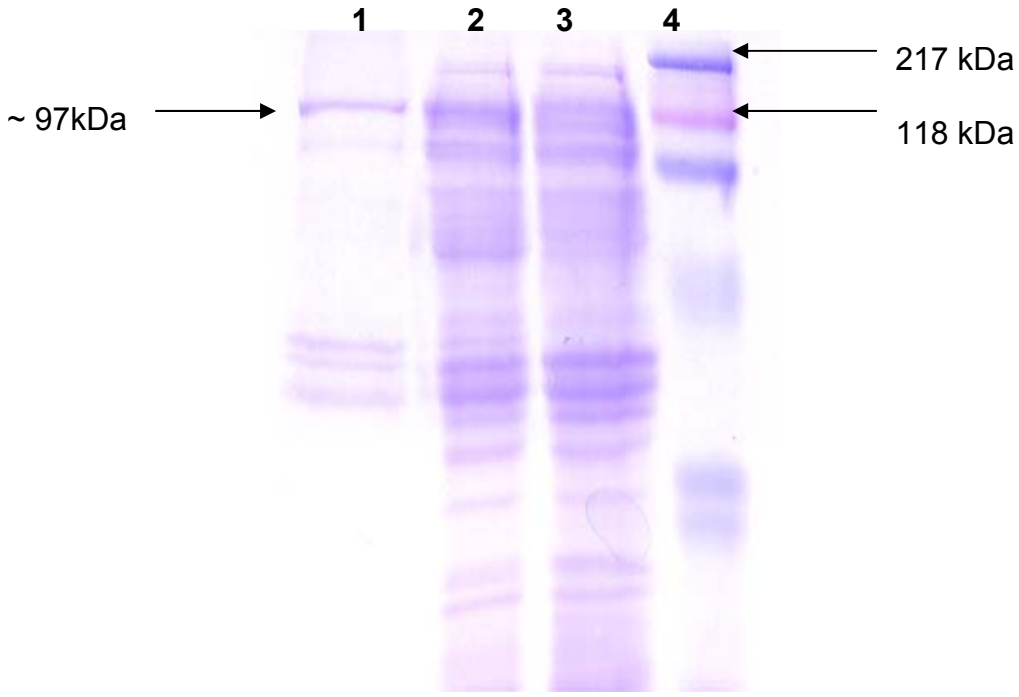


Fig 3.12 SDS−PAGE analysis of the partially purified DNA polymerase I from *Thermus scotoductus* SA-01. Lane 1: partially purified DNA polymerase I protein, lane 2: soluble protein cell extract from *E. coli* pLysS+pETDNAPolI clone, lane 3: uninduced IPTG soluble protein cell extract from *E. coli* pLysS+pETpolI and lane 4: Prestained Protein Marker.

The DNA polymerase protein was expressed as a soluble form in the cytosol. During the purification of the protein of the sonicated cells and His-tag purified proteins, we then used the thermophilic property of DNA polymerase proteins and eliminated most of the *E. coli* proteins by heating at 100°C for 5 mins and centrifuging to eliminate denatured proteins. However, several *E. coli* proteins still remained soluble after heating.

This necessitated the use of the size exclusion chromatography purification step. SDS-PAGE revealed a protein with an approx. molecular weight of 97 kDa, which agrees with the deduced *T. scotoductus* SA-01 DNA polymerase size calculated from the amino acid sequence. At this stage a significantly purified protein was obtained which could be used for a PCR activity assay.

The protein concentration determined using the BCA assay estimated the protein content at 700 $mg.ml^{-1}$.

3.3.5 DNA Polymerase Activity Assay

Enzymatic activity assays of the thermostable DNA polymerase was analysed by PCR titration from undiluted protein and 1:100, 1:200, 1:400, 1:800 and 1:1600 diluted protein concentrations. A standard PCR protocol for a PCR amplifying the internal 16S rRNA product of bacteria was used.

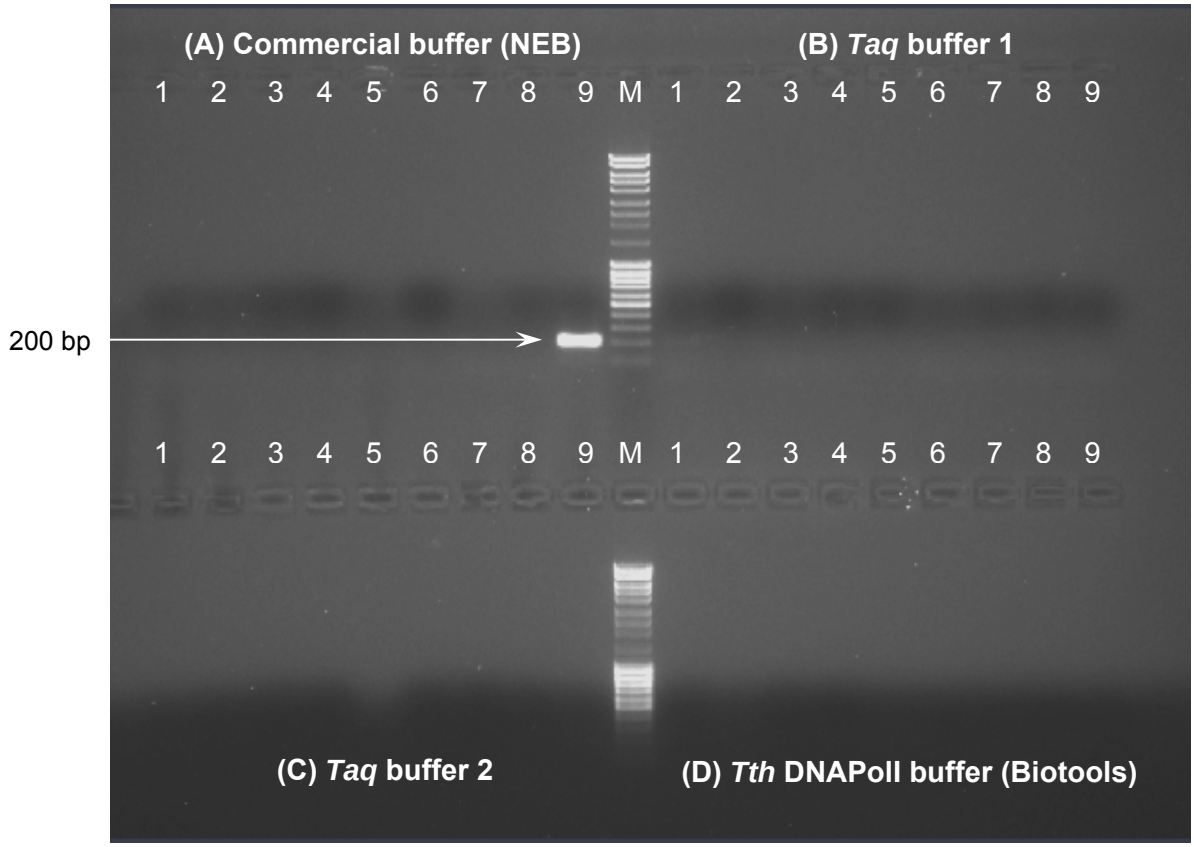


Fig 3.13 Agarose gel electrophoresis of partially purified DNA polymerase in the DGGE PCR titration. Gel A. Lanes 1: undiluted DNA polymerase protein, lanes 2-7: 1:10; 1:100; 1:200; 1:400; 1:800 and 1:1600 diluted DNA polymerase in commercial buffer (NEB), lane 8: negative control (dH_2O) and lane 9: positive control (commercial Taq (NEB). Gel B: same as Gel A however, using Taq Buffer 1 in PCR. Gel C: same as Gel A however, using Taq Buffer 2 in PCR and Gel D: same as Gel A however, using *Tth* DNA Poll buffer in PCR.

The partially purified DNA polymerase proved to be inactive as no PCR products were obtained using either undiluted or diluted DNA polymerase protein in a suitable buffer system. Titrations were performed as previous studies have shown that large amounts of

enzyme is purified from 100 ml culture of cells and excessively high levels of enzyme can inhibit the DNA polymerase reaction (Desai and Pfaffle, 1995). Optimization of the PCR was also unsuccessful in obtaining a PCR product (Fig 3.13). Attempts made using the standard DGGE PCR protocol, a Hotstart DGGE PCR and reducing the initial denaturation time also proved unsuccessful. Different buffer systems employed also indicated that providing the appropriate buffer system seems crucial in the PCR reaction as the positive control using the commercial Taq with the commercial buffer was the only product obtained. Employing an alternative buffer system with the commercial Taq proved unsuccessful.

Similar results were also found with the *Apy* DNA polymerase from the hyperthermophile *Aquifex pyrophilus* and *Aquifex aeolicus* (Choi and Kwon, 2004). Although a thorough characterization was performed on the purified protein, the authors noted that it is not clear whether the lower thermostability of the DNA polymerases reflects the nature of the enzymes *in vivo* or whether it reflects the difficulties of reproducing biological conditions *in vitro*.

3.3.6 Expression of the Recombinant pETSSB Protein

As with the DNA polymerase protein, the level of expression was examined with both the BL21(DE3) and BL21(DE3) pLysS strains. The SSB protein cloned into pET28b, did not give results after expression using the BL21(DE3) competent cells. These results suggest that the pETSSB leaks toxic *Tsc*SSB protein as was seen in the expression of the SSB protein from *T. aquaticus* (Dąbrowski *et al*., 2002), *T. thermophilus* (*Tth*SSB) (Perales *et al*., 2003) and *E. coli* (Dąbrowski and Kur, 1999). However, the use of BL21 pLysS competent cells did not result in a significant expression level, thus indicating that other expression systems that provide stringent control of expression should be employed. In addition, *Thermus thermophilus* is currently being studied as a putative host for expression of thermophilic enzyme complexes. Also, these extreme thermophiles have been employed successfully in the thermostabilisation of enzymes by functional selection at increasing temperatures (de Grado *et al*., 1999). This alternative host could also possibly help in the efficient expression and transformation of the SSB protein.

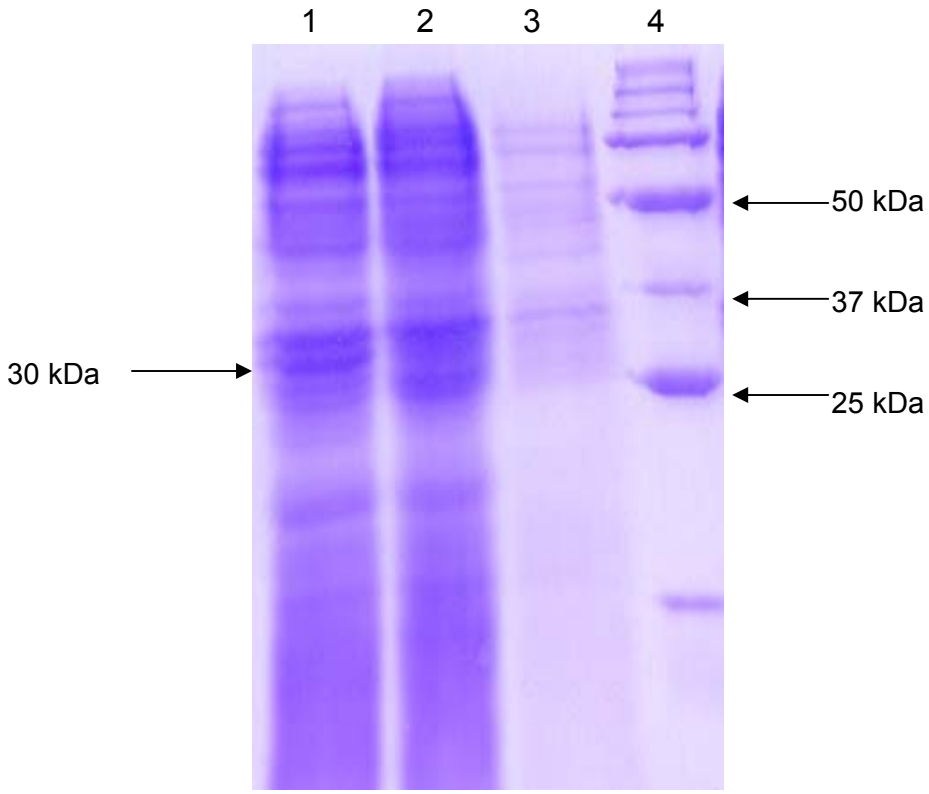


Fig 3.14 SDS-electrophoresis in 10% polyacrylamide gel of the *E. coli* cell extracts after expression of pETSSB constructs. Lane 1: soluble protein cell extract from *E. coli* pLysS+pETDNASSB clone; lanes 2: uninduced IPTG soluble protein cell extract from *E. coli* pLysS+pETpolI, lane 3: pET28b and lane 4: Precision Plus Protein Unstained Standard Marker (Biorad).

3.3.7 Recombinant SSB His-Tag purification

The SSB protein was expressed in a soluble form in the cytosol (Fig 3.14). The elution profile obtained from the His-tag purification of the protein indicated that a significant amount of protein was purified (Fig 3.15). However, this was not true as the SDS-PAGE gel indicated a non-specific binding by several other proteins.

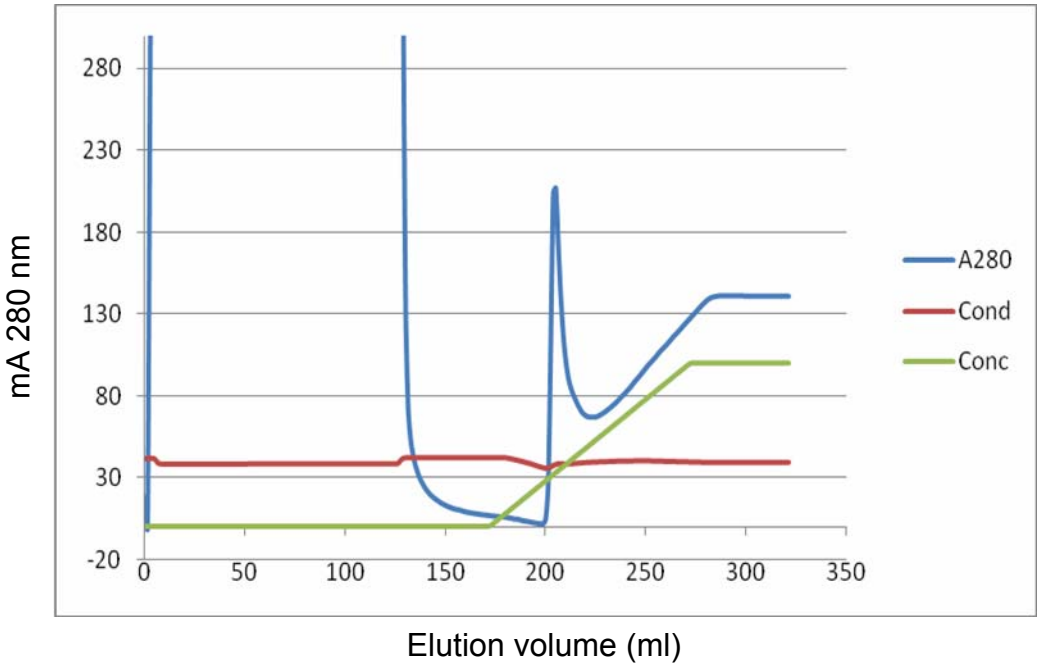


Fig 3.15 Purification of the recombinant soluble SSB protein overproduced in *E. coli* through the Ni-affinity column.

Size exclusion chromatography resulted in loss of the SSB protein. It was then decided to rather increase the imidazole concentration from an initial 20 mM to 30 mM concentration during the His-tag purification for more efficient elution of fewer proteins. However, fractions obtained by a higher-elution gradient from the His-tag column still contained many proteins (Fig 3.16a). The sample was further purified by putting it through the His-tag purification column a second time. Unfortunately, this resulted in the loss of all but one protein of the incorrect size (Fig 3.16b).

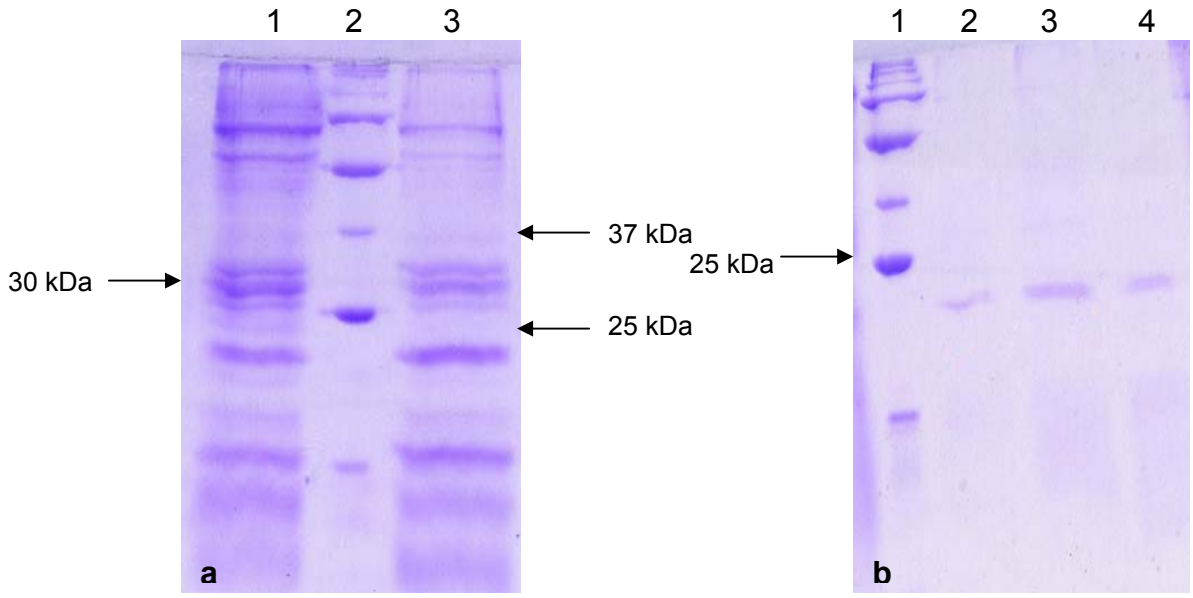


Fig 3.16 SDS-electrophoresis in 10% polyacrylamide gel of the *E. coli* cell extracts after purification through the Ni-affinity column and size-exclusion chromatography of pETSSB constructs. a): Lane 1 and 3: Fractions obtained after His-tag purification and lane 2: Precision Plus Protein Unstained Standard Marker (Biorad). b): Lane 1: Precision Plus Protein Unstained Standard Marker (Biorad) and lanes 2-4: fractions obtained after size-exclusion chromatography.

According to literature, some authors make use of a ssDNA cellulose column after His-tag purification in order for complete purification of the SSB protein (Chilukuri *et al*., 1997). Making use of a column specific for the properties of this particular protein may be crucial in obtaining a pure SSB protein.

3.4 Conclusion

While completing the genome sequence of the extremophile *T. scotoductus* SA-01 (this PhD study), preliminary BLAST results provided complete genes that could be used in molecular applications. Based on its probable thermostable properties, the DNA polymerase I and SSB proteins were chosen and thought to conceivably have a utility in a PCR reaction.

The pET system is one of the most powerful systems developed for cloning and expression of recombinant proteins in *E. coli* (Lu and Erickson, 1997). In this study, a construct was produced in the pET28b vector for expression of the DNA polymerase I and single stranded DNA binding protein in *E. coli.* However, it was found that this plasmid is either toxic or unstable in the expressing strain BL21(DE3), even in the absence of induction. The BL21(DE3) strain carrying a plasmid pLysS, which suppresses expression prior to induction with IPTG, was found to be more useful. Thus, a key to expression in this study appears to be to suppress the basal level of expression of the toxic protein prior to induction. Although many thermostable DNA polymerases have been successfully expressed using the BL21(DE3) pLysS cells, the *Thermus thermophilus* DNA polymerase has been expressed using the BL21 (DE3) cells alone (Moreno *et al*., 2005).

A partially purified DNA polymerase I protein was obtained even though the protein was unable to amplify a 200 bp PCR product. It seems likely that the *T. scotoductus* SA-01 DNA polymerase I may have a lower thermostability as compared to other commercially available DNA polymerases as its growth temperature is about 65°C (Kieft *et al*., 1999) and is less capable of withstanding temperatures used in a PCR reaction.

Unfortunately, the SSB protein was not expressed at significant levels and purification was unsuccessful. Thus, both thermostable proteins could not be employed to enhance a molecular technique such as PCR. Obtaining expression at higher levels and complete protein characterization would be required in order to understand the capabilities of these proteins.

4. Summary

Every genome that has been sequenced to date has provided new insight into the biological processes, activities and potential of these species that had not been evident before (Lioliou *et al*., 2004). In addition to its metal reduction, *T. scotoductus* SA-01 is particularly interesting for genomic analysis for several other reasons. Firstly, it was isolated from an extreme environment i.e. groundwater sampled from 3.2 km below the surface of a South African gold mine. Secondly, its extremophilic nature due to its isolation site being at an ambient temperature of 60°C (Kieft *et al*., 1999). Thirdly, its apparent use of nitrate, Fe(III), Mn(IV) or S^{o} as terminal electron acceptors to grow and its capability of reducing Cr(VI), U(VI), Co(III) and the quinone-containing compound anthraquinone-2,6-disulfonate (Balkwill *et al*., 2004). In addition, the *Thermus* sp. has high biotechnological potential.

This study employed the latest high-throughput DNA sequencing approach called pyrosequencing for the *de novo* sequencing of the *T. scotoductus* SA-01 genome. Initially, the GS20 pyrosequencing system, which produces 100 bp reads, was used to provide 27 Mbp of sequence data. However, complete assembly was not accomplished as 371 large contigs were obtained after assembly using the Newbler assembly software. Automatic annotation by TIGR (J. Craig Venter Institute) was performed on the GS20 contigs resulting in 3 084 ORFs. It was then decided to perform another 50 Mbp pyrosequencing run with the newly released (at the time) FLX pyrosequencing, which produces 200 bp reads. Assembly analysis was performed using GS20, FLX data and a combination of both GS20 and FLX pyrosequencing data with the updated version of the Newbler assembly software. Results revealed a dramatic increase in contig size and number when compared to each data set individually. Automatic annotation performed on the resulting 34 contigs after combining data sets resulted in a total of 2 458 ORFs. These results showed that a large number of frame shifts were present in GS20 data alone as the combined data revealed a decreased number of ORFs.

The draft genome sequence of *T. scotoductus* SA-01 was compared with that of its closest sequenced relative, *T. thermophilus* HB27, using the MUMmer software and the Artemis Comparison Tool (ACT). This program uses BLAST (either blastn or blastx) to compare two or more genomes for the arrangement of homologous genes (Carver *et al*., 2005). Sequence alignment indicated an extensive genome rearrangement as homologous regions were completely scattered. Synteny mapping illustrated that low levels of synteny exist between the two genomes and large sections of the genome are found in reverse orientation. This proved that mapping would not be helpful in obtaining the complete genome sequence due to the genome arrangement.

A genome sequencing strategy was employed to complete the genome of *T. scotoductus* SA-01 by employing Sanger sequencing. In order to determine the contig order, a fosmid library was constructed and clones sequenced from both ends and subsequently aligned to existing contigs. BLAST results of each contig end revealed that most gaps were probably due to the repeat regions; duplicate genes sequences as well as GC rich regions that were difficult to sequence through. Gaps closed were done by employing PCR using specific primers designed at end of contig and primer walking on fosmid clones.

Finally, the complete genome sequence of *T. scotoductus* SA-01 consists of 2 346 803 bp. However, after several attempts to close the last gap on the plasmid, the gap could not be closed. The draft plasmid sequence is 8 383 bp and consists of 12 ORFs. Automatic annotation revealed the chromosome sequence has 2 464 encoded ORFs. However, after using the Artemis software on the chromosome sequence for ORF correction, the number of ORFs increased to 2 482. Manual annotation was also performed on the original draft genome sequence to determine if automatic annotation function prediction was accurate.

Genome comparisons performed using BiBLAST highlighted several features of the genome sequence of *T. scotoductus* SA-01. The general metabolic pathways for *T. scotoductus* SA-01 seem to retain pathways for glycolysis, gluconeogenesis, pentose phosphate pathway, pyruvate dehydrogenase and tricarboxylic acid cycle (TCA) as does *T. thermophilus* HB27 and HB8. The genome sequence indicates that the electron transport system seems quite extensive in *T. scotoductus* SA-01.

This organism was also found be metabolically versatile, able to grow aerobically and anaerobically. This is probably due to the presence of the nitrate reductase gene. It is highly probable that the cytochrome C oxidase subunit cluster, electron transport protein SCO1/SenC and iron-sulfur proteins aid in the organisms growth in anaerobic conditions. In addition the genome comparisons have shown that the c-type cytochromes are also probably involved in metal reduction in *T. scotoductus* SA-01, however, none of the cytochromes present are similar in sequence to the c-type cytochromes found in the metal reducing organisms *G. sulfurreducens* and *S. oneidensis.*

BLAST results of the draft plasmid sequence (pTS01) against the complete chromosome sequence indicated that four ORFs present on the draft plasmid are also present in an identical copy (one or more than one copy) on the *T. scotoductus* SA-01 chromosome, providing evidence of genetic exchange between the chromosome and the extrachromosomal element. The genome re-arrangement observed in *T. scotoductus* SA-01

is probably mediated by the transposons present and can play a role in the genome plasticity.

Several genes of biotechnological applications are present in *T. scotoductus* SA-01. Two such genes were cloned and expressed for potential use in molecular applications. A partially purified DNA polymerase I protein was obtained however the protein was found to be non-functional in a PCR. Expression of the SSB was performed, but the protein could not be purified for further analysis. Obtaining expression at higher levels and complete protein characterization would be required in order to understand the capabilities of these proteins.

This study has highlighted the use of pyrosequencing in obtaining a high coverage of a genome sequence such as *T. scotoductus* SA-01. We expect that a comprehensive understanding of the mechanisms of metal reduction and general metabolism from the extremophile *T. scotoductus* SA-01 will arise from a combination of further comparative genomics of complete annotated genome data and prediction-driven experiments.

Key terms: pyrosequencing, annotation, complete genome, BiBLAST, metabolic pathway.

5. Opsomming

Elke organisme waarvan die DNS basispaaropeenvolging van die genoom bepaal is, het bygedra tot ons insig oor hierdie organismes se biologiese prosesse, aktiwiteit en potensiaal wat tot dusver onbekend was (Lioliou *et al*., 2004). Buiten die feit dat *T*. *scotoductus* SA-01 metale kan reduseer, is daar talle interessante redes waarom hierdie organisme se genomiese DNS basispaaropeenvolging bepaal moes word. Eerstens is hierdie organisme geïsoleer vanuit ‘n ekstreme omgewing naamlik uit grondwater wat 3.2 km onder die oppervlak van ‘n Suid-Afrikaanse goudmyn geneem is. Tweedens: hierdie organisme se ekstremofiliese natuur; dit is geïsoleer vanuit ‘n omgewing waarvan die temperatuur 60˚C is (Kieft *et al*., 1999). Derdens: organisme se skynbare vermoë om nitraat, Fe(III), Mn(IV) of S^0 as terminale elektron akseptor te gebruik vir groei en die vermoë om Cr(VI), U(VI), Co(III) en die kinoon-bevattende verbinding antrakinoon-2,6-disulfonaat se reduseer (Balkwill *et al*., 2004). Ten slotte het die *Thermus* sp. belowende biotegnologiese moontlikhede.

In hierdie studie is die nuutste en mees gevorderde DNS basispaar opeenvolging tegnologie (‘pyrosequencing’) gebruik vir die hoë kapasiteit, *de novo* DNS basispaar oopeenvolgingbepaling van *T*. *scotoductus* SA-01. Aanvanklik is die GS20-sisteem (GS20 ‘pyrosequencing’) gebruik wat 100 basispaar leeslentes geproduseer het en gevolglik 27 Mbp DNS basispaar opeenvolging data gelewer het. Die genoom kon nie volledig

saamgestel word nie, 371 groot saamgestelde dele (contigs) is verkry na samestelling met die Newbler programmatuur. Outomatiese anotering is gedoen by TIGR op die GS20 opeenvolgings wat 3 084 oop leesrame (ORFs) tot gevolg gehad het. ‘n 50 Mbp ‘pyrosequencing’ lopie was toe uitgevoer met die nuutste tegnologie (FLX ‘pyrosequencing’) op daardie tydstip, wat 200 basispaar leeslengtes opgelewer het. ‘n Analise van die samestellings wat gedoen is met GS20, FLX en ‘n kombinasie van beide die ‘pyrosequencing’ data stelle tesame met die opgedateerde weergawe van die Newbler sagteware, is gedoen. Resultate het getoon dat daar ‘n dramatiese toename was in die grootte en hoeveelheid saamgestelde dele, in vergelyking met die twee data stelle op hul eie. Outomatiese annotering is uitgevoer op 34 saamgestelde dele wat 2 458 ORFs tot gevolg gehad het. Die resultate het ook getoon dat daar ‘n groot hoeveelheid leesraam verskuiwings was in die GS20 data aangesien die gekombineerde data (GS20 en FLX stelle) ‘n afname in die getal ORFs getoon het.

Die eerste weergawe (draft) van die genoom van *T*. *scotoductus* SA-01 is vergelyk met die naasverwante DNS basispaaropeenvolging beskikbaar, naamlik die van *T*. *thermophilus* HB27 deur gebruik te maak van MUMmer program en die Artemis Comparison Tool (ACT). Die laasgenoemde program gebruik die BLAST funksies (blasn of blasx) om twee of meer genome se rangskikking van homoloë gene te vergelyk (Carver *et al*., 2005). Vergelyking van die twee genome met mekaar (volgorde belyning) het ‘n omvangryke herrangskikking van gene getoon aangesien die homoloë gebiede heeltemal verspreid was. Sintenie kartering het geillustreer dat lae vlakke van sintenie tussen die twee genome bestaan en dat groot gedeeltes van die een genoom in die omgekeerde oriëntasie gerangskik is. Hierdie bevinding het geïllustreer dat volgordesamestelling deur van ‘n templaat (mapping) nie van hulp sou wees om die hele genoom van *T*. *scotoductus* SA-01 se DNS basispaar opeenvolging te bepaal nie.

Die Sanger metode van DNS basispaaropeenvolgingsbepaling was gebruik om die genoom van *T*. *scotoductus* SA-01 te voltooi. Die volgorde van die saamgestelde dele is bepaal d.m.v. die konstruksie van ‘n fosmiedbiblioteek. Die basispaar opeenvolging van alle klone is bepaal in beide rigtings en is vergelyk met bestaande saamgestelde dele. BLAST resultate van elke saamgestelde deel se eindpunte het getoon dat die gebiede met gapings in die genoom te wydte was aan herhaldende volgordegedeeltes, gedupliseerde gene sowel as GC-ryke gebiede wat die basispaar opeenvolging bepalingsproses bemoeilik. Die gapings was gesluit d.m.v PKR (PCR) deur gebruik te maak van priemstukkie wat gebasseer was op die eindpunte van elke saamgestelde deel en ‘primer walking’ op fosmied klone.

Die finale en volledige chromosoom van *T*. *scotoductus* SA-01 bestaan uit 2 346 803 bp. Na vele pogings kon ‘n enkele gaping op die plasmied nie gesluit word nie. Gevolglik bestaan die huidige DNS basispaar opeenvolging van die plasmied uit 8 383 bp en 12 ORFs. Outomatiese anotering van die chromosoom toon 2 464 ORFs alhoewel die Artemis sagteware ‘n toename in die hoeveelheid ORFs getoon het naamlik 2 482. ‘n Anotering per hand is ook gedoen op die oorspronklike genoom se DNS basispaaropeenvolging om die outomatiese anotering se akuraatheid te verifieër.

Vergelyking van genome met mekaar met BiBLAST het talle kenmerkende eienskappe t.o.v. die *T*. *scotoductus* SA-01 genoom aangedui. Die algemene metaboliese weë van *T*. *scotoductus* SA-01 soos glikolise, glukoneogenese, die pentose fosfaat weg, pirovaat dehidrogenase en die Krebs siklus blyk behoue te gebly het – soos in die geval van *T*. *thermophilus* HB27 and HB8. Die genoom van *T*. *scotoductus* SA-01 toon ook dat hierdie organisme ‘n ekstensiewe elektron transport sisteem het.

Die organisme blyk ook om metabolies baie divers te wees aangesien dit die vermoë het om aerobies sowel as anaerobies te groei. Hierdie is waarskynlik te wydte aan die nitraat reduktase wat die organisme besit. Anaerobiese groei is ook hoogs waarskynlik te wydte aan die sitokroom C oksidase subeenheid groep, elektron transport proteïen SCO1/SenC en die yster-swael proteïene. Vergelyking van genome het getoon dat c-tipe sitokrome hoogs waarskynlik betrokke is by die redusering van metale in *T*. *scotoductus* SA-01 alhoewel geen van die sitokrome se DNS basispaar opeenvolging dieselfde lyk as die van sitokrome van metaal reduserende organismes soos bv. *G*. *sulfurreducens* en *S*. *oneidensis* nie.

BLAST resultate van die plasmied basispaaropeenvolging (pTS01) in vergelyking met die volledige genoom van *T*. *scotoductus* SA-01 het getoon dat daar vier ORFs (een of meer identiese kopiee) teenwoordig is op beide die plasmied sowel as die genoom. Hierdie dui op uitruiling van genetiese materiaal en die moontlikheid van ‘n ekstra-chromosomale element. Hierdie mag dalk die genoomherrangskikking van *T*. *scotoductus* SA-01 verduidelik en dat transposons teenwoordig is wat verantwoordelik is vir die genoom se plastisiteit.

Talle gene van biotegnologiese waarde is teenwoordig in die genoom van *T*. *scotoductus* SA-01. Twee gene is gekies vir klonering en heteroloë uitdrukking vanweë hul potensiële gebruik in molekulêre toepassings. ‘n Semi-gesuiwerde DNS-polimerase I proteien was verkry, maar dit het geen katalitiese aktiwiteit getoon tydens ‘n PKR nie. Uitdrukking van die SSB-proteien was geslaagd, maar die proteïen kon nie gesuiwer word vir verdere analise

nie. Hoër uitdrukking en volledige proteïenkaraktirisering word benodig om hierdie proteïene se katalitiese vermoëns beter te verstaan.

Hierdie studie het spesiale klem gelê op die gebruik van ‘pyrosequencing’ tegnologie vir die hoë deurvloei en ekstensiewe dekking van genoom DNS basispaar opeenvolging soos in die geval van *T*. *scotoductus* SA-01. Ons verwag dat ‘n dieper insig tot die meganismes van metaal reduksie en die algemene metabolisme van die ekstremofiel *T*. *scotoductus* SA-01 bekom sal word d.m.v. ‘n kombinasie van vergelykende genomika op ge-anoteerde genoom data en die van hipotese gedrewe eksperimente.

6. References

Ahmadian, A., M. Ehn and S. Hober. 2006. Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta* **363**: 83-94.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *The Journal of Molecular Biology* **215**: 403-410.

Aury, J-M., C. Cruaud, V. Barbe, O. Rogier, S. Mangenot, G. Samson, J. Poulain, V. Anthouard, C. Sacrpelli, F. Artiguenave and P. Wincker. 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**: 603.

Averhoff, B. 2004. DNA transport and natural transformation in mesophilic and thermophilic bacteria. *Journal of Bioenergetics and Biomembranes* **36(1)**: 25-33.

Balkwill, D.L., T.L. Kieft, T. Tsukuda, H.M. Kostandarithes, T.C. Onstott, S. Macnaughton, J. Bownas and J.K. Fredrickson. 2004. Identification of iron-reducing *Thermus* strains as *Thermus scotoductus*. *Extremophiles* **8**: 37-44.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero and P. Horvath. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709-1712.

Bentley, D.R. 2006. Whole-genome re-sequencing. *Current Opinion in Genetics and Development* **16**: 545-552.

Binnewies, T.T., Y. Motro, P.F. Hallin, O. Lund, D. Dunn, T. La, D.J. Hampson, M. Bellgard, T.M. Wassenaar and D.W. Ussery. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics* **6**: 165-185.

Blakesley, R.W., N.F. Hansen, J.C. Mullikin, P.J. Thomas, J.C. McDowell, B. Maskeri, A.C. Young, B. Benjamin, S.Y. Brooks, B.I. Coleman, J. Gupta, S-L. Ho, E.M. Karlins, Q.L. Maduro, S. Stantripop, C. Tsurgeon, J.L. Vogt, M.A. Walker, C.A. Masiello, X. Guan, NISC Comparative Sequencing Program, G.G. Bouffard and E.D. Green. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Research* **14**: 2235-2244.

Bonfield, J.K., K.F. Smith and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acid Research* **23(24)**: 4992-4999.

Bonetta, L. 2006. Genome sequencing in the fast lane. *Nature Methods* **3(2)**: 141-147.

Braslavsky, I., B. Hebert, E. Kartalov and S.R. Quake. 2003. Sequence information can be obtained from single DNA molecules. *PNAS* **100**: 3960-3964.

Brüggermann, H. and C. Chen. 2006. Comparative genomics of *Thermus thermophilus*: Plasticity of the megaplasmid and its contribution to a thermophilic lifestyle. Journal of Biotechnology **124**: 654-661.

Carver, T.J., K.M. Rutherford, M. Berriman, M-A. Rajandream, B.G. Barrell and J. Parkhill. 2005. ACT: the Artemis comparison tool. *Bioinformatics Application Note* **21(16)**: 3422-3423.

Carver, T., N. Thomson, A. Bleasby, M. Berriman and J. Parkhill. 2009. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* **25(1)**: 119-120.

Chan, E.Y. 2005. Advances in sequencing technology. *Mutation Research* **573**: 13-40.

Chang, J.R., J.J. Choi, H-K. Kim and S-T. Kwon. 2001. Purification and properties of *Aquifex aeolicus* DNA polymerase expressed in *Escherichia coli*. *FEMS Microbiology Letters* **201**: 73-77.

Chen, K. and L. Pachter. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PloS Computational Biology* **1(2)**: e24.

Chilukuri, L.N. and D.H. Bartlett. 1997. Isolation and characterization of the gene encoding single-stranded-DNA-binding protein (SSB) from four marine *Shewanella* strains that differ in their temperature and pressure optima for growth. *Microbiology* **143**: 1163-1174.

Chivian, D., E.L. Brodie, E.J. Alm, D.E. Culley, P.S. Dehal, T.Z. DeSantis, T.M. Gihring, A. Lapidus, Li-H. Lin, S.R. Lowry, D.P. Moser, P.M. Richardson, G. Southam, G. Wanger, L.M. Pratt, G.L. Andersen, T.C. Hazen, F.J. Brockman, A.P. Arkin and T.C. Onstott. 2008. Environmental Genomics Reveals a Single-Species Ecosystem Deep Within Earth. *Science* **322(5899)**: 275–278.

Choi, J.J., H-K. Kim and S-T. Kwon. 2001. Purification and characterisation of the 5´→3´ exonuclease domain-deleted *Thermus filiformis* DNA polymerase expressed in *Escherichia coli*. *Biotechnology Letters* **23**: 1647-1652.

Choi, J.J. and S-T Kwon. 2004. Cloning, expression and characterisation of DNA polymerase from hyperthermophilic bacterium *Aquifex pyrophilus*. *Journal of Microbial Biotechnology* **14(5)**: 1022-1030.

Dąbrowski, S., M. Olszewski, R. Piątek, A. Brillowska-Dąbrowska, G. Konopa and J. Kur. 2002. Identification and characterization of single-stranded-DNA-binding proteins from *Thermus thermophilus* and *Thermus aquaticus* – new arrangement of binding domains. *Microbiology* **148**: 3307-3315.

Dąbrowski, S., M. Olszewski, R. Piątek and J. Kur. 2002. Novel thermostable ssDNA-binding proteins from *Thermus thermophilus* and *T. aquaticus* – expression and purification. *Protein Expression and Purification* **26**: 131-138.

Dąbrowski, S. and J. Kur. 1999. Cloning, Overexpression and Purification of the recombinant His-tagged SSB protein of *Escherichia coli* and use in polymerase chain reaction amplification. *Protein Expression and Purification* **16**: 96-102.

Daraselia, N., D. Dernovoy, Y. Tian, M. Borodovsky, R. Tatusov and T. Tatusova. 2003. Reannotation of *Shewanella oneidensis* genome. *OMICS A Journal of Integrative Biology* **7(2)**: 171-175.

Deckert, G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay, R. Huber, R.A. Feldman, J.M. Short, G.J. Olsen and R.V. Swanson. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**: 353-358.

de Grado, M., P. Castán and J. Berenuer. 1999. A high-transformation-efficiency cloning vector for *Thermus thermophilus*. *Plasmid* **42**: 241-245.

Delarue, M., O. Poch, N. Tordo, D. Moras and P. Argos. 1990. An attempt to unify the structure of polymerases. *Protein Engineering* **3(6)**: 461-467.

Delcher, A.L., S. Kasif, R.D. Fleischmann, J. Peterson, O. White and S.L. Salzberg.1999. Alignment of whole genomes. *Nucleic Acids Research* **27(11)**: 2369-2376.

Delcher, A.L., A. Phillippy, J. Carlton and S.L. Salzberg. 2002. Fast algorithms for large-scale alignment and comparison. *Nucleic Acids Research* **30(11)**: 2478-2483.

DeLong, E.F. 2005. Microbial community genomics in the ocean. *Nature Reviews* **3**: 459-469.

Desai, U.J. and P.K. Pfaffle. 1995. Single-step purification of a thermostable DNA polymerase expression in *Escherichia coli*. *BioTechniques* **19**: 780-784.

Edwards, R.A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D.M. Peterson, M.O. Saar, S. Alexander, E.C. Alexander Jr. and F. Rohwer. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**: 57.

Eggington, J.M., N. Haruta, E.A. Wood and M.M. Cox. 2004. The single-stranded DNA-binding protein of *Deinococcus radiodurans*. *BMC Microbiology* **4**: 2.

Ewing B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**: 186-194.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J.D. Gocayne, J. Scott, R. Shirley, L. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillipps, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith and J.C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269(5223)**: 496-512.

Franguel, L., K.E. Nelson, C. Buchrieser, A. Danchin, P. Glaser and F. Kunst. 1999. Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**: 2625-2634.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischman, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, R.D. Fritchman, J.F. Weidman, K.V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T.R. Utterback, D.M. Saudek, C.A. Phillips, J.M.

Merrick, J.F. Tomb, B.A. Dougherty, K.F. Bott, P.C. Hu, T.S. Lucier, S.N. Peterson, H.O. Smith, C.A. Hutchison and J.C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium. Science* **270(5235):** 397-403.

Fraser, C.M. J.A. Eisen, K.E. Nelson, I.T. Paulsen and S.L. Salzberg. 2002. The value of complete microbial genome sequencing (you get what you pay for). *Journal of Bacteriology* **184(23)**: 6403-6405.

Fraser-Liggett, C.M. 2005. Insights on biology and evolution from microbial genome sequencing. *Genome Research* **15**: 1603-1610.

Fredrickson, J.K. and M.F. Romine. 2005. Genome-assisted analysis of dissimilatory metal-reducing bacteria. *Current Opinion in Biotechnology* **16**: 269-274.

Forterre, P. 2002. A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends in Genetics* **18(5)**: 236-237.

Friedrich, A., J. Rumszauer, A. Henne and B. Averhoff. 2003. Pilin-like proteins in the extremely thermophilic bacterium *Thermus thermphilus* HB27: Implication in competence for natural transformation and links to Type IV pilis biogenesis. *Applied and Environmental Microbiology* **69(7)**: 3695-3700.

Friedrich, A., T. Hartsch and B. Averhoff. 2001. Natural transformation in mesophilic and thermophilic bacteria: Identification and characterization of novel, closely related competence genes in *Acinetobacter* sp. strain BD413 and *Thermus thermphilus* HB27*Applied and Environmental Microbiology* **67(7)**: 3140-3148.

Gharizadeh, B., Z.S. Herman, R.G. Eason, O. Jejelowo and N. Pourmand. 2006. Large-scale pyrosequencing of synthetic DNA: A comparison with results from Sanger dideoxy sequencing. *Electrophoresis* **27**: 3042-3047.

Goldberg, S.M.D., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S.A. Kravitz, F.M. Lauro, K. Li, Y-H Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier and J.C. Venter. 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *PNAS* **103(30)**: 11240-11245.

Green, R.E., J. Krause, S.E. Ptak, A.W. Briggs, M.T. Ronan, J.F. Simons, L. Du, M. Egholm, J.M. Rothberg, M. Paunovic and S. Pääbo. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**: 330-336.

Haft, D.H., J.D. Selengut, L.M. Brinkac, N. Zafar and O. White. 2005. Genome Properties: A system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* **21(3)**: 293-306.

Hall, N. 2007. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology* **209**: 1518-1525.

Havlak, P., R. Chen, K.J. Durbin, A. Egan, Y. Ren, X-Z. Song, G.M. Weinstock and R.A. Gibbs. 2004. The Atlas genome assembly system. *Genome Research* **14**: 721-732.

Heidelberg, J.F., I.T. Paulsen, K.E. Nelson, E.J. Gaidos, W.C. Nelson, T.D. Read, J.A. Eisen, R. Seshadri, N. Ward, B. Methe, R.A. Clayton, T. Meyer, A. Tsapin, J. Scott, M. Beanan, L. Brinkac, S. Daugherty, R.T. DeBoy, R.J. Dobson, A.S. Durkin, D.H. Haft, J.F. Kolonay, R. Madupu, J.D. Peterson, L.A. Umayam, O. White, A.M. Wolf, J. Vamathevan, J. Weidman, M. Impraim, K. Lee, K. Berry, C. Lee, J. Mueller, H. Khouri, J. Gill, T.R. Utterback, L.A. McDonald, T.V. Feldblyum, H.O. Smith, J.C. Venter, K.H. Nealson and C.M Fraser. 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis. Nature Biotechnology* 20: 1118-1123.

Henne, A., H. Brüggermann, C. Raasch, A. Wiezer, T. Hartsch, H. Liesegang, A. Johann, T. Lienard, O. Gohl, R. Martinez-Arias, C. Jacobi, V. Starkuviene, S. Schlenczeck, S. Dencker, R. Huber, H-P. Klenk, W. Kramer, R. Merkl, G. Gottschalk and H-J. Fritz. 2004. The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nature Biotechnology* **22(5)**: 547-553.

Hongou, Y., V.K. Sharma, T. Prakash, S. Noda, T.D. Taylor, T. Kudo, Y. Sakaki, A. Toyoda, M. Hattori and M Ohkuma. 2008. Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *PNAS* **105(14)**: 5555-5560.

Huang, X., J. Wang, S. Aluru, S-P Yang and L. Hillier. 2003. PCAP: A whole-genome assembly program. *Genome Research* **13**: 2164-2170.

Huse, S.M., J.A. Huber, H.G. Morrison, M.L. Sogin and D.M. Welch. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**: R143.

Jaffe, D.B., J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J.P. Mesirov, M.C. Zody and E.S. Lander. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Research* **13**: 91-96.

Jansen, R., J.D.A. van Embden, W. Gaastra and L.M. Schouls. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* **43(6)**: 1565-1575.

Jeong, H. and J.F. Kim. 2008. An optimised strategy for genome assembly of Sanger/pyrosequencing hybrid data using available software. *Genomics and Informatics* **6(2)**: 88-90.

Keith, J.M., D.A.E. Cochran., G.H. Lala., P. Adams, D. Bryant and K.R. Mitchelson. 2004. Unlocking hidden genomic sequence. *Nucleic Acid Research* **32(3)**: e35.

Kieft, T.L., J.K. Fredrickson, T.C. Onstott, Y.A. Gorby, H.M. Kostandarithes, T.J. Bailey, D.W. Kennedy, S.W. Li, A.E. Plymale, C.M. Spadoni and M.S. Gray. 1999. Dissimilatory reduction of Fe(III) and other electron acceptors by a *Thermus* isolate. *Applied and Environmental Microbiology* **65(3)**: 1214-1221.

Kieleczawa, J. 2006. Fundamentals of Sequencing of Difficult Templates—An Overview. *Biomol Tech.* **17(3)**: 207–217.

Kim, Y.J., H.S. Lee, S.S Bae, J.H. Jeon, J.K. Lim, Y. Cho, K.H. Nam, S.G. Kang, S-J. Kim, S-T. Kwon and J-H. Lee. 2007. Cloning, purification and characterisation of a new DNA polymerase from a hyperthermophilic archaeon, *Thermococcus* sp. NA1. *J. Microbial. Biotechnol.* **17(7)**: 1090-1097.

Kur, J., M. Olszewski, A. Długołęcka and P. Filipkowski. 2005. Single-stranded DNA-binding proteins (SSBs) – sources and applications in molecular biology. *Acta Biochimica Polonica* **52(3)**: 569-574.

Kurtz, S., A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S.L.Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome Biology* **5**: R12.

Kwon, S-T., J.S. Kim, J.H. Park, H.K. Kim and D-S. Lee. 1997. Cloning and analysis of the DNA polymerase-encoding gene from *Thermus caldophilus* GK24. *Mol. Cells* **7(2)**: 264-271.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680−685.

Lane, D.J. (1991) 16S/23S rRNA sequencing In: Nucleic Acid Techniques in Bacterial Systematics (Stackebrandt, E. and Goodfellow, M. Eds), pp. 115-175. Wiley, Chichester, UK.

Langaee, T. and M. Ronaghi. 2005. Genetic variation analyses by pyrosequencing. *Mutation Research* **573**: 96-102.

Latreille, P., S. Norton, B.S. Goldman, J. Henkhaus, N. Miller, B. Barbazuk, H.B. Bode, C. Darby, Z. Du, S. Forst, S. Gaudriault, B. Goodner, H. Goodrich-Blair and S. Slater. 2007. Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC Genomics* **8**: 321.

Lawyer, F.C., S. Stoffel, R.K. Saika, K. Myambo, R. Drummond and D.H. Gelfand. 1989. Isolation, characterisation and expression in *Escherichia coli* of the DNA polymerase gene from *Thermus aquaticus*. The Journal of Biological Chemistry **264(11)**: 6427-6437.

Lee, J.E., R.J. Potter and D. Mandelman. 2007. World Intellectual Property Organization. WO/2007/029200.

Lioliou, E.E., A.A. Pantazaki and D.A. Kyriakidis. 2004. *Thermus thermophilus* genome analysis: benefits and implications. Microbial Cell Factories **3**: 5.

Lister, R., B.D. Gregory and J.R. Ecker. 2008. Next is now: new technologies for sequencing of genomes, transcriptomes and beyond. *Current Opinion in Plant Biology* **12**: 1-12.

Lu, C. and H.P. Erickson. 1997. Expression in *Escherichia coli* of the thermostable DNA polymerase from *Pyrococcus furiosus*. *Protein Expression and Purification* **11**: 179-184.

Makarova, K.S., L. Aravind, Y.I. Wolf, R.L. Tatusov, K.W. Minton, E.V. Koonin and M.J. Daly. 2001. Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol. Mol. Biol. Rev.* **65(1)**: 44-79.

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

Marshall, M.J., A.S. Beliaev, A.C. Dohnalkova, D.W. Kennedy, L. Shi, Z. Wang, M.I. Boyanov, B. Lai, K.M. Kemner, J.S. McLean, S.B. Reed, D.E. Culley, V.L. Bailey, C.J. Simonson, D.A. Saffarini, M.J. Romine, J.M. Zachara and J.K. Fredrickson. 2006. c-Type Cytochrome-dependent formation of U(VI) nanoparticles by *Shewanella oneidensis*. *PLoS Biology* **4(8)**: e268.

Methe, B.A., K.E. Nelson, J.A. Eisen, I.T. Paulsen, W. Nelson, J.F. Heidelberg, D. Wu, M. Wu, N. Ward, M.J. Beanan, R.J. Dodson, R. Madupu, L.M. Brinkac, S.C. Daugherty, R.T. DeBoy, A.S. Durkin, M. Gwinn, J.F. Kolonay, S.A. Sullivan, D.H. Haft, J. Selengut, T.M. Davidsen, N. Zafar, O. White, B. Tran, C. Romero, H.A. Forberger, J. Weidman, H. Khouri, T.V. Feldblyum, T.R. Utterback, S.E. Van Aken, D.R. Lovely and C.M. Fraser. 2003. Genome of *Geobacter sulfurreducens*: Metal reduction in subsurface environments. *Science* **302**: 1967-1969.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Research* **15**: 1767-1776.

Meyer, F., A. Goesmann, A.C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich and A. Pühler. 2003. GenDB-an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research* **31(8)**: 2187-2195.

Miller, J.R., A.L. Delcher, S. Koren, E. Venter, B.P. Walenz, A. Brownley, J. Johnson, K. Li, C. Mobarry and G. Sutton. 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24(24)**: 2818-2824.

Möller, C. and E. van Heerden. 2006. Isolation of a soluble and membrane-associated Fe(III) reductase from the thermophile, *Thermus scotoductus* (SA-01). *FEMS Microbiol. Lett.* **265**: 237-243.

Moreira, L.M. and Sá-Correia. 1997. Megaplasmids in *Thermus oshimai* isolates from two widely separated geographical areas: restriction fragment profiling and DNA homology. *Arch. Microbiol* **168**: 473-479.

Moreira, L.M., M.S. Da Costa and Sá-Correia. 1997. Comparative genomic analysis of isolates belonging to the six species of the genus *Thermus* using pulse-field gel electrophoresis and ribotyping. *Arch. Microbiol* **168**: 92-101.

Moreno, R., A. Haro, A. Castellanos and J. Berenguer. 2005. High-level overproduction of His-tagged *Tth* DNA polymerase in *Thermus thermophilus*. *Applied and Environmental Microbiology* **71(1)**: 591-593.

Murzin, A.G. 1993. OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* **12**: 861-867.

Nierman, W.C., D. DeShazer, H.S. Kim, H. Tettelin, K.E. Nelson, T. Feldblyum, R.L. Ulrich, C.M. Ronning, L.M. Brinkac and S.C. Daugherty. 2004. Structural flexibility in *Burkholderia mallei* genome. *Proc. Natl. Acad. Sci* **101**: 14246-14251.

Noonan, J.P., G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J.K. Pritchard and E.M. Rubin. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* **314**: 1113-1118.

Okamoto, S. and M. Ohmori. 2002. The cyanobacterial PilT protein responsible for cell motility and transformation hydrolyzes ATP. *Plant Cell Physiol.* **43(10)**: 1127-1136.

Omelchenko, M.V., Y.I. Wolf, E.K. Gaidamakova, V.Y. Matrosova, A. Vasilenko, M. Zhai, M.J. Daly, E.V. Koonin and K.S. Makarova. 2005. Comparative genomics of *Thermus*

thermophilus and *Deinococcus radiodurans*: divergent routes of adaptation to thermophily and radiation resistance. BMC Evolutionary Biology **5**: 57.

Opperman, D.J. and E. van Heerden. 2007. Aerobic Cr(VI) reduction by *Thermus scotoductus* strain SA-01. *Journal of Applied Microbiology* **103**: 1907-1913.

Opperman, D.J. and E. van Heerden. 2008. A membrane-associated protein with Cr(VI)-reducing activity from *Thermus scotoductus* SA-01. *FEMS Microbial. Lett.* **208**: 210-218.

Opperman, D.J., L.A. Piater and E. van Heerden. 2008. A novel chromate reductase from *Thermus scotoductus* SA-01 related to old yellow enzyme. *Journal of Bacteriology* **190(8)**: 3076-3082.

Oshima, T. and K. Imahori. 1974. Description of *Thermus thermophilus* (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. *International Journal of Systematic Bacteriology* **24(1)**: 102-112.

Overbeek, T. Begley, R.M. Butler, J.V. Choudhuri, H-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E.D. Frank, S. Gerdes, E.M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jenson, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A.C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G.D. Pusch, D.A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko and V. Vonstein. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* **33(17)**: 5691-5702.

Pantazaki, A.A., A.A. Pritsa and D.A. Kyriakidis. 2002. Biotechnologically relevant enzymes from *Thermus thermophilus*. *Appl. Microbiol. Biotechnol* **58**: 1-12.

Park, J.H., J.S. Kim, S-T. Kwon and D-S. Lee. 1993. Purification and characterisation of *Thermus caldophilus* GK24 DNA polymerase. *Eur. J. Biochem.* **214**: 135-140.

Park, J.H., B.C. Park, S.H. Koch, J.S. Kim, J.H. Koh, M.H. Yang, Y.S. Kim, C.H. Kim, M.H. Kim, S.T. Kwon and D-S. Lee. 2003. Genome mapping of an extreme thermophile, *Thermus caldophilus* GK24. *Genomics and Informatics* **1(1)**: 50-54.

Park, H-S., K.J. Kayser, J-H. Kwak and J.J. Kilbane II. 2004. Heterologous gene expression in *Thermus thermophilus*: β-galactosidase, dibenzothiophene monooxygenase, PNB carboxy esterase, 2-aminobiphenyl-2,3-diol dioxygenase and chloramphenicol acetyl transferase. *J. Ind. Microbiol. Biotechnol.* **31**: 189-197.

Patel, P.H., M. Suzuki, E. Adman, A. Shinkai and L.A. Loeb. 2001. Prokaryotic DNA polymerase I: Evolution, Structure and "Base Flipping" mechanism for nucleotide selection. *J. Mol. Biol.* **308**: 823-837.

Patterton, H-G. and S. Graves. 2000. DNAssist: the integrated editing and analysis of molecular biology sequences in Windows. *Bioinformatics Information Note* **16(7)**: 652-653.

Perales, C., F. Cava, W.J.J. Meijer and J. Berenguer. 2003. Enhancement of DNA, cDNA synthesis and fidelity at high temperatures by a dimeric single-stranded DNA-binding protein. *Nucleic Acids Research* **31(22)**: 6473-6480.

Pfiffner, S.M., J.M. Cantu, A. Smithgall, A.D. Peacock, D.C. White, D.P. Moser, T.C. Onstott and E. van Heerden. 2006. Deep subsurface microbial biomass and community structure in Witwatersrand basin mines. *Geomicrobiology Journal* **23**: 431-442.

Poinar, H.N., C. Schwarz, J. Qi, B. Shapiro, R.D.E. MacPhee, B. Buigues, A. Tikhonov, D.H. Huson, L.P. Tomsho, A. Auch, M. Rampp, W. Miller and S.C. Schuster. 2006. Metagenomics to Paleogenomics: Large-scale sequencing of mammoth DNA. *Science* **311**: 392-394.

Pop, M., A. Phillipy, A.L. Delcher and S.L. Salzberg. 2004. Comparative genome assembly. *Briefing in Bioinformatics* **5(3)**: 237-248.

Quinn, N.L. N. Levenkova, W. Chow, P. Bouffard, K.A. Boroevich, J.R. Knight, T.P. Jarvie, K.P. Lubieniecki, B.A. Desany, B.F. Koop, T.T. Harkins and W.S. Davidson. 2008. Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* **9**: 404.

Ramírez-Arcos, S., L.A. Fernández-Herrero and J. Berenguer. 1998. A thermophilic nitrate reductase is responsible for the strain specific anaerobic growth of *Thermus thermophilus* HB8. *Biochimica et Biophysica Acta* **1396**: 215-227.

Roh, Y., H. Gao, H. Vali, D.W Kennedy, Z.K. Yang, W. Gao, A.C. Dohnalkova, R.D. Stapleton, J-W. Moon, T.J. Phelps, J.K. Fredrickson and J. Zhou. 2006. Metal reduction and iron biomineralization by a psychrotolerant Fe(III)-reducing bacterium, *Shewanella* sp. Strain PV-4. *Applied and Environmental Microbiology* **72(5)**: 3236-3244.

Ronaghi, M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Research* **11**: 3-11.

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

Salzberg, S.L. and J.A. Yorke. 2005. Beware of mis-assembled genomes. *Bioinformatics* **21(24)**: 4320-4321.

Scheibye-Alsing, K., S. Hoffmann, A. Frankel, P. Jensen, P.F. Stadler, Y. Mang, N. Tommerup, M.J. Gilchrist, A-B. Nygård, S. Cirera, C.B. Jørgensen, M. Fredholm and J. Gorodkin. 2009. Sequence Assembly. *Computational Biology and Chemistry* **33**: 12-136.

Schmidt, C.L. 2004. Rieske Iron–Sulfur Proteins From Extremophilic Organisms. *Journal of Bioenergetics and Biomembranes.* **36(1)**: 107-113.

Schuster, S.C. and G. Gottschalk. 2005. Microbial genomics in its second decade. *Current Opinion in Microbiology* **8**: 561-563.

Schwarzenlander, C. and B. Averhoff. 2006. Characterization of DNA transport in the thermophilic bacterium *Thermus thermophilus* HB27. *FEBS Journal* **273**: 4210-4218.

Seedorf, H., W.F. Fricke, B. Veith, H. Brüggemann, H. Liesegang, A. Strittmatter, M. Miethke, W. Buckel, J. Hinderberger, F. Li, C. Hagemeier, R.K. Thauer and G. Gottschalk. 2008. The genome of *Clostridium kluyveri*, a strict anaerobe with unique metabolic features. *PNAS* **105(6)**: 2128-2133.

Shendure, J., R.D. Mitra, C. Varma and G.M. Church. 2004. Advanced sequencing technologies: Methods and Goals. *Nature Reviews* **5**: 335-344.

Smith, P.K., Krohn, R.I., Hermanson, G.T., Mallia, A.K., Gartner, F.H., Provenzano, M.D.,Fujimoto, E.K., Goeke, N.M., Olson, B.J. and Klenk, D.C. (1985) Measurement of protein using bicinchoninic acid. *Analytical Biochemistry* **150**: 76-85.

Smith, M.G., T.A. Gianoulis, S. Pukatzki, J.J. Mekalanos, L.N. Ornston, M. Gerstein and M. Synder. 2007. New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposons mutagenesis. *Genes and Development* **21**: 601-614.

Staden, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acid Res.* **8:** 3673-3694.

Staden, R., Beal, K.F., and J.K. Bonfield. 2000. The Staden package, 1998. *Methods Mol. Biol.* **132**: 115-130.

Takami, H., Y. Takaki, G-J. Chee, S. Nishi, S. Shimamura, H. Suzuki, S. Matsui and I. Uchiyama. 2004. Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. *Nucleic Acids Research* **32(21)**: 6292-6303.

Tettelin, H., D. Radune, S. Kasif, H. Khouri and S.L. Salzberg. 1999. Optimized multiplex PCR: efficiently closing whole-genome shotgun sequencing project. *Genomics* **62**: 500-507.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

van Heerden, E., D.J. Opperman, A.P. Bester, J. van Marwijk, E.D. Cason, D. Litthauer, L.A. Piater and T.C. Onstott. 2008. Metabolic promiscuity from the deep subsurface: a story of survival or superiority. *Instruments, Methods and Missions for Astrobiology XI*, Proc. of SPIE, 7097: 7079S-1.

van Hijum, S.A.F.T., A.L. Zomer, O.P. Kuipers and J. Kok. 2005. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Research* **33**: W560-566.

Venter, J.C., S. Levy, T. Stockwell, K. Remington and A. Halpern. 2003. Massive parallelism, randomness and genomic advances. *Nature Genetics Supplement* **33**: 219-227.

Villbrandt, B., H. Sobek, B. Frey and D. Schomburg. 2000. Domain exchange: chimeras of *Thermus aquaticus* DNA polymerase, *Escherichia coli* DNA polymerase I and *Thermotoga neopolitana* DNA polymerase. *Protein Engineering* **13(9)**: 645-654.

Wang, Y., D.E. Prosen, L. Mei, J.C. Sullivan, M. Finney and P.B.V. Horn. 2004. A novel strategy to engineer DNA polymerases for enhanced processivity and improved performance *in vitro*. *Nucleic Acids Research* **32(3)**: 1197-1207.

Wicker, T., E. Schlagenhauf, A. Graner, T.J. Close, B. Keller and N. Stein. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275**.**

Wollherr, A. and H. Leisegang. 2008. Bacterial pan-and core-genomes: a comparative genomics approach based on bi-directional BLAST. Göttingen Genomics Laboratory, Georg-August-Universität Göttingen, unpublished data.

Xu, J. 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools and recent advances. *Molecular Ecology* **15**: 1713-1731.