UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Genome assembly and metabolic pathway reconstruction of *Pantoea ananatis* LMG 20103

by

Wai Yin Chan

Submitted in partial fulfilment of the
requirements for the degree
Magister Scientiae

In
The Faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria

Supervisor: Prof S.N. Venter
Co-supervisor: Prof T.A. Coutinho

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Declaration

I, the undersigned, hereby declare that the thesis submitted herewith for the degree **Magister Scientiae** to the University of Pretoria contains my own independent work. This work has not previously been submitted for any other degree at any other University.

____

Wai Yin Chan

April 2012

Genome assembly and metabolic pathway reconstruction of *Pantoea ananatis* LMG 20103

by

Wai Yin Chan

Supervisor: Prof S.N. Venter

Co-supervisor: Prof T.A. Coutinho

Department: Microbiology and Plant Pathology

Degree: MSc (Microbiology)

SUMMARY

Next generation of sequencing (NGS) technologies have taken life science research into a new era. With the rapid advances in these technologies and the associated reduction in overall costs, the sequencing and assembly of genomes have come within reach of most laboratories. Studies related to the evolution, ecology and biology of an organism now rely heavily on genomic data and obtaining a genome sequence has become an essential resource for the rapid progress and success of these studies.

Pantoea ananatis is recognised as an emerging but rather unconventional pathogen capable of infecting a wide range of different hosts. Numerous plants of agricultural and economic importance including maize, rice, onion, pineapple, melon, sudan grass and *Eucalyptus* trees have been affected. With the outbreak of *P. ananatis* in a South African *Eucalyptus* nursery in 1998, it was realised that very little is known about this pathogen. A better understanding of the pathogenicity, metabolism and ecology of the bacterium is required to develop strategies for the control of the disease.

During this study, the genome sequence of *P. ananatis* strain LMG 20103 was obtained using the Roche 454 technology. To aid in the assembly of this *Eucalyptus* pathogen's genome sequence, the type strain of *P. ananatis* LMG 2665 was also sequenced using Illinima's Genome Analyzer (GA). A draft assembly of *P. ananatis* LMG 20103, consisting of 117 contigs, was generated after optimization of the Newbler assembly parameters and comparison with other genome assemblies and genomes. This study demonstrated that the assembly could be completed using both *in-vitro*, and *in-silico* approaches such as contig scaffolding, gap closure with conventional PCR reactions and sequencing, manual curation and automated genome annotation. The final complete genome consisted of a 4 386 227 bp chromosome and a 317 146 bp mega-plasmid.

With the complete genome sequence available, the reconstruction of metabolic network of *P. ananatis* LMG 20103 was attempted using two pathways reconstruction pipelines namely, Pathway Tools and Model SEED. It was found that missing metabolic reactions and incomplete pathways in the draft metabolic networks were mainly caused by incorrect gene annotations or bioinformatic errors during the automated network reconstruction. These two pipelines differed substantially in the way network reconstruction is undertaken. Performing a comparison between the two proposed networks, annotation errors could be detected and corrected. Although some improvement could be made to the predicted network further experimental data is still required to improve the accuracy of the draft metabolic network.

Despite the amount of effort and cost, it is believed that the complete genome and a draft metabolic network of *P. ananatis* LMG 20103 will be a valuable resource for many subsequent studies to investigate the evolution and biology of this emerging plant pathogen. This information will be essential for the development of strategies to predict and control future disease outbreaks associated with this pathogen.

ACKNOWLEDGEMENTS

- I will like to take this opportunity to express my appreciation to my family and friends (Tony, Andy, Grace and Ricardo) for their financial support and encouragement to further my research career.

Also with special recognition to:

- My supervisors, Prof. Stephanus N. Venter and Prof. Teresa A. Coutinho for their patience and special guidance throughout my MSc.
- Dr. Pieter De Maayer and Dr. Carrie Brady for being my mentors, as I have learnt many practical skills under their guidance.
- The Bioinformatics and Computational Biology Units of University of Pretoria (especially to Prof. Fourie Joubert, Dr. Charles Hefer and Mr Keoagile Oliver Ignatius Bezuidt) for all the computation support.
- All my lab mates and colleges in the Phytobacteriology Research Group and the Department of Microbiology and Plant Pathology (Adéle Prinsloo, Divine Shyntum, Eric Birkholz, Gabrielle Carstensen, Gina Shin, Lunghile Mthombeni, Karabo Sereme, Marike Du Plessis, Olga Makhari, Pieter De Maayer, Sarah MacRae, Tarren Seale, Zander Human) for the joyful work environment.
- Prof. Ian K. Toth and Prof. Paul R.J. Birch, for their hospitality during the research exchange program at the Scottish Crop Research Institute (SCRI) in Scotland.
- All my funding bodies (National Research Foundation (NRF), Centre of Excellence in Tree Health Biotechnology (CTHB) and Tree Protection Co-operative Programme (TPCP)) for their financial support of my studies.
- Tanzelle Oberholster and Douglas Martin from the Tree Protection Co-operative Programme (TPCP) mentorship program, as I am appreciative of their assistance with my research work.
- Forestry and Agricultural Biotechnology Institute (FABI), for the wonderful opportunity of networking with people all over the world.

TABLE OF CONTENTS

Chapter 1

Genome assembly and metabolic pathways reconstruction of bacterial genomes

Chapter 2

Draft Genome assembly of *Pantoea ananatis* strain LMG 20103 and LMG 2665

Chapter 3

Complete genome assembly of *Pantoea ananatis* LMG 20103

Chapter 4

Genomic metabolic pathway reconstruction of *Pantoea ananatis* LMG 20103

Chapter 5

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC transporters	-	ATP-binding cassette transporter
ADP	-	Adenosine diphosphate
AI	-	Autoinducer
Ala	-	Alanine
API	-	Analytical profile index
Asp	-	Asparic acid
BASys	-	Bacterial Annotation System
BLAST	-	Basic Local Alignment Search Tool
bp	-	Base pair
BRENDA	-	Braunschweig Enzyme Database
BsubCyc	-	Encyclopaedia of *Bacillus subtilis* Genes and Metabolisms
CDP	-	Cytosine diphosphate
CFA	-	Cyclopropane fatty acid
CMP-KDO	-	Cytidine 5'-monophosphate-3-deoxy-D-manno-octulosonate
CMR	-	Comprehensive Microbial Resource
CNA	-	CellNetAnalyzer
Cys	-	Cysteine
DADNt2	-	Deoxyadenosine
ddNTPs	-	Dideoxynucleotide triphosphates
DMSO	-	Dimethyl sulfoxide
DNA	-	Deoxyribonucleic acid
dNTP	-	Deoxyribonucleotide triphosphate
dTDP	-	Thymidine diphosphate
E. coli	-	*Escherichia coli*
EBI	-	European Bioinformatics Institute
EC	-	Enzyme commission
EcoCyc	-	Encyclopaedia of *Escherichia coli K-12* Genes and Metabolisms
EtOH	-	Ethanol

FABI	-	Forestry and Agriculture Biotechnology Institute
FBA	-	Flux balance analysis
Fe	-	Iron
fm	-	From
FVA	-	Flux balance analysis
GA	-	Genome Analyzer
GAM	-	Growth-associated ATP maintenance reaction
GDP	-	Guanosine diphosphate
GEM	-	Genome-scale model
Glimmer	-	Gene Locator and Interpolated Markov ModelER
Gln	-	Glutamine
Glu	-	Glutamic acid
Gly	-	Glycine
GPR	-	Gene-protein-reaction
GS	-	Genome sequencer
H	-	Hydrogen
Hg	-	Mercury
His	-	Histidine
HumanCyc	-	Encyclopaedia of *Homo sapiens* Genes and Metabolisms
IMG	-	Integrated Microbial Genome
INSt2r	-	Inosine
kb	-	Kilo base pair
KDO	-	3-Deoxy-D-manno-oct-2-ulosonic acid
KEGG	-	Kyoto Encyclopaedia of Genes and Genomes
Leu	-	Leucine
MaGe	-	Microbial Genome Annotation
Met	-	Methionine
MOPS	-	3-(N-morpholino)propanesulfonic acid
Na	-	Sodium
NAD	-	Nicotinamide adenine dinucleotide
NADH	-	Reduced form of nicotinamide adenine dinucleotide
NaOH	-	Sodium hydroxide
NCBI	-	National Center for Biotechnology Information

NGS	-	Next generation sequencing
NMN	-	Nicotinamide ribonucleotide
OLC	-	Overlap-layout-consensus
ORF	-	Open reading frames
P. ananatis	-	*Pantoea ananatis*
Pb	-	Lead
PCR	-	Polymerase chain reaction
PEP	-	Phosphoenolpyruvate
PGDB	-	Pathway Genome Database
Phe	-	Phenylalanine
ppGpp	-	Guanosine pentaphosphate
Pro	-	Proline
PTS	-	Phosphotransferase system
Pyr	-	Pyruvate
Q40 +	-	Number of base with an quality score above 40
RAST	-	Rapid annotations using subsystems technology
rDNA	-	Ribosomal DNA
RNA	-	Ribonucleic acid
S. aureus	-	*Staphylococcus aureus*
SCRI	-	Scottish Crop Research Institute
SNP	-	Single-nucleotide polymorphism
SRI	-	Stanford Research Institute
TC number	-	Transporter Classification
TCA	-	Krebs cycle
THF	-	Tetrahydrofuran
Thr	-	Threonine
TransportDB	-	Transporter Protein Analysis Database
TRDR	-	Thioredoxin reductase
tRNA	-	Transfer RNA
Tyr	-	Tyrosine
UDP	-	Uridine diphosphate
UWC	-	University of Western Cape

PREFACE

Genome research and sequencing are currently the leading drivers in biological research. With the development and advances in high throughput sequencing platforms (Margulies *et al*., 2005; McKernan *et al*., 2009; Quinn *et al*., 2008) and the reduction in associated costs, sequencing and assembly of genomes has now become available to most laboratories. The genome sequence of an organism not only serves as the blue print of all the genes present but provides the opportunity to expand our knowledge and understanding of the organism's biology. Through data mining, the functioning and metabolism of the organism can be determined and an understanding of the functional role of the organism in a specific community can be developed. This data also provides for new opportunities to answer more complex questions about the ecology of an organism. Studying the metabolic capability of an organism allows us to comprehend the possible interactions between the organism and its surrounding environment. This valuable information has also been beneficial in other research areas such as metabolic pathway engineering (Durot *et al*, 2009; Thiele and Palsson, 2010).

In order to answer questions related to the evolution, population dynamics, metabolic capabilities and regulatory network of an organism a fully assembled complete genome of high quality is often required. An incomplete draft genome sequence might lack details such as the orientation of genes and their order, the location of promoter or regulation sites and the various repetitive sequence copies (Chain *et al*., 2009; Fraser *et al*., 2002). Assembly of a complete genome is, however, a painstaking process and the quality of the assembly will differ depending on the quality of the raw sequence data, the assembler algorithm used as well as the nature of the genome (Nagarajan *et al*., 2010).

Manual curation of the assembly and annotation is another crucial step in obtaining a complete genome sequence. The construction of metabolic pathways from the genome data can provide important assistance during this curation process (Osterman and Overbeek, 2003). With the help of the pathway map for an organism, the crucial enzymes missing from the annotation can be identified. This information is then used to find alternative enzymes, genes that were incorrectly annotated or to locate open reading frames that were not previously predicted by the initial approach used for this purpose.

Pantoea ananatis is a yellow pigmented bacterium, which is recognised as an emerging but rather unconventional pathogen capable of infecting a wide range of different hosts. Numerous plants of agricultural and economic importance including maize, rice, onion, pineapple, melon, sudan grass and *Eucalyptus* trees have been affected (Coutinho *et al*., 2002; Coutinho and Venter, 2009). Little information on what makes *P. ananatis* such a versatile bacterium is? available and no information on the pathogenicity and virulence factors has been published.

It was strongly believed that the complete genome sequence and the reconstructed metabolic network of a *P. ananatis* strain could offer a solid foundation to better understand this bacterium as a plant pathogen and to support the development of disease control strategies for this organism in the future. The aim of this study was, therefore, to produce a fully annotated complete genome sequence of *P. ananatis* that complied with the requirements as was set out by the Bermuda standard (National Human Genome Research Institute, 2001 and 2002; Human Genome Sequencing Consortium International, 2004). For this reason the genome of *P. ananatis* strain LMG 20103 (the eucalyptus pathogenic strain) was sequenced. A number of bioinformatics approaches and molecular methods were used to construct the complete genome and a metabolic network. The complete genome sequence can serve as the reference sequence for a number of other closely related genome projects and studies into the pathogenicity and ecology of this bacterium. The draft genome of the *P. ananatis* type strain LMG 2665 was also completed during this study.

The approach followed during this study was obtained from the literature. Chapter 1 provides a review on the recent developments in and problems experienced with genome assembly projects and pathways reconstruction. Chapter 2 describes the sequencing approaches (Roche 454 and Illumina) and how Newbler and Velvet assembly were investigated and optimized to improve the assembly quality. The completion of the genome assembly was done through combining the scaffolding of draft contigs, resolution of repetitive sequences and gap closure, using contig graphics, comparisons of closely related draft genome assemblies and various PCR methods (Chapter 3). In Chapter 4 a metabolic network was reconstructed by combining the outputs of two pathway reconstruction software programmes. The draft network was based on a comparison of the different software predictions and this information was then used to improve the genome annotation by filling in incomplete pathway reactions and cross referencing to available experimental results (e.g. Biolog and API test).

References

Chain, P.S.G., Grafham, D.V., Fulton, R.S., FitzGerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., Cole, J.R., Ding, Y., Dugan, S., Field, D., Garrity, G.M., Gibbs, R., Graves, T., Han, C.S., Harrison, S.H., Highlander, S., Hugenholtz, P., Khouri, H.M., Kodira, C.D., Kolker, E., Kyrpides, N.C., Lang, D., Lapidus, A., Malfatti, S.A., Markowitz, V., Metha, T., Nelson, K.E., Parkhill, J., Pitluck, S., Qin, X., Read, T.D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R.L., Sutton, G., Thomson, N.R., Tiedje, J.M., Weinstock, G., Wollam, A., Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, and Detter, J.C., 2009. Genome Project Standards in a New Era of Sequencing. Science 326, 236 –237.

Coutinho, T.A., Preisig, O., Mergaert, J., Cnockaert, M.C., Riedel, K.H., Swings, J., and Wingfield, M.J., 2002. Bacterial blight and dieback of Eucalyptus species, hybrids, and clones in South Africa. Plant disease 86, 20–25.

Coutinho, T.A., and Venter, S.N., 2009. *Pantoea ananatis*: an unconventional plant pathogen. Molecular plant pathology 10, 325–335.

Durot, M., Bourguignon, P.Y., and Schachter, V., 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS microbiology reviews 33, 164–190.

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

Human Genome Sequencing Consortium International, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

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

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M., and Blanchard, A.P., 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research 19, 1527–1541.

Nagarajan, N., Cook, C., Di Bonaventura, M., Ge, H., Richards, A., Bishop-Lilly, K., DeSalle, R., Read, T., and Pop, M., 2010. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. BMC Genomics 11, 242.

National Human Genome Research Institute. 2001. Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project. National Institutes of Health. Available at: http://www.genome.gov/10001812 [Accessed June 22, 2011]

National Human Genome Research Institute. 2002. Human Genome Sequence Quality Standards. National Institutes of Health. Available at: http://www.genome.gov/10000923 [Accessed June 22, 2011]

Osterman, A., and Overbeek, R., 2003. Missing genes in metabolic pathways: a comparative genomics approach. Current Opinion in Chemical Biology 7, 238–251.

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

Thiele, I., and Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5, 93–121.

Chapter 1

Genome assembly and metabolic pathways reconstruction of bacterial genomes

Genome assembly and metabolic pathways reconstruction of bacterial genomes

1.1. Introduction

After the invention of chain-termination sequencing and its use to obtain the first bacteriophage Phi X174 genome sequence (Sanger *et al*., 1977; Staden, 1979), efforts to sequence the genomes of more complex organisms, e.g. bacteria and human genomes followed soon thereafter. Once a few genome sequences were available, it was soon realised how important this information was for our understanding of biology at the molecular and cellular level. This has led to a concerted effort to make genome sequencing technology more affordable and within reach of most scientists. With the introduction of next generation of sequencing (NGS) this dream was realised and the genome sequence of an organism has become one of the essential data resources in all biological science studies.

A genome sequence covers all the encoded gene sequences that are involved in the different cellular functions and metabolic activities within an organism and can assist in our understanding of the organism’s growth and survival strategies as well as its biology. The genome sequence can also serve as a database for gene searches, where putative cellular functions can be predicted, and novel discovery can be made through comparison of genome sequences. As all the cell component information is being stored/recorded in its genomic sequence, an increasing number of life science researchers now focus on data mining and computational analyses of the genome sequence prior to any experimental design or investigation.

With the increasing number of genome sequences available, studying the metabolic pathways of an organism on a genomic scale is currently developing rapidly and various tools have become available to reconstruct the metabolic network of an organism using its genome sequence. A reconstructed metabolic network prediction, with all the possible known pathways of the organism, is often used as a foundation for future advancements. It has been used in medicine to design strategies to interrupt critical metabolic reactions as a means of controlling certain pathogens and diseases. Understanding the metabolic network of an organism also has numerous advantages in the field of biotechnology as scientists can manipulate the phenotypic expression of the organisms for the production of compounds and other industrial benefits (Feist *et al.*, 2008; Hara *et al.*, 2011).

In this review, the improvement of genome sequencing and assembly methods between the first and next generation of sequencing (NGS) is reviewed in an attempt to understand the existing problems and challenges of genome research projects. This is followed by the discussion of various types of solutions that aid with complete genome assembly. Furthermore, the development of metabolic pathways reconstruction is also reviewed. All this information will provide the necessary background to determine the best approach to obtain the complete genome sequence of *Pantoea ananatis* strain LMG 20103 (Coutinho *et al.*, 2002) and to reconstruct its metabolic network.

1.2. Genome sequence research

1.2.1. Genome sequencing

DNA sequencing was in part initiated by Frederick Sanger, who developed the chain-termination method (Sanger *et al.*, 1977; Staden, 1979). Initially Sanger sequencing was performed by synthesizing the target DNA with the addition of dideoxynucleotide triphosphates (ddNTPs) in four different reactions to terminate the elongation of the DNA strand. The products of the four reactions were then separated by gel electrophoresis. To reduce the complexity of sequencing and assist in the automation of the method, fluorescently labelled dye terminator sequencing was introduced as an alternative approach where all the different ddNTPs were added in one reaction. Sanger sequencing is, however, still limited to short fragments of DNA, approximately 1000 – 2000 bp. In order to obtain the sequence of a whole genome, approaches such as chromosome walking and shotgun sequencing were developed. DNA fragments with overlapping edges were sequenced and assembled together based of the sequence alignments (Staden, 1979; Anderson, 1981). Using this approach small virus and bacteria genomes were initially sequenced, followed by other larger genomes including the human genome.

In 2005, next generation sequencing was launced by Roche / 454 Life Sciences (Margulies *et al.*, 2005), followed by Illumina Ltd (Quinn *et al.*, 2008) in 2006 and other technologies such as SOLiD (McKernan *et al*., 2009) and Ion Torrent (Rothberg *et al*., 2011). The new sequencing technologies have brought down the time and expense of sequencing. They also produce short sequencing reads in parallel fashion, resulting in higher coverage for assembly. These technologies have decreased sequencing efforts dramatically and ensured that many larger genomes could be targeted. On the other hand more complex problems with regards to sequencing also arose. High or low GC content had an effect on the sequencing and the presence of repetitive elements affected genome assembly. Issues that have to be addressed include the suitability of certain technologies for *de novo* or re-sequencing projects; whether the quality of the sequenced reads will be good enough for downstream assembly and analysis; what the required coverage needs to be to ensure the assembly of the whole genome. With no initial standards, these issues had to be resolved empirically through genome assembly endeavours.

1.2.2. Genome assembly

New assembler programs were developed to deal with genome assembly using massive amounts of short sequence data. The main objectives of these programmes are to perform genome assemblies using either a *de novo* or reference assembly. These two approaches have different advantages and disadvantages which will determine the selection of a specific approach based on the information requirements of each genome sequencing project.

With reference assembly, also known as comparative assembly, a reference genome sequence is used as a backbone for the alignment of the new sequence reads. This method can reduce the time spent on assembly, but a closely related completed genome sequence is required to serve as reference sequence for the genome assembly. This approach is ideal for the genome assembly of closely related species. The major drawback of reference assembly is that genome regions with variations are not assembled, which leads to non-specific assembly of an organisms’ genome and is not ideal for in-depth comparison studies between distantly related organisms (Pop, 2009).

For *de novo* assembly, no reference genome is utilized during the assembly process. The genome assembly is done using pairwise alignment of reads based on nucleotide similarity. These generated genome assemblies are rather organism-specific and will highlight issues such as genome rearrangement, horizontally transferred islands, the orientation of genes within the operon and unique insertions and deletions (Pop, 2009).

The quality of the genome assembly is influenced by the nature of the genome, the type of sequencing reads as well as the selected genome assembler. A variety of algorithms are utilized in genome assemblers in response to the different assembly problems and in order to optimise the assembly of the different reads. Algorithms such as the Greedy (VCAKE and SSAKE,), Overlap-layout-consensus (OLC) (Newbler and Celera Assembler) or Eulerian path algorithms (Velvet) are used with the goal to generate a single consensus sequence (contig). Depending on the performances/capacity of the available server as well as the algorithm implemented by the different assembly programs, the generated contigs' length and error rate varies between assemblies (Farrer *et al*., 2009; Pop, 2009). Scientists soon realised that these assemblers lack the ability to resolve complex sequence regions within the genome, resulting in misassembled contigs and an inability to assemble repeat sequences (Chaisson *et al.*, 2004; Pop, 2009).

1.2.3. Genome assembly errors

While common AT homopolymer sequencing errors with Roche 454 sequencing are well documented, low GC regions and transposon like elements are also often found at the edge of the contigs at contig breakages (Wicker *et al*., 2006). In addition, other DNA secondary structure formations e.g. hairpin and looped structures may also cause sequencing and assembly errors. Repetitive elements are, however, the major obstacle in genome assembly (Schmutz *et al.*, 2004). According to Phillippy *et al*. (2008) two types of assembly errors are caused by repeat sequences. They are referred to as the "collapse and expansion of repeat copies" or "rearrangement and inversion of sequence".

The collapse or expansion of repeat copies appears in almost any genome assembly. These phenomena occur due to the fact that repeat sequences play an important role in phenotypic variation (Vinces *et al*., 2009). During assembly, the collapse of the repeat is characterised by a stacking of the repeat sequences in a particular region

(Figure 1.1). This manifests itself in a sudden increase in the coverage of that specific region. In the case where a unique sequence is present between the repeats, a repeat unit collapse will exclude this unique sequence, therefore, two contigs are generated instead of one contig (Figure 1.2). These types of mis-assemblies are often recognized by the presence of single nucleotide polymorphisms (SNPs) between the repeat copies. The rearrangement and inversion of repeat sequences are usually mistaken as biological rearrangement events within the genome. These may, however, arise when the order of the repeats is shuffled during assembly, resulting in the rearrangement of the unique sequences in between the repeats as shown in Figure 1.3 (Phillippy *et al*., 2008).

1.2.4. Improvement on genome assembly

The current sequencing strategy to overcome this type of mis-assembly problems is to use mate pairs/paired reads, where both ends of a fragment, for which the estimated size is known, are sequenced. During the assembly process, these reads are assembled together according to the mate-paired constraints. However, this approach cannot accommodate long repeats that stretch longer than twice the insert length of the mate-pair. In addition, mis-assembly can still occur even though all the mate-paired constraints have been met (Arner *et al*., 2006).

To improve a draft genome assembly, Farrer (2009) suggested that the integration of contigs generated by several different assembly programs might improve the quality of the assembly. According to his experiment, *de novo* assembled contigs generated when using Edena and Velvet, using both single and paired-end reads, were being compared to completed genome sequence of *Pseudomonas syringae* pv. *syringae* B728a. It was found that by using different assembly programmes, some programmes managed to provide the correct sequences for certain gaps, while other programmes failed to do so. However, regions that encoded for the mobile genetic elements and noncoding RNA sequences were found to be difficult to assemble (Farrer *et al*., 2009).

1.2.5. Improvement of sequencing errors

A typical problem with pyrosequencing is the difficulty in determining the correct sequence in homopolymeric regions. The correction of this sequencing error can be done according to two approaches. These errors can be corrected with the aid of additional reads from other sequencing platforms, such as Illumina or Sanger sequencing reads. As these extra reads typically have no homopolymer sequence error problems, an assembly combining these reads can eliminate the sequence errors. The other approach to correct these errors is through annotation of the genome. After predicting the possible open reading frames (ORF) within the genome, the DNA and amino acid sequences of the predicted ORF can be annotated based on the similarity with coding sequences from protein databases or by direct comparison to anonymous genomic sequence from NCBI (Baxevanis and Ouellette, 2001). Through this approach, incorrect frame shifts within any protein sequence caused by the indels from the homopolymer error can be pinpointed. This approach is only applicable for errors that occur within ORF regions (Chain *et al*., 2009; Kislyuk *et al*., 2010) and can't be used for non-coding sequences.

1.2.6. Complete genome sequence

Obtaining a complete genome sequence remains a painstaking process (Chain *et al*., 2009). The standard requirements for a complete genome sequence were compiled by the International Human Genome Consortium when the first Human Genome Project was started. Based on the documents "Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project" (http://www.genome.gov/10001812) and "Human Genome Sequence Quality Standards" (http://www.genome.gov/10000923), the Bermuda standard for genome finishing was introduced and currently serves as the best practise guidelines for genome finishing. According to these guidelines the ultimate goal is to obtain a complete genome consisting of one contiguous sequence, with quality scores higher than 30 for each base, as well as with the absence of any gaps, N's or X's characters, according to the National Human Genome Research Institute (2001; 2002).

Finishing the genome assembly therefore requires scaffolding of the contigs by determining their orientation, filling all the gaps that are difficult to assemble by the assembly software, improvement of the low quality/high error rate regions of the contig, addressing the poor contig coverage areas, checking the abnormal high or low CG content regions and correcting the misaligned or misassembled sequences. This process has always been described as the expensive, time and labour intensive part of the any assembly project. During this process additional experimentally produced sequence reads could be of great assistance. These reads are typically produced through amplification of targeted areas using either universal or, custom made primers, followed by sequencing of the product (Gordon *et al*., 2001). These reads are typically added to the existing sequence data after which the re-assembly process is repeated in an attempt to cut down the number of contigs.

1.2.7. Software that aid the completion of a genome sequence

While the traditional process of contig examination and gap filling is done manually through validation using individual PCR sequences, computational tools specially designed to accelerate this process could be used. A number of bioinformatics software packages such as Consed and its Autofinish package, Gap5, DNPTrapper, Tablet, Eagleview and Amosvalidate, can assist in this process (Arner *et al*., 2006; Bonfield and Whitwham, 2010; Gordon *et al*., 1998 and 2001; Huang and Marth, 2008; Milne *et al*., 2010; Phillippy *et al*., 2008). These programs form part of either contig editor or genome assembly viewer software. For a better visualisation of the assembled information, most of the assembly viewer software (e.g. Tablet and Eagleview) include a birds-eye view of the ace assembly file, where all the reads associated with the contig are presented along with the contig's depth and length information. Such an overview of the contig provides a means to rapidly identify high coverage regions within the contig. In addition, colour representation of misassembled and abnormal regions provides a more user-friendly interface for misalignment and identification of repeat sequences in high throughput sequence data. Contig editors (e.g. Consed and Gap5) have most of the functions typically associated with assembly viewers, but they also allow the end user to have direct interactions with the contig for editing proposes such as to remove mis-assembled reads from the contig or resolving and separating the different copies of repetitive sequences (Arner *et al*., 2006; Bonfield and Whitwham, 2010; Gordon *et al*., 1998; Huang and Marth, 2008; Milne *et al*., 2010). Other more advanced software also includes various validation tools. The Amosvalidate pipeline, allows for a draft genome assembly to be passed through a series of tests, such as mate-pair validation, repeat sequence and coverage analysis, identification of micro-heterogeneities, read breakpoint analysis and integration of validation signatures (Phillippy *et al*., 2008) in order to validate the present assembly. Although these test results and inspection only service as indicators for mis-assembly, manual curation and experimental confirmation are required to address some of the conflicting reports and errors.

1.2.8. Annotation of genomes

The next phase, once the genome has been completely assembled, is that the information within the genome needs to be interpreted, by means of annotating the genome. This involves the recognition of the gene coding regions and identification of the protein functions of these predicted gene sequences. Open reading frame (ORF) predictions can be done with a number of bioinformatics tools such as Glimmer (Gene Locator and Interpolated Markov ModelER) (http://www.cbcb.umd.edu/software/glimmer) or fgenesh (http://linux1.softberry.com/all.htm). These programmes search for start and stop codons in the different frame shift translations of the nucleotide sequence (Delcher *et al*., 2007; Salamov and Solovyev, 2000). Once the ORFs are determined, the annotation of the gene/gene prediction can be done in a number of ways using either the DNA or protein sequence. With the DNA sequence approach, the prediction is done by searching for nucleotide signals associated with specific genes, finding homology to known coding sequences, or direct comparisons to anonymous genomic sequences. When using the protein sequence for annotation there is a strong focus on the physical properties of the protein and motifs and patterns by which the function of the protein can be predicted (Baxevanis and Ouellette, 2001).

1.3. Metabolic pathway and network studies

A metabolic pathway is a series of biochemical reactions that alters a particular chemical compound within a cell, so it can be utilised directly by the cell, become part of a different pathway for further processing or be stored within the cell. These biochemical reactions are catalysed by enzymes in the presence or absence of co-factors, metal ions or vitamins. Pathways function together to form the metabolic/biochemical network of the specific cell (Covert *et al*., 2001; Papin *et al*., 2003).

Network reconstructions are primarily done for well-studied organisms (*Escherichia coli* K-12 MG1655, *Bacillus subtilis*, laboratory mouse and human) (Oh *et al*., 2007; Romero *et al*., 2004; Thiele and Palsson, 2010) but can also be attempted for newly sequenced organisms. With the advancement of technology, in both sequencing and computational analysis, a substantial amount of data is generated that can be used for the reconstruction of these metabolic networks. With the reconstructed metabolic network, a better understanding of the functioning of the organism can be obtained by relating its phenotypic characteristics to metabolic functional pathways. Issues that can be addressed include the type of nutrient and resource utilisation, an understanding of the mechanisms involved in the interactions of the organism with its biotic and abiotic environment, as well as symbiotic or pathogenic behaviour. Metabolic pathways also provide information that can be utilized in metabolic engineering for the optimisation or suppression of certain pathways as to control and prevent undesirable reactions. Metabolic network comparisons between different strains or species can also lead to the discovery of new pathways or functional proteins. Lastly, the metabolic network of an organism can provide an important foundation for experimental design such as lock-out experiments, where the targeted gene/s are removed or suppressed in a pathway, or the prediction or explanation of experimental outcomes including issues such as their interactions with specific chemicals compounds or extreme environmental conditions (Durot *et al*., 2009; Hanisch *et al*., 2002; Imielinski and Belta, 2008; Karp *et al*., 2011; Risso *et al*., 2008; Tomita and Arakawa, 2006; Zamboni *et al*., 2008).

The reconstruction of any network can be broken down into the four major steps as shown in Figure 1.4 (Feist *et al*., 2008):

- Network reconstruction
- Curation of the network
- Conversion of network into a computational representation, and
- Use of the constructed network with high-throughput data

1.3.1. Construction of a draft network

This is the initial step for the construction of a draft pathway network and it has also been referred to as metabolic pathway reconstruction. During this phase, genome information is obtained from a variety of databases such as EntrezGene, Comprehensive Microbial Resource (CMR), Genome reviews of European

Bioinformatics Institute (EBI) or Integrated Microbial Genome (IMG) (Feist *et al.*, 2008).

From the annotated genome sequence, the presence of enzymes and metabolites serve as an indication of the presence of specific pathways, and provide clues for the type of biochemical reactions and enzyme interactions to be expected. Genes that are involved in the same pathway are often regulated by the same promoter and are clustered together in the same operon, therefore the order and location of the genes provides vital information for biochemical network studies. Through examination of these annotated enzymes and their stoichiometry, a draft network can be constructed manually. Alternatively, automated reconstruction tools (e.g. PathwayTools, metaShark, SEED and GEM System) are also available for the construction of draft networks (Feist *et al.*, 2008).

Pathway Tools is one of the most widely used software packages used for automated metabolic network reconstruction (Karp *et al.*, 2002 and 2005). The reconstructed metabolic pathways are stored in the Pathway/Genome Database (PGDB), which is generated by the PathoLogic software using the "gene-reaction association" approach. Each PGDB consists of the overall cellular network including the transport proteins and reactions as well as the metabolic and signalling reactions within the targeted organism. With Pathway Tools, the simple organism network is displayed as a 2-D network representation (Figure 1.5), where the locations and characterised description of the reactions and metabolic pathways can be simultaneously displayed (Paley and Karp, 2006).

Model SEED is a metabolic pathway network reconstruction software specially designed for genome annotation comparison and analysis. Instead of applying the general steps of network reconstruction, the "subsystems approach" is implemented by using the coherent reaction sub-network database. The subsystem of a pathway network is identified as a scenario, with a set of scenario inputs (well defined substrates) and scenario outputs (reaction products) that are interconnected with biochemical reactions, according to the KEGG database. Once a genome is annotated, the search for similar sub-networks from the database is initiated. Thereafter SEED's Path-finding tool comes in to connect the selected subnetworks/scenarios to complete the different pathways. Overall, the re-construction of the network is done according to the functional roles of the annotated genome (Figure 1.6). This provides major advantages for gap fillings and network verification and therefore reduces some of the intensive manual curation time required (DeJongh *et al.*, 2007).

Direct metabolic network comparisons between organisms can also be done by both Pathway tools and Model SEED, where shared pathways or reactions are highlighted based on the user's request, to aid with curation of the network. Some of the metabolic reconstruction software also includes editing interfaces that allow users to edit their organisms' network. With the recent developments, most of these programs also provide other verification methods for pathway network e.g. flux balance analysis (Kauffman *et al.*, 2003).

1.3.2. Errors in network predictions

The genome sequence holds the key to a comprehensive understanding of the organism and a complete genome sequence is required for any thorough reconstruction of the network. Draft genomes only provide a starting point for reconstruction studies, and may still contain numerous errors due to the effect of a poor genome annotation. Incorrect annotation of genes could be due to poor sequencing data or the use of an inappropriate gene-finding algorithm. The use of an incomplete genome sequence for reconstruction results in missing genes and its corresponding reactions in the network. In some cases the gene could be missing from the network because the annotation does not contain any information on the types of substrate it interacts with and therefore no association could be made between the gene and the reactions it would be involved in (Feist *et al.*, 2008).

Another source of errors in the draft metabolic network could be that the data contained in the databases used during the automatic construction of the network was incomplete. When it comes to metabolic pathways, the general information on issues such as Gene-protein-reaction (GPR) might be lacking. In some cases the specificity and directionality of a reaction are not well defined, as a general group of substrates were assigned to the enzyme and some of the reactions are not reversible. In many cases only the neutral state of the compound is considered for the reactions, which might not be a good representation for its protonated form present at a different pH level. For some pathways the organism studied will have unique biosynthesis pathways compared to the reference organisms contained in the database and these reactions will also be missing from the network (Feist *et al.*, 2008).

1.3.3. Curation of a draft network

Curation is a crucial step in the reconstruction of the metabolic network, especially if the draft network was constructed automatically. Since the draft network only consists of a list of candidate pathways that are selected based on the information extracted from the genome, the incomplete sections or incorrect areas require manual modification. Detailed resources such as related organism-specific databases, journal publications and review articles are useful for the manual curation of the network.

During the curation of the network, information from a number of comprehensive databases (e.g. TransportDB, MetaCyc, KEGG, and BRENDA) are utilized (Caspi *et al*., 2006 and 2010; DeJongh *et al*., 2007; Kanehisa *et al.*, 2006; Ogata *et al.*, 1999; Ren *et al.*, 2007; Schomburg *et al.*, 2004). Each of these databases focuses on specific biological data, which are essential components of the cell functions and thus provide vital information for metabolic reconstruction. These publicly available databases are preferred as they are well-structured and provide specific and well curated information for which the quality and accuracy have been assured. Furthermore information on these databases are continuously updated and allow for easy access to the data as they have their enzyme and transporter data hyperlinked to observed biochemical reactions for the of studied organisms (Feist *et al.*, 2008).

Protein localization studies, biochemical studies, reversibility and substrate specificity of enzymes also act as solid evidence to alter and improve an inconclusive draft network. The goal of this step is to generate a knowledge base structured network by filling in the missing gaps and dead-end pathways. In addition, inappropriate pathways are removed or rather replaced by more suitable pathways. SMILEY algorithm, GapFind/GapFill and PathoLogic are common bioinformatics tools used for gap filling (Feist *et al.*, 2008).

1.3.4. Mathematic representation of a metabolic pathway network

A conversion of the network to a computational model is typically performed, where a mathematical representation of the genome-scale model (GEM) is implemented after the genome-scale reconstructed network has been curated. Most of the computational verification and metabolic pathways analysis is performed by the popular Flux balance analysis (FBA) method. This constraint-based approach is done base on a number of assumptions related to mass and energy balances in its network, flux limitations, and stoichiometric constraints required by the metabolic network to reach a steady state (Kauffman *et al.*, 2003). The mathematic representation model, can serve as a platform for experimental predictions and data comparisons. The models are typically used to determine whether certain parts of the network will be executed under different environmental conditions and constraints. The computational evaluations are done by powerful mathematical software (e.g. CellNetAnalyzer (CNA)) (Klamt *et al*, 2007).

To verify the model and the original metabolic network, actual experimental data has to be generated for the comparison. Minimal growth experiments and growth on selective media are standard procedures for obtaining data on the growth, uptake and secretion rate of the studied organism. The chemical composition of the organism's biomass can also provide clues to the list of actual metabolic pathways present. This data can be obtained from experiments where the weight percentage of macromolecular content (DNA or RNA, lipid and protein) in the organism of interest is determined. The molar fractions of each of the macromolecule's building blocks can be calculated and used for further analysis (Feist *et al.*, 2008).

1.4. Future usage of genome sequences and metabolic pathway networks

With the rapid development in biological research, scientists have turned to genomic research as the new approach to understand the fundamentals of life. Through processing the massive amount of information within the genomic sequence and co-relating this data with experimental findings and observations, biologists have begun to recognise new components and their putative functional roles.

The easily accessible data will allow metabolic pathway reconstruction to grow at a fast pace and more in-depth pathway studies can be done. Through the integration of other high-throughput data generated from additional experiments (i.e. metabolomics, proteomics and transcriptomics data; enzymatic assays; genomic neighbourhood, synthetic lethal interactions and knockout experiments) the expansion of metabolic networks will promote the discovery of new proteins and their functions (Feist *et al.*, 2008).

There are still numerous obstacles in the development of metabolic pathway reconstruction, where most of the reconstruction methods assume the metabolic pathways function independently and that enzymes with multi-functions remain difficult to be represented in the network. The complexity of mapping multiple types of data therefore still requires new approaches and algorithms (Khatri *et al*., 2012).

The dynamic integration of both genomic and metabolic knowledge with experimental data has introduced scientists to new opportunities in system biology research. Such knowledge can bring us one step closer to disease control at the metabolic level, e.g. the undesired effect can be suppressed, while the beneficial pathways can be optimised e.g. to promote the production of certain compounds for use in the biotechnology industry (Hara *et al.*, 2011).

References

Anderson, S., 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Research 9, 3015.

Arner, E., Tammi, M., Tran, A.N., Kindlund, E., and Andersson, B., 2006. DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. BMC bioinformatics 7, 155.

Baxevanis, A.D., and Ouellette, B.F.F., 2001. Bioinformatics: a practical guide to the analysis of genes and proteins. John Wiley and Sons.

Bonfield, J.K., and Whitwham, A., 2010. Gap5—editing the billion fragment sequence assembly. Bioinformatics 26, 1699–1703.

Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F., Kaipa, P., Karthikeyan, A.S., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Paley, S., Popescu, L., Pujar, A., Shearer, A.G., Zhang, P., and Karp, P.D., 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research 38, D473 –D479.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P., and Karp, P.D., 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic acids research 34, D511–D516.

Chain, P.S.G., Grafham, D.V., Fulton, R.S., FitzGerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., Cole, J.R., Ding, Y., Dugan, S., Field, D., Garrity, G.M., Gibbs, R., Graves, T., Han, C.S., Harrison, S.H., Highlander, S., Hugenholtz, P., Khouri, H.M., Kodira, C.D., Kolker, E., Kyrpides, N.C., Lang, D., Lapidus, A., Malfatti, S.A., Markowitz, V., Metha, T., Nelson, K.E., Parkhill, J., Pitluck, S., Qin, X., Read, T.D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R.L., Sutton, G., Thomson, N.R., Tiedje, J.M., Weinstock, G., Wollam, A., Genomic

Standards Consortium Human Microbiome Project Jumpstart Consortium, and Detter, J.C., 2009. Genome Project Standards in a New Era of Sequencing. Science 326, 236 –237.

Chaisson, M., Pevzner, P., and Tang, H., 2004. Fragment assembly with short reads. Bioinformatics 20, 2067–2074.

Coutinho, T.A., Preisig, O., Mergaert, J., Cnockaert, M.C., Riedel, K.H., Swings, J., and Wingfield, M.J., 2002. Bacterial blight and dieback of Eucalyptus species, hybrids, and clones in South Africa. Plant disease 86, 20–25.

Covert, M.W., Schilling, C.H., Famili, I., Edwards, J.S., Goryanin, I.I., Selkov, E., and Palsson, B.O., 2001. Metabolic modeling of microbial strains in silico. Trends in Biochemical Sciences 26, 179–186.

DeJongh, M., Formsma, K., Boillot, P., Gould, J., Rycenga, M., and Best, A., 2007. Toward the automated generation of genome-scale metabolic networks in the SEED. BMC bioinformatics 8, 139.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L., 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673.

Durot, M., Bourguignon, P.Y., and Schachter, V., 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS microbiology reviews 33, 164–190.

Farrer, R.A., Kemen, E., Jones, J.D.G., and Studholme, D.J., 2009. De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads. FEMS microbiology letters 291, 103–111.

Feist, A.M., Herrgåard, M.J., Thiele, I., Reed, J.L., and Palsson, B.Ø., 2008. Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology 7, 129–143.

Gordon, D., Abajian, C., and Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome research 8, 195–202.

Gordon, D., Desmarais, C., and Green, P., 2001. Automated Finishing with Autofinish. Genome Research 11, 614–625.

Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T., 2002. Co-clustering of biological networks and gene expression data. Bioinformatics 18, S145 –S154.

Hara, Y., Kadotani, N., Izui, H., Katashkina, J.I., Kuvaeva, T.M., Andreeva, I.G., Golubeva, L.I., Malko, D.B., Makeev, V.J., Mashko, S.V., and Kozlov, Y.I., 2011. The complete genome sequence of *Pantoea ananatis* AJ13355, an organism with great biotechnological potential. Applied Microbiology and Biotechnology 93, 331–341.

Huang, W., and Marth, G., 2008. EagleView: A genome assembly viewer for next-generation sequencing technologies. Genome Research 18, 1538–1543.

Imielinski, M., and Belta, C., 2008. Exploiting the pathway structure of metabolism to reveal high-order epistasis. BMC systems biology 2, 40.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic acids research 34, D354–D357.

Karp, P.D., Keseler, I.M., Altman, T., Caspi, R., Fulcher, C.A., Subhraveti, P., Kothari, A., Krummenacker, M., Latendresse, M., Lee, T., Paley, S.M., Shearer, A.G., and Trupp, M., 2011. BioCyc: Microbial Genomes and Cellular Networks. Microbe 6, 176-182.

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N., 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic acids research 33, 6083–6089.

Karp, P.D., Paley, S., and Romero, P., 2002. The pathway tools software. Bioinformatics 18, S225–S232.

Kauffman, K.J., Prakash, P., and Edwards, J.S., 2003. Advances in flux balance analysis. Current Opinion in Biotechnology 14, 491–496.

Khatri, P., Sirota, M., and Butte, A.J., 2012. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology 8, e1002375.

Kislyuk, A.O., Katz, L.S., Agrawal, S., Hagen, M.S., Conley, A.B., Jayaraman, P., Nelakuditi, V., Humphrey, J.C., Sammons, S.A., Govil, D., Mair, R.D., Tatti, K.M., Tondella, M.L., Harcourt, B.H., Mayer, L.W., and Jordan, I.K., 2010. A computational genomics pipeline for prokaryotic sequencing projects. Bioinformatics 26, 1819 –1826.

Klamt, S., Saez-Rodriguez, J., and Gilles, E., 2007. Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC systems biology 1, 2.

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

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M., and Blanchard, A.P., 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research 19, 1527–1541.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D., 2010. Tablet—next generation sequence assembly visualization. Bioinformatics 26, 401–402.

National Human Genome Research Institute. 2001. Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project. National Institutes of Health. Available at: http://www.genome.gov/10001812 [Accessed June 22, 2011]

National Human Genome Research Institute. 2002. Human Genome Sequence Quality Standards. National Institutes of Health. Available at: http://www.genome.gov/10000923 [Accessed June 22, 2011]

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 27, 29–34.

Oh, Y.-K., Palsson, B.O., Park, S.M., Schilling, C.H., and Mahadevan, R., 2007. Genome-scale Reconstruction of Metabolic Network in Bacillus subtilis Based on High-throughput Phenotyping and Gene Essentiality Data. Journal of Biological Chemistry 282, 28791–28799.

Paley, S.M., and Karp, P.D., 2006. The pathway tools cellular overview diagram and omics viewer. Nucleic acids research 34, 3771–3778.

Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., and Palsson, B.O., 2003. Metabolic pathways in the post-genome era. Trends in Biochemical Sciences 28, 250–258.

Phillippy, A.M., Schatz, M.C., and Pop, M., 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome biology 9, R55.

Pop, M., 2009. Genome assembly reborn: recent computational challenges. Briefings in bioinformatics 10, 354–366.

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

Ren, Q., Chen, K., and Paulsen, I.T., 2007. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic acids research 35, D274–D279.

Risso, C., Van Dien, S.J., Orloff, A., Lovley, D.R., and Coppi, M.V., 2008. Elucidation of an Alternate Isoleucine Biosynthesis Pathway in Geobacter sulfurreducens. Journal of Bacteriology 190, 2266–2274.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T., and Bustillo, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., and Karp, P.D., 2004. Computational prediction of human metabolic pathways from the complete human genome. Genome biology 6, R2.

Salamov, A.A., and Solovyev, V.V., 2000. Ab initio gene finding in Drosophila genomic DNA. Genome Research 10, 516–522.

Sanger, F., Nicklen, S., and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y.M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., Rettere, J., Rodriguez, A., Rogers, S., Salazar, A., Tsai, M., and Myers, M., 2004. Quality assessment of the human genome sequence. Nature 429, 365–368.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D., 2004. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Research 32, 431D–433.

Staden, R., 1979. A strategy of DNA sequencing employing computer programs. Nucleic acids research 6, 2601.

Thiele, I., and Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5, 93–121.

Tomita, M., and Arakawa, K., 2006. Large-scale Modeling for Systems Biology. BIOforum Europe 10, 54-55.

Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K.J., 2009. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. Science 324, 1213–1216.

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

Zamboni, N., Kümmel, A., and Heinemann, M., 2008. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. BMC bioinformatics 9, 199.

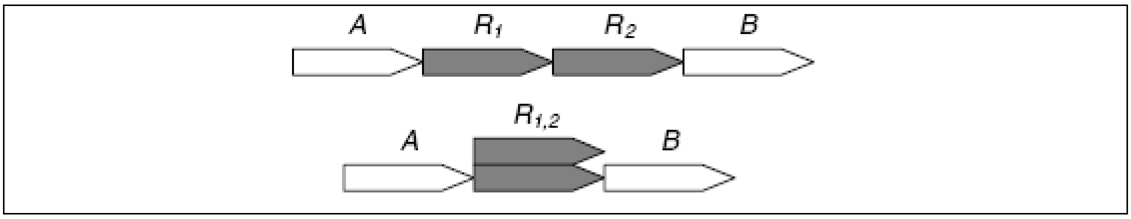


Figure 1.1: Example of a collapsed tandem repeat. The top diagram shows the correct assembly, with R representing the repeat sequences. The bottom diagram shows the incorrect assembly of the repeat into one sequence (Phillippy *et al*., 2008).

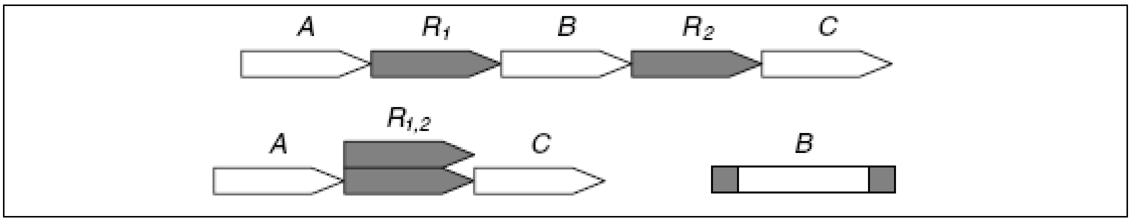


Figure 1.2: Example of a collapsed repeat sequences which contains a unique sequence between the repeat elements. The top diagram represents the correct assembly; R stands for the repeat sequence and B represents a unique sequence. The bottom diagram depicts the classic example, where the repeat sequences are collapsed while the unique sequence B becomes an individual contig (Phillippy *et al*., 2008).

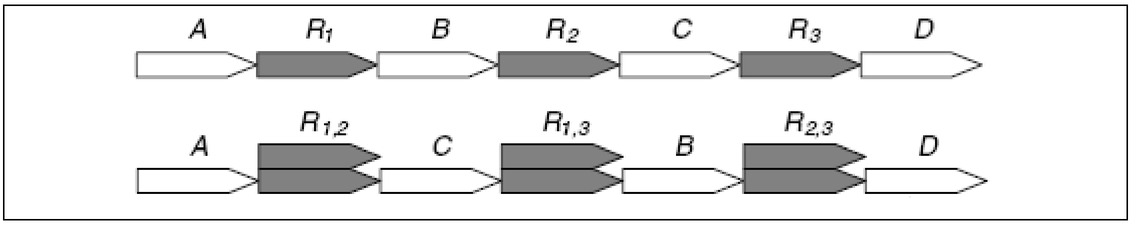


Figure 1.3: Example of rearrangement of repeat sequences. The top assembly represents the correct assembly with 3 repeat sequences separated by two unique sequences B and C. The bottom assembly shows the common mis-assembly that results in mis-placing the unique sequence B and C (Phillippy *et al*., 2008).

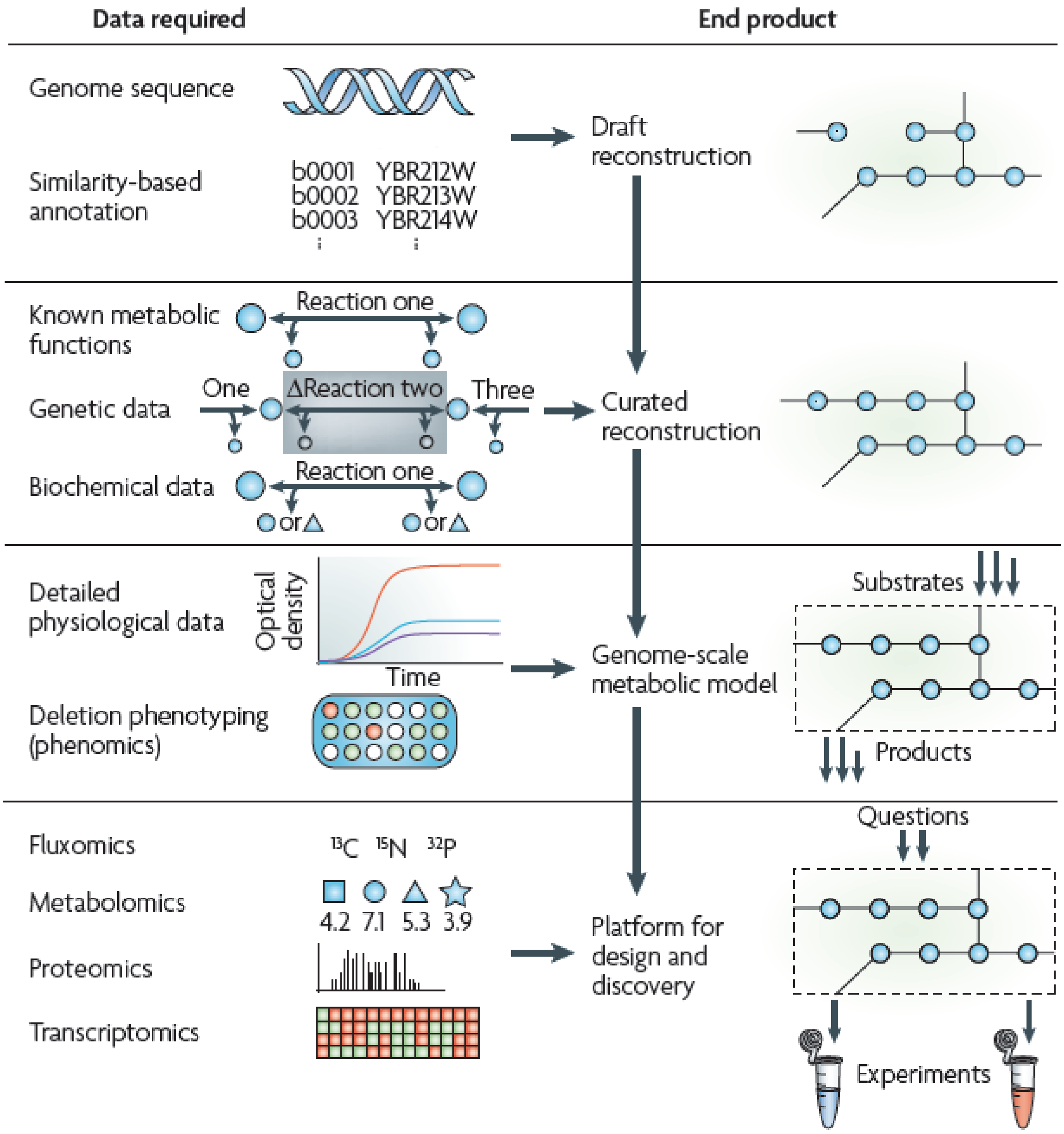


Figure 1.4: A simplified flow diagram depicting the metabolic pathway reconstruction process, indicating the data requirements and end products for each of the four stages of reconstruction (Feist *et al*., 2008).

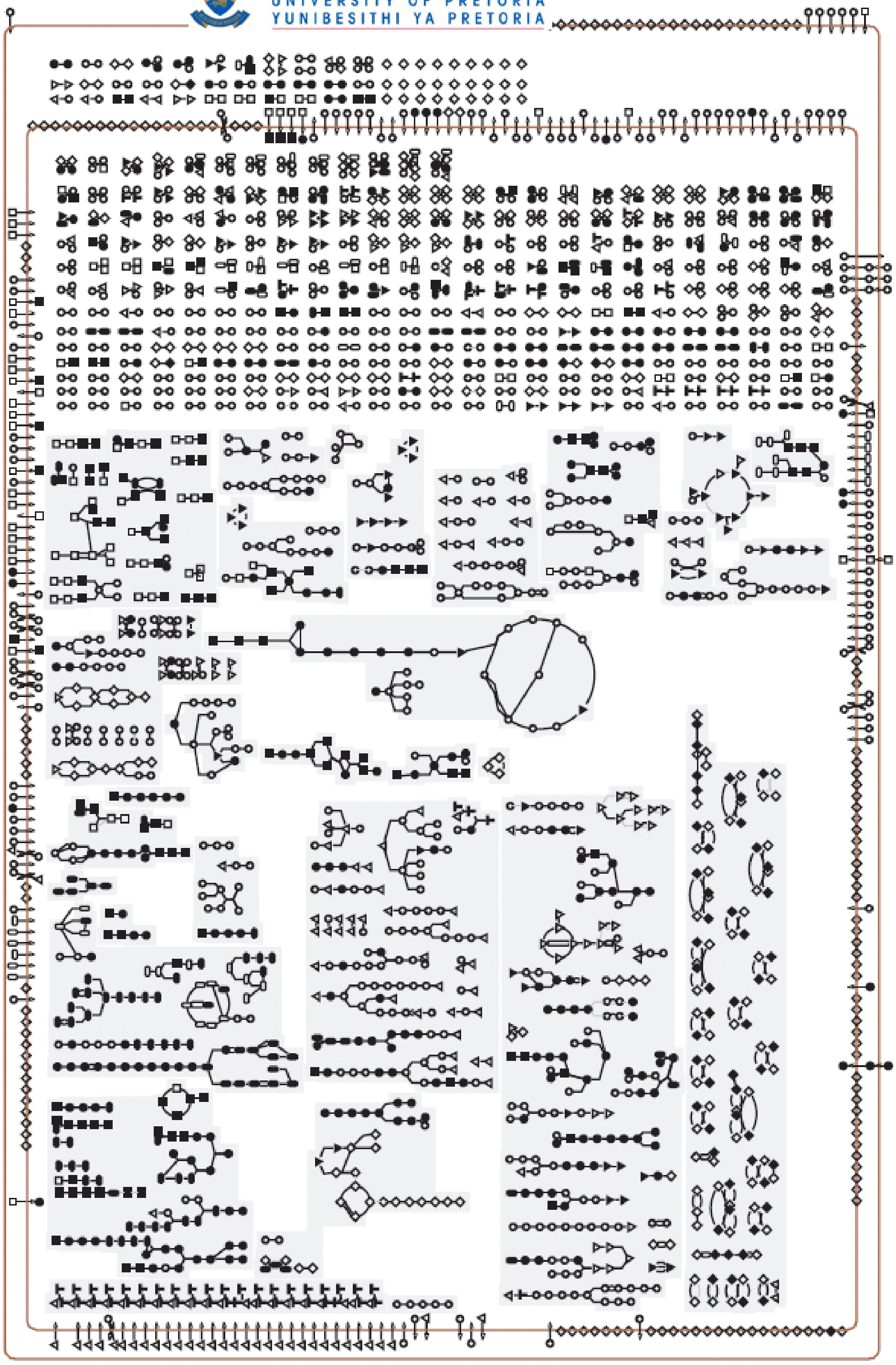


Figure 1.5: The Cellular overview diagram for *E. coli* as generated by Pathway Tools. The border surrounding the diagram represents the cell membrane of the bacteria. Each node represents metabolites or membrane proteins. Edges/lines between nodes are transport reactions or stoichiometric reactions. Pathways are divided into three groups: biosynthetic pathways, energy metabolism pathways, degradation pathways and any other pathways that do not belong to any of the mentioned groups, are arranged from left to right inside the border of the overview diagram. Transport reactions and proteins are included in the border (Paley and Karp, 2006).

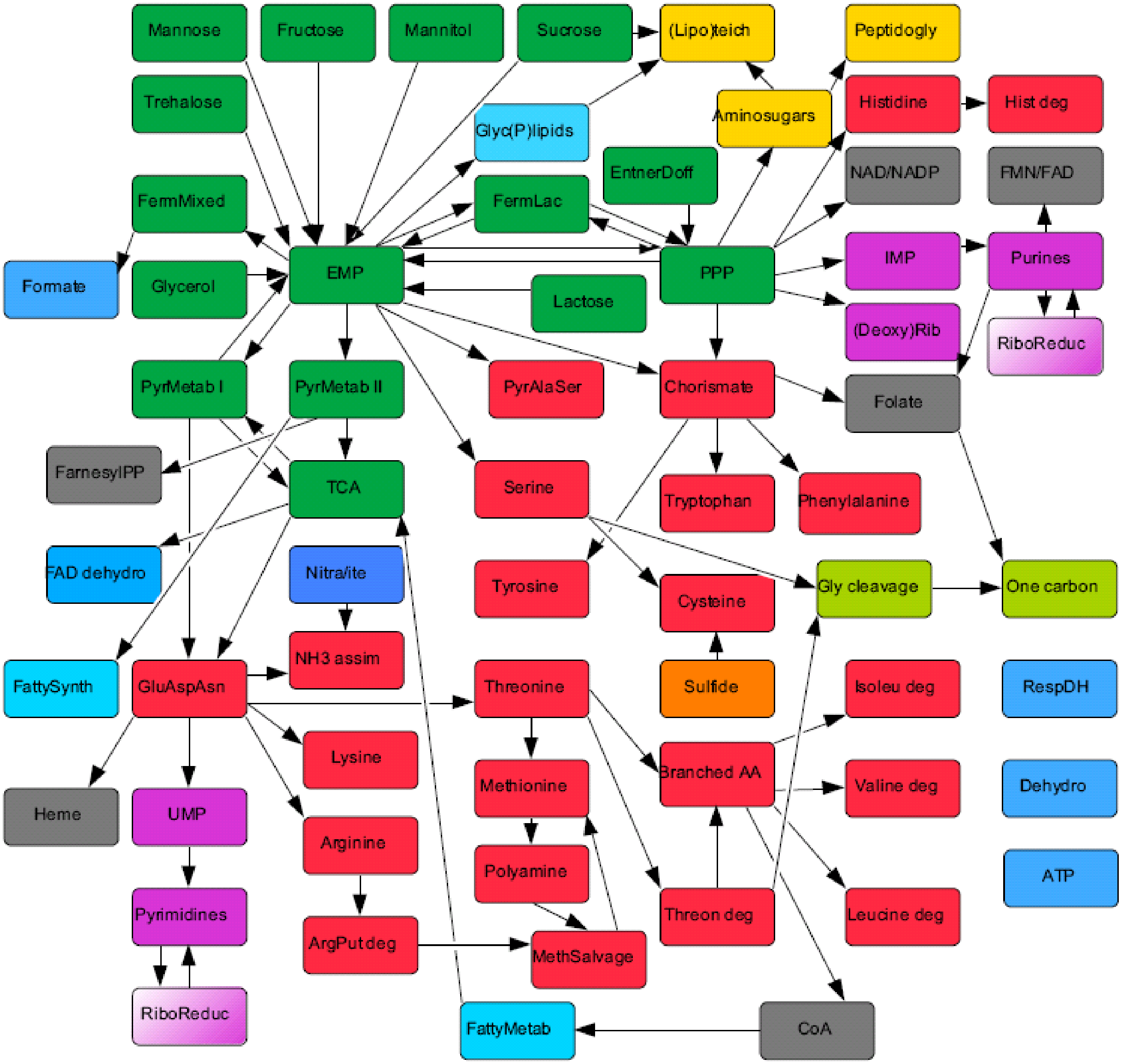


Figure 1.6: Example of *S. aureus* model generated by SEED, with the connection of 65 subsystems represented as individual text boxes (DeJongh *et al.*, 2007).

Chapter 2

Draft Genome assembly of *Pantoea ananatis* strains LMG 20103 and LMG 2665

Draft Genome assembly of *Pantoea ananatis* strains LMG 20103 and LMG 2665

2.1. Introduction

Modern biological research is currently driven by information obtained from sequencing the full genome of a variety of organisms. The genome sequence captures all of the coded gene sequences that form part of the organism and its functions. It is often used to provide possible explanations for biological questions, such as to elucidate the differences between closely related or distantly related organisms though comparison and evolutionary studies. Lastly, the genome data can also be used for gene or protein discovery.

Genome sequencing began in 1979, when shotgun sequencing was proposed by Roger Staden for the sequencing of small genomes. This method involved the random splicing and shearing of the genomic DNA of the targeted organisms, followed by the sequencing of these DNA fragments using the chain termination sequencing method (Sanger *et al*., 1977; Staden, 1979). The overlap between these DNA sequence fragments was then used to assemble the whole genome sequence like a jigsaw puzzle. The launch of next generation sequencing (NGS) in the 2005 resulted in a dramatic increase in genome sequencing as it allowed for the simultaneous sequencing of millions of short reads in parallel. Platforms such as the Genome Sequencer from Roche based on 454's pyrosequencing technology (Margulies *et al*., 2005), Illinima's Genome Analyzer (GA) (Quinn *et al*., 2008), sequencing by the hybridization approach using the SOLiD system (Voelkerding *et al*., 2009) and the latest 'post-light' sequencing by Ion Torrent (Rothberg *et al*., 2011) are currently commonly used in genome sequencing projects.

The pyrosequencing approach, developed by Margulies *et al*. (2005) was initially the main platform used for NGS, due to the huge reduction in the cost of and time spent on genome sequencing. In addition, 454 Roche's longer read length (100 bp to 450 bp) allowed *de novo* assembly of larger genomes and better detection of indel variation in the genome (Whiteford *et al*., 2005; Voelkerding *et al*. 2009). However, mis-estimation of the number of repeats of the same base when occurring in series, also referred to as photopolymers, is recognised as the major error with the pyrosequencing technique (Gharizadeh *et al*., 2006).

Currently the sequencing by synthesis approach as used by Illumina (previously known as Solexa) is also a common method used. Data from this platform is either used on its own or in combination with sequence data obtained from runs with the Genome sequencer (Roche / 454), where homopolymer sequence error from the 454 Roche data are corrected by Illumina's higher quality reads. Illunina's short read length (36 - 72 bp) is still a drawback during the *de novo* assembly of genomes and high coverage depth (50x or more) combined with paired-end reads or a reference sequence is required to improve such assemblies (Voelkerding *et al*., 2009).

Once the sequence data has been collected, the genome sequence assembly can be performed using two approaches, reference assembly or *de novo* assembly. The two approaches provide different advantages and disadvantages. With reference assembly or comparison assembly, the genome assembly is achieved by aligning the available reads to a reference genome sequence which is used as the scaffold. The success of this approach depends mainly on the distance between or how divergent the target genome and the reference genome are from each other. Insertions, deletions, rearrangements and divergent regions within any of the two genomes could have an impact on the accuracy of the new assembly and certain parts of the genome can remain unresolved (Pop *et al*., 2004). On the other hand with *de novo* assembly, no reference genome is utilized during the assembly process. Having full coverage of the genome is crucial to ensure that most of the regions are sequenced and covered by at least one sequencing read, and over-sampling of the genomic DNA is therefore required.

The quality of the assembly is not only dependant on the quality of the sequence data and the assembly approach selected but is also closely linked to the performance and capacity of the available assembler program and computational server. Redundancy, the generated contig length and error rate have been noted to differ between assemblies constructed based on the same initial raw data but using different programmes (Farrer *et al*., 2009). This could be due to the algorithm implemented by the selected assembly programme. The assembly programme also has to deal with other challenges. As the sequence read size decreases, the repetitive nature within the genome becomes harder to resolve and the genomes being sequenced are getting larger in size with a subsequent increase in the amount of sequence reads to be handled (Pop, 2009).

The Newbler assembler was uniquely developed to handle 454 pyrosequence data (Margulies *et al*., 2005; Quinn *et al*., 2008). The Newbler assembler software provides a range of functions to aid with both *de novo* and reference assembly. Low quality reads that include ambiguous values or with uncertain signals are excluded and filtered, leaving only the high quality reads for base calling. A number of algorithms are incorporated in the newer version of this assembler, including the Overlap-Layout-Consensus (OLC) algorithm which uses multiple alignment to construct contigs, the corrective algorithm for the correction of multiple alignments that have conflict regions and the detangling algorithm that resolves tandem repeat regions that are shorter than the read length (Margulies *et al*., 2005; Miller *et al*., 2010; Quinn *et al*., 2008; Zerbino and Birney, 2008).

The Velvet assembler is distributed free of charge and is designed especially for short reads used for *de novo* genome assembly such as the sequencing data obtained from the Illunima/Solexa platform (DiGuistini *et al*., 2009; Zerbino and Birney, 2008). Due to the extreme short length (35 bp to 72 bp) of Illumina reads, the Eulerian path strategy is implemented for the genome assembly by the Velvet assembler. Utilising the coverage cut-off and expected coverage estimation, Velvet will attempt to remove possible errors caused by polymorphisms, SNPs or associated with other biological variants (Pop, 2009; Zerbino and Birney, 2008).

The objective of any genome assembler is to generate an assembly as close as possible to the complete genome sequence. A desirable draft genome assemble will have the least number of contigs and a high N50 statistic value. A number of approaches to optimise a genome assembly have been suggested. Running the assembly with different parameter settings to test the sensitivity and specificity of the assembler should be done. Farrer (2009) found that the integration of different contigs generated by different assemblers do occasionally provide additional information. The use of paired-end reads to assist the assembly has also become common practice with current genome sequencing projects (Ratnakurmar *et al*., 2010; Zerbino and Birney, 2008).

In this study the objective was to get the best genome assemblies for *Pantoea ananatis* strains LMG 20103 and LMG 2665. The genome of *Pantoea ananatis* strain LMG 20103 was sequenced using the genome sequencer (Roche) and strain LMG 2665 was sequenced using the Illumina platfrom. The *de novo* genome assembly of these genomes with Newbler or Velvet was investigated using different parameters. Comparison between closely related draft genomes were also investigated to aid with scaffolding of the draft assembly.

2.2. Materials and Methods

2.2.1. Sequencing and assembly of the *Pantoea ananatis* strain LMG 20103 genome

The *Pantoea ananatis* strain LMG 20103 genomic DNA was extracted using the genomic DNA extraction kit (Fermentas). The genomic DNA was pyrosequenced using the Roche 454 GS20 first generation sequencer at the Inqaba Biotech. Three runs were performed in the March, April and June of 2007 to obtain an estimated 20 times coverage of the genome. The raw 454 sequencing data and quality scores were collected in a .sff file.

The genome assembly of *P. ananatis* LMG 20103 was done using Newbler. A *de novo* assembly of the *P. ananatis* LMG 20103 genome was run on the University of Pretoria Bioinformatics Unit’s servers, Sep and Anjie. Raw 454 reads with the estimated size of 100 bp, were submitted to the Newbler with the following command:

```
runAssembly [options] reads.sff
```

Additional options “–g” and “–consed” were included during the assembly with Newbler 2.0.00. Further bioinformatics approaches to improve the *de novo* genome assemble were attempted, by investigating the use of different parameter settings for the assemblies. Three different values for the parameters: Seed step, Seed length, Minimal overlap length and Minimal overlap identity were selected (Table 2.1). The combination of the different parameters resulted in 81 *de novo* assemblies. The assemblies were run at the Scottish Crop Research Institute (SCRI) cluster server (Gruffalo) with Newbler 2.3.00.

2.2.2. Sequencing and assembly of the *Pantoea ananatis* strain LMG 2665 genome

The *P. ananatis* strain LMG 2665 genome DNA was isolated using the genomic DNA extraction kit (Fermentas). The genomic DNA was sequenced using the Illumina Genome Analyzer, at the University of Western Cape (UWC) in 2008. Since the fastq file format has changed between the release of the different versions of the Illunima technology, the original fastq files could not be used by the latest version of Velvet (Cock *et al.*, 2010). The original single-end read raw data was, therefore, extracted and parsed into .fasta file format. Meanwhile the paired-end read raw data was further parsed into the specific fasta format according to the Velvet manual.

A *de novo* genome assembly of *P. ananatis* LMG 2665 was performed using the Velvet 1.0.01 assembler. The short 36 bp, single-end and paired-end reads with the insert of 200 bp were assembled on the University of Pretoria’s Bioinformatics Server: Zoidberg. Genome assembly with Velvet was divided into two steps, first both single-end and paired-end reads were submitted as fasta files and the *k*-mer spectrum was prepared by the velveth script. The processed data was then used to construct the de Bruijn graph using the velvetg script. The expected coverage and coverage cut-off values were set to the automated estimation for all the assemblies as shown in the example:

```
./velveth /localdata 23 –fasta –shortPaired “paired_end read file name” –short “single_end read file name”
```

```
./velvetg /localdata -exp_cov auto –cov_cutoff auto –ins_length 200
```

Since the Velvet assembler is sensitive to the hash length, further investigation into the optimisation of the *de novo* assembly of *P. ananatis* LMG 2665 was done by determining the suitable hash length for the assembly. Five Velvet assemblies were run with different hash lengths (19-mers, 21-mers, 23-mers, 25-mers and 27-mers). The 5 genome assemblies were thereafter assessed based on their individual assembly statistics.

2.2.3. Nucleotide sequence comparison between the draft Genome assemblies

Three draft genome assemblies of *P. ananatis* LMG 20103 were constructed using different versions of the Newbler software (version 1.0.53, 2.0.00 and 2.3.00). The default parameter settings of Newbler were used in all cases. Contigs that were larger than 500 bp were considered as large contigs and were used for nucleotide sequence comparisons. Comparisons between the three assemblies were done with Mauve (Darling *et al.*, 2004), using the default parameter setting. The aim was to identify possible contigs that could form a bridge between contigs. For the same purpose the draft assembly was also compared with the genomes of closely related *Pantoea* species. The draft genome assemblies of *Pantoea stewartii* subsp. *stewartii DC283* and *Pantoea sp. At-9b* were obtained from the NCBI ftp server and stored as fasta files. For the genome comparison, Mauve and the local alignment algorithm (BLAST) (Altschul *et al*., 1990) were used with default settings. The aim was again to identify possible gap sequences that could form a bridge between contigs in the draft assemblies.

2.3. Results

2.3.1. Genome sequencing and assembly of *P. ananatis* LMG 20103

A total of 991 246 single-ended reads, with an average size of 100 bp, were generated using the Roche pyrosequencing platform. The reads and their quality scores were stored in a .sff file format. D*e novo* assembly was done with three versions of Newbler (1.0.53, 2.0.00 and 2.3.00) using the default parameter settings. For each assembly a separate set of files were generated (454AllContigs.fna, 454AllContigs.qual, 454Contigs.ace, 454LargeContigs.fna, 454LargeContigs.qual, 454NewblerMetrics.txt and 454ReadStatus). These files contained information such as the sequences of the assembled contigs, the quality score for the individual bases within each contig, the actual assembly and other information such as the manner in which the reads were assembled. Additional files were generated, the “Consed folder” was required for the genome visualisation program Consed and the “Contig graph" provided information of possible contig lineages or orientations. The statistical analyses for each assembly were summarized in Table 2.2. An overall improvement in the genome assembly was observed when the new version of Newbler was used. This was demonstrated by a clear increase in the number of large contigs, the average size of the contigs and the N50 value.

The final draft genome assembly (done with Newbler 2.0.00) consisted of a total of 117 contigs, 75 of these contigs were larger than 500 bp in size. A total of 965 020 reads were assembled, and 18 520 reads were excluded. The draft assembly consisted of 4 658 782 bases. Further optimisation of the genome assembly has become a practice in genome research as it was shown that the assembly generated by the default parameter settings might not be the best assembly (Farrer *et al.*, 2009). Using the latest release Newbler version (2.3.00), eighty-one draft genome assemblies were generated and their statistical values were recorded (Table 2.3). These assemblies were evaluated on the basis of various statistics such as N50 and Q40+ values, the number of large contigs, the number of bases assembled, the number of reads assembled, the mean value of contigs and the size of the largest contig, were also recorded. The mean of N50 and Q40+ values of the 81 assemblies were plotted against the different assembly parameter settings are shown in Figures 2.1.

From Figure 2.1 it can be seen that as the seed step increased, the N50 and Q40+ value decreased significantly. However, the opposite behaviour was noted for N50 and Q40+ values when there is an increase in the Seed length. According to the minimum overlap length graph (Figure 2.1), the N50 and Q40+ values reached their optimums with the minimum overlap length value set at 40 %. For the minimum overlap identity the N50 and Q40+ values showed opposite trends. The Q40+ values increased with an increase in the the min overlap identity, while the N50 value drops slightly when the value was set at 90% identity.

2.3.2. Genome sequencing and assembly of *P. ananatis* LMG 2665

The genome of the type strain of *P. ananatis* was also sequenced. This strain (LMG 2665) was isolated as the infectious agent causing disease in pineapple. Instead of the Genome sequencer (Roche) the sequence was determined using the Illumina platform. In total, 3 182 288 single-end reads and 6 4219 504 paired-end reads, with a 200 bp insert, were obtained using the Illumina Genome Analyser. These reads were all 36 bp in size. Using Velvet, 5 genome assemblies were performed using different hash lengths. The statistics of the different assemblies were summarized in Table 2.3.

Evaluation of the different Velvet genome assemblies showed that there is no indication of which assembly is the most optimal, as all the statistics varied greatly between different assemblies. No steady trend was observed, as an increase in the hash length resulted in an increase in the number of nodes but a decrease in the total assembly size. There was also a drop in the total number of reads assembled linked to an increase in the hash length. The N50 and Maximum node size values showed an initial increasing trend as the hash length value increased, followed by a decrease in these values when the hash length was larger than 23 bp. The assembly with the highest N50 statistical value was the third draft assembly, with a hash length of 23 bp and was further analysed. A total of 782 nodes were generated with an approximate 61 times coverage, using 8 327 416 of the total of 9 601 792 short reads. Within these 782 nodes, 297 of them were larger than 45 bp.

2.3.3. Genome assembly comparisons

One of the aims of this study was to produce a complete genome for *P. ananatis* LMG 20103. For this reason a nucleotide sequence comparison between a number of different draft genome assemblies and other genomes was done with Mauve 2.3.1 software (Darling *et al*, 2004). Three draft genome assemblies which included the *P. ananatis* LMG 20103 genome as obtained with Newbler version 2.0.00, 2.3.00, and 1.0.53 respectively, were included. The draft assemblies of *P. ananatis* LMG 2665, *Pantoea stewartii subsp. stewartii DC283* and *Pantoea* sp. At-9b were also included in this comparison (Figure 2.2).

During the comparison, a number of contigs from the other draft assemblies provided suggestions for the improvement of the final draft genome assembly (done with Newbler 2.0.00). Suggestions for bridging between contig edges as well as information on the contig orientation and possible gap sequences were obtained. The comparison between the three closely related *Pantoea* spp., provided only 29, 25 and

42 orientation suggestions from *Pantoea stewartii* subsp. *stewartii DC*283, *Pantoea* sp. At-9b and *Pantoea ananatis* LMG 2665, respectively (Table 2.4). This extra information was recorded for further use when dealing with the manual completion of the genome assembly, described in Chapter 3.

2.4. Discussion

As the ecology and pathogenicity of *Pantoea ananatis* is not well understood, the genome of *P. ananatis* strain LMG 20103, known to infect *Eucalyptus,* was sequenced and assembled to provide a complete genome that could be used as a valuable resource to investigate the biology of this bacterium. The genome of the *P. ananatis* type strain LMG 2665 was also sequenced and the data was used to complement the assembly of LMG 20103. Two different next generation sequencing (NGS) platforms were used to sequence these two isolates. For LMG 20103 the sequencing was performed on the Genome sequencer GS 20 (Roche) and the single-ended read was assembled using different versions of Newbler. The LMG 2665 genome was sequenced using the Illumina Platform. The data consisted of both single and paired-end reads. The two approaches differed substantially and each data set and assembly procedure had their own challenges. As expected, the two draft assemblies were not complete and because the sequencing and assembly were performed using different sequences and assembler programmes the assemblies differed between the two genomes.

Newbler uses what is commonly referred to as an Overlap-Layout-Consensus (OLC) agirothom (Miller *et al*., 2010; Pop, 2009), where reads with the same sequence are aligned and grouped together as one alignment, regardless of the coverage of the alignment. Repeat regions that are larger than the average read length are, therefore, aligned together with a local high coverage. This is a typical problem, especially with a single-end read assembly as was the case for the LMG 20103 assembly. The copies of repeat sequences were stacked together and this resulted in the creation of missing sequences elsewhere within the genome assembly and gaps between contigs. Repetitive sequences problems are, however, not only limited to Newbler assemblies.

Due to the additional benefit of paired-end reads most small repetitive sequences were resolved by Velvet. However, repetitive sequences that stretch longer than the length of the paired-end fragment were still mis-assembled or broken into individual nodes. Another problem was that Velvet assembly's algorithm created additional nodes from poor quality sequencing data as the assembler possibly recognises this data as SNPs or sequence variations. This may be one of the reasons why a high number of small nodes were found within the genome assembly of LMG 2665.

Optimisation of the genome assembler remains one of the critical issues to be addressed in order to create an optimal assembly. Some of the assemblers do not optimise the assembly conditions for the end user. For example, optimisation of the parameters in Velvet is a necessary step for genome assembly. Zerbino and Birney (2008) demonstrated that the value of the hash length does affect the sensitivity and specificity of the Velvet assembler. Commonly the assembly with the highest N50/median value is considered to be the ideal assembly. For the Velvet assembly (Table 2.5), the ideal combination of hash length and other parameters had to be determined empirically.

With the Newbler assembly, the proposed parameters for an optimal assembly are provided as the default setting for all Newbler assemblies. In order to determine whether these paramaters provided the optimal assembly for the *P. ananatis* LMG 20103 genome, four parameters were tested. It was found that only the seed step value differed from the suggested settings for Newbler's 2.0.00 version This slight difference could be due to the fact that the raw data reads originated from an earlier version of the Genome sequencer which provided shorter reads. The data set also did not include any paired-end reads.

The different versions of the Newbler assembler produced different assemblies of varying quality. As can be seen from Table 2.2 there was a huge improvement in the quality of the assemblies between version 1.0.53 and 2.0.00. This is most likely due to the fact that with subsequent upgrades of the Genome sequencer platform certain components of the assembly process used by the first version of Newbler was upgraded to make use of a more suitable algorithm (Quinn *et al.,* 2008). The "Overlapper" approach was modified to increase Newbler's capacity. A reduction in the number of larger contigs and fewer mis-assembly errors were also observed with other released versions of Newbler. Other small improvements included the generation of additional files for other assembly process programs such as consed. The extension of tandem repeat sequences also helped in improving the draft assembly. Due to the continuous and fast developments in the assembly algorithm, not all improvements were captured in the literature (Kislyuk *et al*., 2010; Miller *et al*., 2010) and it will in most cases not be possible to make a specific link between an improvement in the assembly and a change in the algorithm.

As the reads generated by the Genome sequencer are relatively short, high coverage is demanded for an accurate assembly. The use of the OLC algorithm during assembly has unfortunately also became a major draw-back. Due to the nature of this algorithm, assembly of repeat units has become difficult. Stacking of the repeats or termination of the contigs at the start of the repeat units was typically observed (Margulies *et al*., 2005; Quinn *et al*., 2008). Therefore, alternative solutions to resolve the assembly of repetitive sequences still remain a challenge. A major improvement was the inclusion of longer reads and paired-end data (Pop and Salzberg, 2008; Whiteford *et al*., 2005), which improved the draft genome assembly by reducing the amount of contigs. Apart from using these approaches, the addition of long reads provided from other sources such as the PCR and sequencing of individual fragments has also been widely used. For the current project this approach was followed and will be discussed in Chapter 3.

The comparison of the different assemblies of LMG 20103 provided information that could be used to close the gaps between contigs or change the order in which the contigs have been arranged. The data showed that the assemblies did not have major difference in their nucleotide sequences, but that the appearances of gaps varied. Collapsed repeats sequences and, in particular, the ribosomal DNA, could be linked to any one of the correct locations on the genome by Newbler. This data provided useful information for gap closure, contig orientation, the order of the contigs and possible branch points of polymorphisms and was used in the next part of the study to complete the genome assembly of *P. ananatis* LMG 20103.

Gaps and scaffolding suggestions were also provided by comparing the draft genome with those of other closely related *Pantoea* species. More suggestions for the orientation of contigs were obtained from the comparison with another strain of the same species (*Pantoea ananatis* LMG 2665) than with the comparison to other species (*Pantoea stewartii subsp. stewartii DC283* and *Pantoea* sp. At-9b). This was not unexpected as the occurrence of genome rearrangements within the species has been well documented and is less than what is observed at the inter-species level (Darling *et al*, 2008). The comparison with other species could, however, still be useful. Genome rearrangement events often involve clusters of genes (Darling *et al*., 2008), and the synteny of the genes within a cluster could improve the assembly of that cluster in the draft assembly of the targeted genome.

2.5. Conclusions

With the current next generation sequencing technologies and improvement in genome assembler programmes, millions of reads can be rapidly generated and assembled to provide a draft genome of an organism. These draft genome sequences are, however, never complete and gaps and un-assembled fragments are typically still present. This study has shown that optimisation of the parameters used by the assembler programmes and comparisons between different assemblies or with related genomes could provide crucial information that could be used to resolve and improve the final assembly.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38, 1767 –1771.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T., 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Research 14, 1394–1403.

Darling, A.E., Miklós, I., and Ragan, M.A., 2008. Dynamics of genome rearrangement in bacterial populations. PLoS Genetics 4, e1000128.

DiGuistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., Mardis, E., Marran, M.A., Hamelin, R.C., Bohlmann, J., Breuil, C., and Jones, S.J.M., 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10, R94.

Farrer, R.A., Kemen, E., Jones, J.D.G., and Studholme, D.J., 2009. De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads. FEMS microbiology letters 291, 103–111.

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

Kislyuk, A.O., Katz, L.S., Agrawal, S., Hagen, M.S., Conley, A.B., Jayaraman, P., Nelakuditi, V., Humphrey, J.C., Sammons, S.A., Govil, D., Mair, R.D., Tatti, K.M., Tondella, M.L., Harcourt, B.H., Mayer, L.W., and Jordan, I.K., 2010. A computational genomics pipeline for prokaryotic sequencing projects. Bioinformatics 26, 1819 –1826.

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

Miller, J.R., Koren, S., and Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327.

Pop, M., 2009. Genome assembly reborn: recent computational challenges. Briefings in bioinformatics 10, 354–366.

Pop, M., Phillippy, A., Delcher, A.L., and Salzberg, S.L., 2004. Comparative genome assembly. Briefings in bioinformatics 5, 237–248.

Pop, M., and Salzberg, S.L., 2008. Bioinformatics challenges of new sequencing technology. Trends in Genetics 24, 142–149.

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

Ratnakurmar, A., McWilliam, S., Barris, W., and Dalymple, B., 2010. Using paired-end sequences to optimise parameters for alignment of sequence reads against related genomes. BMC Genomics 11, 458.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T., and Bustillo, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352.

Sanger, F., Nicklen, S., and Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74, 5463.

Staden, R., 1979. A strategy of DNA sequencing employing computer programs. Nucleic acids research 6, 2601.

Voelkerding, K.V., Dames, S.A., and Durtschi, J.D., 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry 55, 641–658.

Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., Essex, J.W., Roach, P.L., Bradley, M., and Neylon, C., 2005. An analysis of the feasibility of short read sequencing. Nucleic Acids Research 33, e171.

Zerbino, D.R., and Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821–829.

Table 2.1: Different parameters used for the Newbler assembly runs

Type of parameter setting	**Values used**
Seed step	6, 12, 18
Seed length	10, 13, 16
Minimal overlap length	20, 40, 60
Minimal overlap identity	70, 80, 90

Table 2.2: A Summary of assembly statistics of *P. ananatis* LMG 20103 associated with the different versions of Newbler

	Newbler 1.0.53	**Newbler 2.0.00**	**Newbler 2.3.00**
Number of large contigs	89	75	66
Average size of contigs	52 300	62 014	70 464
N50 value	98 193	131 487	131 961
Q40 + value	4 651 466	4 642 884	4 639 592

Table 2.3: Summary of 81 draft genome assemblies with the different combination of parameter value

			Min	**Min**				**Mean**	**N50**	**Largest**	
Job	**Seed**	**Seed**	**overlap**	**overlap**	**Reads**	**Large**	**Bases**	**contig**	**contig**	**contig**	
#	**step**	**length**	**length**	**identity**	**assembled**	**contigs**	**assembled**	**size**	**size**	**size**	**Q40+**
1	6	10	20	70	964754	78	4650217	59618	130195	248356	4638363
2	12	10	20	70	956959	78	4650428	59620	131957	248439	4637979
3	18	10	20	70	944758	90	4649411	51660	105809	269920	4637598
4	6	10	40	70	967552	63	4650098	73811	132835	282603	4638240
5	12	10	40	70	965443	75	4650002	62000	131485	282600	4638132
6	18	10	40	70	957314	86	4650607	54076	105830	305209	4638842
7	6	10	60	70	965926	67	4650029	69403	160216	282552	4638142
8	12	10	60	70	965574	75	4650323	62004	131959	282600	4638802
9	18	10	60	70	962254	91	4649855	51097	105552	237252	4638059
10	6	10	20	80	964696	67	4650521	69410	140963	355139	4638842
11	12	10	20	80	957017	78	4650920	59627	131485	248356	4638480
12	18	10	20	80	944977	83	4650545	56030	108044	313484	4638047
13	6	10	40	80	967532	63	4650020	73809	132835	282603	4638167
14	12	10	40	80	965445	75	4650003	62000	131485	282601	4638140
15	18	10	40	80	957314	86	4650607	54076	105830	305209	4638842
16	6	10	60	80	965927	67	4650029	69403	160216	282552	4638142
17	12	10	60	80	965574	75	4650323	62004	131959	282600	4638802
18	18	10	60	80	962308	91	4649912	51097	105552	237252	4638108
19	6	10	20	90	964861	72	4651661	64606	128677	248347	4639345
20	12	10	20	90	956547	73	4649773	63695	133970	248353	4638316
21	18	10	20	90	943925	89	4649556	52242	104372	282543	4638422
22	6	10	40	90	967322	66	4651540	70477	137021	282603	4640211
23	12	10	40	90	964769	73	4649435	63690	131960	282600	4637727
24	18	10	40	90	957686	79	4650969	58873	121163	305206	4638546
25	6	10	60	90	965864	65	4649863	71536	160192	282604	4638391
26	12	10	60	90	965327	76	4650623	61192	132108	282602	4639340
27	18	10	60	90	961881	92	4650216	50545	101913	237191	4638428
28	6	13	20	70	963389	69	4650754	67402	131958	313586	4638686
29	12	13	20	70	961324	81	4650489	57413	119538	231236	4638706
30	18	13	20	70	956658	79	4650126	58862	124428	258015	4638109
31	6	13	40	70	967610	66	4650985	70469	131485	282602	4639021
32	12	13	40	70	966927	63	4651581	73834	133184	282601	4639713
33	18	13	40	70	965048	67	4650559	69411	131815	313485	4637975

34	6	13	60	70	965475	69	4650476	67398	131485	282557	4638438
35	12	13	60	70	965346	70	4650605	66437	119532	280633	4638703
36	18	13	60	70	964783	73	4651307	63716	130330	269821	4639383
37	6	13	20	80	963509	77	4651067	60403	126719	342794	4638143
38	12	13	20	80	961878	80	4650691	58133	131485	253115	4638946
39	18	13	20	80	956379	74	4649713	62833	126513	258011	4637447
40	6	13	40	80	967621	64	4651380	72677	131960	282601	4640117
41	12	13	40	80	966840	67	4650534	69410	131814	282603	4638656
42	18	13	40	80	965062	68	4650220	68385	131962	282667	4637617
43	6	13	60	80	965479	69	4650476	67398	131485	282557	4638437
44	12	13	60	80	965345	70	4650608	66437	119536	280633	4638667
45	18	13	60	80	964779	73	4651307	63716	130330	269821	4639360
46	6	13	20	90	962787	88	4650611	52847	114808	269920	4638809
47	12	13	20	90	961360	95	4649328	48940	90288	230219	4637665
48	18	13	20	90	957189	75	4649735	61996	126224	248352	4637692
49	6	13	40	90	967025	65	4651269	71557	131815	282599	4640084
50	12	13	40	90	966556	69	4650143	67393	131485	282598	4638936
51	18	13	40	90	964832	67	4650711	69413	131960	282599	4639219
52	6	13	60	90	965496	68	4651507	68404	137024	282598	4639465
53	12	13	60	90	964976	68	4650192	68385	131483	282607	4638616
54	18	13	60	90	964607	77	4650768	60399	110153	237207	4638794
55	6	16	20	70	963522	60	4650472	77507	169766	355133	4638898
56	12	16	20	70	960085	75	4650760	62010	131721	355089	4638481
57	18	16	20	70	4638481	70	4649877	66426	131959	282556	4638800
58	6	16	40	70	966798	69	4651340	67410	161652	282602	4638830
59	12	16	40	70	965402	63	4650487	73817	131957	313483	4638627
60	18	16	40	70	963064	75	4650413	62005	131484	282600	4637970
61	6	16	60	70	965493	69	4649324	67381	132757	282559	4637456
62	12	16	60	70	965102	67	4650521	69410	132776	282603	4638834
63	18	16	60	70	964242	74	4651703	62860	131485	269919	4639797
64	6	16	20	80	963522	60	4650472	77507	169766	355133	4638898
65	12	16	20	80	960088	75	4650760	62010	131721	355089	4638482
66	18	16	20	80	954340	72	4650944	64596	131483	282555	4639223
67	6	16	40	80	966798	69	4651339	67410	161651	282602	4638833
68	12	16	40	80	965552	62	4650495	75007	131958	313483	4638640
69	18	16	40	80	963062	75	4650412	62005	131484	282600	4637968
70	6	16	60	80	965493	69	4649324	67381	132757	282559	4637456
71	12	16	60	80	965102	67	4650521	69410	132776	282603	4638834
72	18	16	60	80	964242	74	4651703	62860	131485	269919	4639797
73	6	16	20	90	963010	68	4650637	68391	144352	355152	4638915

74	12	16	20	90	959923	66	4649952	70453	162268	355201	4638144
75	18	16	20	90	954418	76	4650542	61191	131483	276230	4638002
76	6	16	40	90	966458	66	4650679	70464	131961	282603	4639592
77	12	16	40	90	965397	69	4651365	67411	131486	282603	4640161
78	18	16	40	90	963199	73	4650633	63707	119561	282597	4638137
79	6	16	60	90	965375	66	4650613	70463	162265	282604	4638760
80	12	16	60	90	964659	72	4649579	64577	131485	282603	4637518
81	18	16	60	90	964032	79	4651145	58875	125867	248350	4639239

Table 2.4: Contig orientation suggested by the genome comparison of different *Pantoea* spp. as well as strains of *P. ananatis*

Draft genome	**Scaffold information suggested**
Pantoea stewartii subsp. *stewartii DC*283	29
Pantoea sp. At-9b	25
P. ananatis LMG 2665	42
P. ananatis LMG 20103 (1.0.53 version)	77
P. ananatis LMG 20103 (2.3.00 version)	23

Table 2.5: A summary of the *P. ananatis* LMG 2665 Velvet assemblies indicating the impact of a difference in hash length

Assembly #	**Hash length**	**Number of Node**	**N50**	**Max**	**Total reads**	**Total**
1	19	1926	68 272	238 433	8 900 303/ 9 601 792	4 934 378
2	21	933	108 765	328 162	8 646 391/ 9 601 792	4 936 208
3	23	782	119 599	313 830	8 327 416/ 9 601 792	4 936 769
4	25	626	106 526	277 021	7 981 474/ 9 601 792	4 939 434
5	27	578	76 038	277 741	7 615 930/ 9 601 792	4 940 416

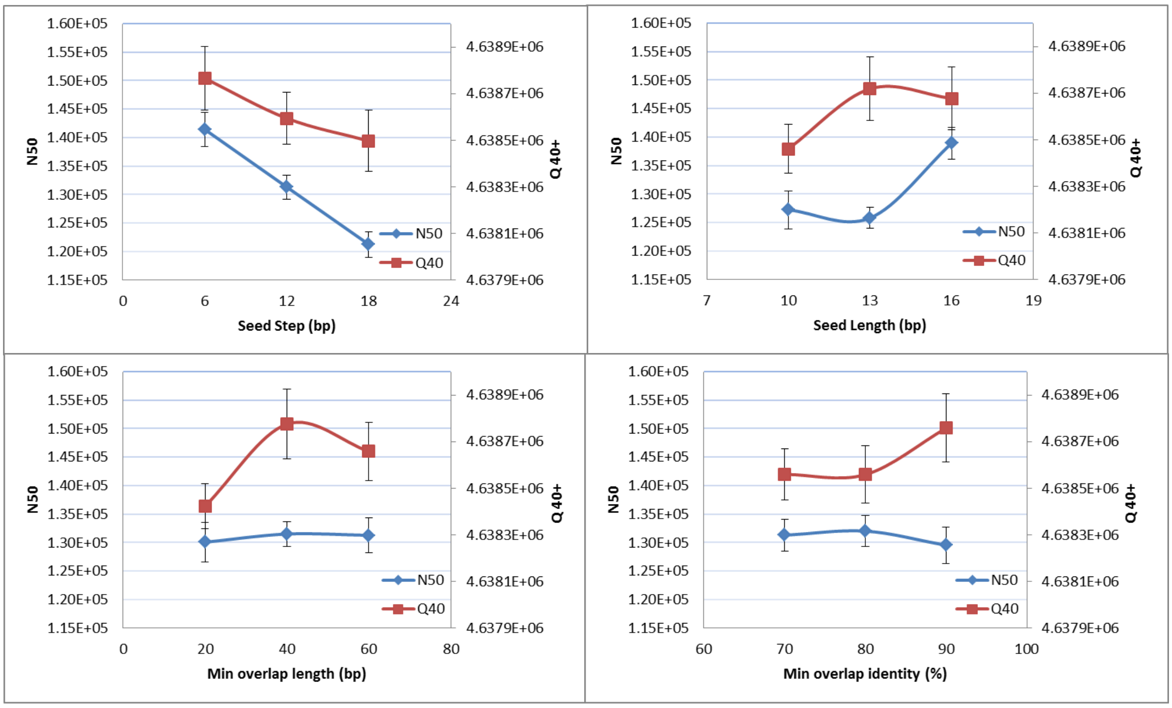


Figure 2.1: The mean values of N50 and Q40+ against the different parameter settings of Newbler (with standard error from the mean) on the *P. ananatis* LMG 20103 pyrosequence data.

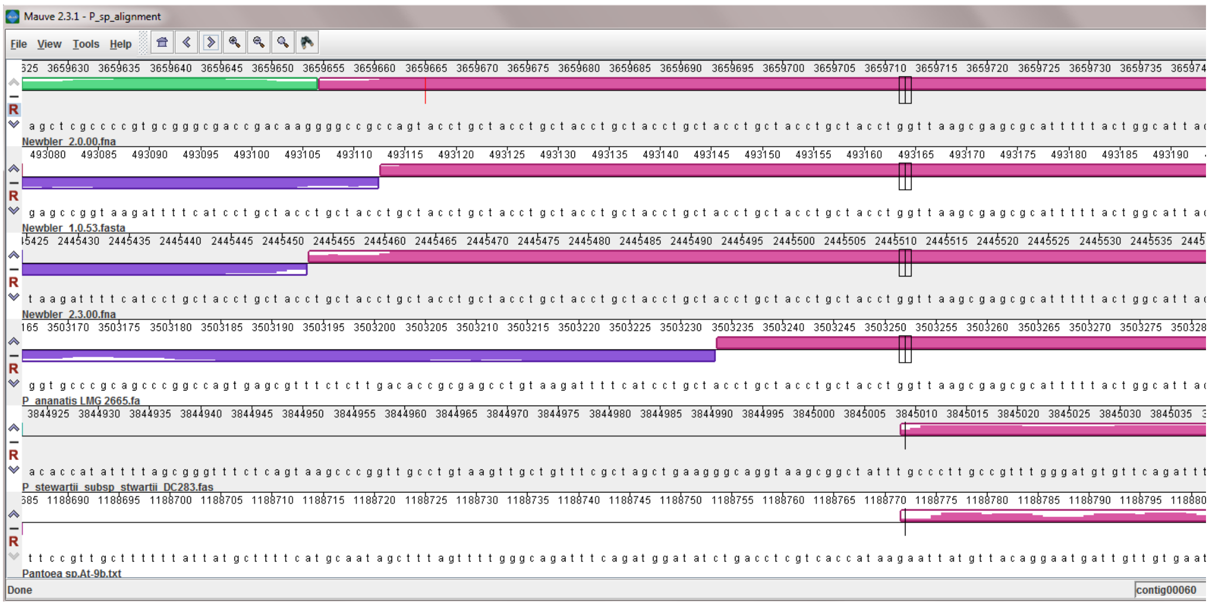


Figure 2.2: Nucleotide sequence comparison between the different draft genome assemblies as performed by Mauve 2.3.1 – Genome Alignment Visualization software. The red line divides individual contigs, and identical sequences were indicated with the same colour blocks. The draft genome sequence was arranged in the following order from top to bottom: Draft genome sequence of *P. ananatis* LMG 20103 assembled by Newbler version 2.0.00, draft genome sequence of *P. ananatis* LMG 20103 assembled by Newbler version 2.3.00, draft genome sequence of *P. ananatis* LMG 20103 assembled by Newbler version 1.0.53, draft genome sequence of *P. ananatis* LMG 2665, draft genome sequence of *Pantoea stewartii subsp. stewartii DC283* and draft genome sequence of *Pantoea sp.*At-9b.

Chapter 3

Complete genome assembly of the *Pantoea ananatis* LMG 20103

Complete genome assembly of the *Pantoea ananatis* LMG 20103

3.1. Introduction

As the fundamental processes of life are captured in an organism's DNA, scientists can analyse and utilise the complete genome sequence of an organism as a useful resource to answer many biological questions related to the targeted organism. With the advancement of next generation sequencing (NGS) technology, there has been a significant increase in the number of available genomes and genome projects which has transformed biological research into a whole new direction. With the genome data and its wide range of applications, researchers are exploring the genome sequence for gene and protein discovery. Through comparison studies between genomes, new mechanisms and evolution patterns of the organism can be revealed and the organism's interaction toward its surrounding environments can be predicted. In other words, a genome sequence is a valuable data resource for exploring the biology of the organism.

The necessity of completing or finishing a genome assembly is a debatable issue (Chain *et al*., 2009; Fraser *et al*., 2002) and depends on the scientific questions posed in the study. For example, a higher quality genome assembly is preferred for genomic comparisons related to biological diversity studies such as evolutionary and gene comparison studies as well as pathway reconstruction (Chain *et al*., 2009; Fraser *et al*., 2002; Mardis *et al*., 2002; Pop, 2009). In addition, mapping/assembly of closely related genomes can be done using the completed genome as the reference (Fraser *et al*., 2002; Mardis *et al*., 2002). As the genome sequence often forms the basis for many subsequent studies it is beneficial and recommended to complete the target genome assembly in order to create a reliable data resource (Mardis *et al*., 2002).

The aim of a finished genome assembly is to produce a correct and complete annotated genome sequence for the targeted organism. Since the first human genome assembly project was introduced, the Bermuda standard was recommended for completion of a genome. According to the standard a finished genome should consist of one contiguous sequence, with Phred quality score for each base to be higher than 30. Additional requirements of the finished sequence are that it should contain no gaps or any characters other than A, T, G and C. The resolution of all repetitive elements and mis-assembled regions should also have been attempted (Mardis *et al*., 2002; National Human Genome Research Institute, 2001 and 2002).

The Human Genome Project and other early genome assembly projects, for example, “Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project” (http://www.genome.gov/10001812) and “Human Genome Sequence Quality Standards” (http://www.genome.gov/10000923) provided some guidance for the completions of genomes. These guidelines were originally addressing low-throughput genome assembly data, where the assembly relied on low coverage and long cloning library sequences for the assembly. Maintaining the golden standard has become a challenging task as genome research switched from low-throughput sequencing to high-throughput techniques and finishing the genome remains one of the most time consuming and expensive steps in the genome assembly process (Chain *et al.*, 2009).

Most assembly errors are caused by repetitive elements. These repeat sequences can be tandem repeat elements, polymorphisms, transposon elements or long terminal repeats within the genome (Phillippy *et al*., 2008; Quinn *et al*., 2008; Wicker *et al*., 2006). Consequently, they are associated with mis-assembly as they may stack together leading to breakage of contigs, deletion, false genome arrangement and false inversion within the draft genome assembly (Phillippy *et al*., 2008). To resolve these undesirable events, different completing strategies were designed for gap closure, to identify and resolve mis-assembly regions and repeats sequence within the draft genome assembly (Nagarajan *et al*., 2010; Phillippy *et al*., 2008). As the sizes of the targeted genomes are increasing, the chances for mis-assembly have also increased. To close the gap between genome assembly and completion, both biological and bioinformatics solutions have been developed, including both *in-vitro*, and *in-silico* approaches.

The *in-vitro* approach, which uses the traditional method of polymerase chain reaction (PCR) and sequencing for gap closure, is still the most reliable solution (Roux, 1995). Optical mapping is another approach followed to achieve a finished genome which relies on a number of restriction enzyme digestions of the genomic sequence. Through correlating the digested fragments and their estimated sizes the correct order of each contigs could be determined (Latreille *et al*., 2007; Nagarajan *et al*., 2010).

With the *in-silico* approach, a number of computational analyses such as mate-pair validation, coverage analysis and repeat analysis are commonly applied as part of the completion step in genome assembly (Phillippy *et al*., 2008). Mate-pair validation is used to resolve repeat sequences and the scaffolding of contigs (Phillippy *et al*., 2008; Pop, 2009). To reduce the number of mis-assemblies within the draft genome, most of the newly sequenced genome projects include mate-paired reads. The major advantage of paired-end reads is that they can reveal the order and the correct distance/gap size between adjacent contigs. At the same time they can also resolve repeat sequences that were stacked together, while normal single read assemblies cannot resolve this problem (Phillippy *et al*., 2008; Pop, 2009; Wetzel *et al*., 2011).

Other computational analyses can also assist with the completion of genome assemblies when mate-paired reads are not available. Coverage analysis helps with the identification of repeat sequences and mis-assembled regions. This coverage analysis data is often provided within the statistical analysis report of the genome assembly. Through coverage analysis, an estimation of the number of repeat/polymorphism copies can be determined. DNPTapper is a validation program, which uses both the coverage and SNPs information within the collided repeat sequences to identify and resolve the polymorphism sequences (Arner *et al*., 2006). Contig adjacency information can assist with scaffolding of the contigs. This information is often included in the 454 contig graphic file, where the contigs' order or scaffolding informatics is provided to aid with completion of the assembly (Nagarajan *et al*., 2010). One should keep in mind that computational analyses are not 100% reliable and manual inspection by the curator of difficult regions is still required.

Visualisation tools that support the inspection of the genome assembly include Gap5, Consed, EagleView and Tablet (Bonfield and Whitwham, 2010; Gordon *et al*., 1998; Huang and Marth, 2008; Milne *et al*., 2010). For detailed inspection or extensive modification of the genome assembly, additional functions or other visual aids have been employed. This includes coverage graphs for coverage analysis, quick search functions for low quality score base reads, and repeat sequence searches for possible reads that span between contigs. Pipelines or software (e.g. Gap resolution) for assembly validation (DOE Joint Genome Institute; Phillippy *et al*., 2008) are in high demand. Furthermore, primer design, SNPs identifier and other useful built-in packages have been included in some of these programs, to speed-up the process of completion of the genome assembly (Bonfield and Whitwham, 2010; Gordon *et al*., 1998).

As a fully assembled genome sequence of high quality is required for taxonomic and evolutionary studies as well as the metabolic network re-construction of the organism, the aim of this part of the study was to complete the genome assembly of *P. ananatis* LMG 20103. In this chapter, the nature of the gaps, the cause of contig breakage and possible assembly errors were investigated. Since mate-paired reads were not available both *in vitro* and *in silico* approaches were used to create a reliable genome assembly.

3.2. Materials and Methods

In Chapter 2, optimisation of the Newbler assembly was performed, resulting in a draft genome of *Pantoea ananatis* LMG 20103 consisting of 117 contigs of which 75 were larger than 500 bp. This genome served as the input data used during the completion of the genome assembly which will be discussed in this chapter.

3.2.1. Coverage and adjacent contig analysis

Assembly analyses were undertaken to identify indicators for mis-assembly at the edges of the contigs. Bioinformatics software such as Gap4, Tablet, Consed and Eagleview were used for the analysis. The .ACE assembly file was loaded onto Tablet, Consed and Eagleview directly, while a library was generated from the .ACE file to run on Gap4.

Coverage graphs and diagrams provided by Consed and Gap4 were used for quick detection of low and high coverage regions. Search function targeting overlapping or repeat regions in Consed and Gap4 were run to identify contigs that could be joined. The edges of the individual contigs were examined for the presence of SNPs, coverage variation and reads that could bridge between two contigs. Reads that could bridge between contigs were duplicated and renamed by Newbler in the .ACE file, with either the word “to” or “from” attached to the original read name. The scaffolding information of the contigs stored in the 454ContigGraph.txt output file was also checked. The collected details were drawn onto a contig graph, so that the scaffold of each contig could be reviewed.

3.2.2. Gap closure

3.2.2.1. Polymerase chain reactions (PCR)

The culture of *Pantoea ananatis strain* LMG 20103 (Coutinho *et al*., 2002), was obtained from the Bacterial Culture collection of Forestry and Agriculture Biotechnology Institute (FABI). The culture was incubated overnight at 28 ^{o}C on nutrient agar and the genomic DNA of this strain was extracted with Quick-gDNA MiniPrep kit (Zymo research) for use in subsequent PCR reactions.

After the scaffolding of the contigs was done, PCR sequencing was used to resolve problematic regions (e.g. orientations conflict, low coverage region) of the draft genome. Depending on the characteristic of the gap regions, different approaches were used to resolve the gaps within the draft genome assembly. This included different types of polymerase chain reactions, contigs edge extension and resolution of long repeats sequences.

DNA sequences that are 500 bp away from the edges of a target contig were extracted (Figure 3.1) and compared with the draft genome assembly, to identify unique sequence regions for primer targets. The primer sequences were then based on these unique regions, preferably 200 bp to 500 bp away from the edge of the large contig. The primer sequence regions were inspected with genome visualisation tools during primer design for any coverage variation and to avoid possible repeat regions. For gaps that were larger than 1000 bp, primer walking was done, and internal primers also had to be designed.

Additional criteria were applied during primer design. Primers were not designed for any contigs that were smaller than 500 bp in size. Primers varied in size between 18 – 25 bp, and had melting temperature between 55 ^{o}C to 62 ^{o}C. The online program OligoAnalyzer 3.1 (http://eu.idtdna.com/analyzer/applications/oligoanalyzer/) was used to ensure no secondary structures would form. All the primers were synthesized at Inqaba (Table 3.1).

Conventional PCR reactions were used first as most of the gaps were expected to be small. The PCR was run with half reactions of 2.5 µl of 10X reaction buffer, 2 µl of 25 mM $MgCl_2$, 2 µl of 10 mM primer I and primer II, 0.15 µl of 5u/µl of Super-Therm Polymerase and 100 ng of genomic DNA template. The PCR cycle consisted of denaturation at 95 ^{o}C for 5 minutes, follow by 25 cycles with 95 ^{o}C for 5 seconds, 60 ^{o}C for 30 seconds, 72 ^{o}C for 30 seconds and final elongation at 72 ^{o}C for 7 minutes.

Gaps that were not resolved by normal PCR were subjected to Long-range PCR. This approach was used for gaps larger than 4 kbp. The Long PCR Enzyme Mix (Fermentas) was used. The half reaction was made up of 2.5 µl of 10X long PCR Buffer with $MgCl_2$, 2.5 µl of 2 mM dNTP mix, 2 µl of 10 mM primer I and primer II, 0.1 µg of Genomic DNA template and 0.25 µl of Long PCR Enzyme Mix. The PCR cycle consisted of denaturation at 95 ^{o}C for 5 minutes, follow by 25 cycles with 95 ^{o}C for 5 second, 60 ^{o}C for 30 seconds, and 62 ^{o}C for 4 minutes and 30 seconds, followed by a final elongation at 62 ^{o}C for 7 minutes.

Optimisations of the PCR reactions were done for problematic gap regions with poor PCR amplification or poor sequencing results. For GC rich regions, normal PCR reactions were performed with the addition of 4 % dimethyl sulfoxide (DMSO) (Frackeman *et al*., 1998) added to the PCR reaction mixture before the PCR cycle.

The PCR products were cleaned with exonuclease Eco I (Fermentas) and FastAP (Fermentas) enzymes to remove the remaining dNTP and polymerase. The PCR product was then sequenced in a 1/16 sequencing reaction, the 12 µl reaction consisted of 2.5 µl of Sequencing buffer, 0.5 µl of Big Dye 3.1, 0.3 µl of the 100 mM primer (either one of the PCR reaction primer), and 4 µl of clean PCR template.

The PCR cycle for sequence was then carried out with an initiation temperature of 95 ^{o}C for 5 seconds, follow by a 25 cycles of 94 ^{o}C for 10 seconds, 55 ^{o}C for 10 seconds and 60 ^{o}C of 4 minutes.

The sequencing products were cleaned by sodium acetate precipitation (NaOH/EtOH), where the 12 µl of product was precipitated with 16 µl of 100% ethanol and 2 µl of 3 M of sodium acetate after 30 minutes of centrifugation at maximum speed. This was then followed by a double wash step with 150 µl of 70 % ethanol and centrifuge at maximum speed for 5 minutes. The excess ethanol was removed and the pellet was dried on a 90 ^{o}C heating block for 5 minutes. The final product was sequenced at the sequencing facility of the University of Pretoria.

For the sequence analysis, the sequencing results were compared against the draft genome assembly database using BioEdit. Poor sequencing regions were clipped off, and all the sequences were collected in a fasta format file. This data was later used during the reassembly of the genome.

3.2.2.2. Contig edge extension

Contigs that were not joined or scaffolded by PCR sequencing were subjected to contig edge extension. By extending the partial aligned hidden reads at the end of the contigs, hidden information of the gaps can be obtained. The raw 454 reads were extracted with a python script: sff_extract.py, where the raw reads were converted into fasta file format. A local database was created with the converted fasta file using the accessory application of BioEdit.

The search for unassembled reads that could align to the hidden sequences at the edge of the contig was done by using a local blast alignment (Altschul *et al.*, 1990) search of the partially aligned reads against the raw 454 read database in BioEdit or Genious. The identified reads were extracted and then assembled using either Genius or manually in BioEdit based on the high similarity between the sequences. These regions were checked for errors and a further decision on the strategies to resolve these regions was made depending on the character of the gap regions. Tandem repeats were typically joined.

3.2.3. Genome re-assembly

With the original raw data and the longer additional reads, an attempt to reduce the number of gaps and resolve the repeat region during the re-assembly was performed. The .abi sequences generated from the PCR for gap closures were converted into fasta file format, after the poor sequences section were clipped off. These sequences were collected in a single fasta file as single long reads with sizes ranging from 300 bp to 700 bp.

Using the default settings, the original raw data file .sff along with the long reads fasta file were submitted to Newbler assembler for re-assembly. Other repeat sequences, such as the resolved rDNA sequences were also submitted for re-assembly. The newly generated assembly sequence and the original draft genome assembly sequence were both submitted to Mauve (Darling *et al.*, 2004) for DNA sequence comparison and to verify contig joining.

3.2.4. Resolving the ribosomal DNA fragments

With any bacterial genome, the largest repeat sequence to be expected is the ribosomal DNA. Each individual copy has an approximate size of 5.5 kbp, with a slight variation between the different copies. Without mate-paired reads, these copies are often assembled together as contigs with an unusually high coverage. The identification of these rDNA contigs was done using a blast alignment of the large contigs sequences against the NCBI database. The number of rDNA copies was estimated by dividing the coverage of the identified rDNA contigs with the over-all coverage of the draft genome. The contigs that connect to the rDNA were located by identifying the reads at the edge of the rDNA contigs that caused branching of the contig. Primers were designed based on the targeted contigs and normal PCRs were performed to confirm the linkage between the contigs and the rDNA. Once the confirmation of the contigs was completed, optimised long-range PCRs were then used to resolve the position of the rDNA copies within the genome.

3.2.5. Final/manual assembly

The remaining gaps that could not be joined by Newbler were merged manually, once the correct gaps closure sequences were obtained. The overlapping regions of the long sequences were located using a local blast alignment against the remaining super-contigs. The gaps sequences were then inserted in-between the super-contigs.

3.3. Results

3.3.1. Genome assembly analysis and contig adjacency information

Thorough analysis of all 117 contigs from the draft genome assembly of *P. ananatis* strain LMG 20103, as generated by Newbler, was performed. Using the genome visualising program, the contig boundaries were grouped as follows:

- Low coverage (coverage below 14) (Table 3.2)
- High coverage with the presence of multiple SNPs (stacking of repeat sequences) (Figure 3.2)
- High coverage without SNPs (possible collapse of identical tandem repeats) (Figure 3.3)

During the inspection of the contigs, reads that could bridge two contigs were duplicated and renamed by Newbler in the .ACE assembly file with the addition of characters (e.g. ENC8TDH02GXHF2**.18-1.fm91**). The information of the reads was also found in the “Contig graph.txt” or in the .ACE file. Contigs with consistently high coverage were also identified, and these typically ranged from 85 bp to 3 kbp in size. According to the coverage graphs, the highest coverage for the draft genome assembly was 215x.

A contig graph that could aid with completing the genome assembly was constructed using Microsoft Visio (Figure 3.4). Each contig was represented by a node, where additional information with regards to the length of the contig, the coverage of the contig boundaries, the direction of the contig, the presence of tandem repeat sequences, the estimated copy number of repetitive elements and the contig adjacency information were incorporated into the graph. Each node was rearranged according to the adjacency information, and linked with edges between the nodes. A number of nodes had multiple edges that branched off to different nodes. Conflicting edges between nodes were also observed on the contig graph (Figure 3.4). Cyclic linkages present in the diagram were found to be associated with contigs that have repeative sequence at their boundary edges and in some cases, inverted repeat sequences were also observed at the end of these contigs.

One hundred and seven possible links between neighbouring contigs were recognised (Figure 3.4). Some of the contigs' edges had more than 1 possible linkage or rather branched off to more than one contig. The orientation of the remaining contigs was unknown. However, by including the comparison results between closely related draft genomes performed in the previous chapter, a total of 109 contigs could be scaffolded. Once the orientation of the contigs had been determined, experiments were performed to confirm this information

3.3.2. Gap closure with polymerase chain reactions (PCR)

Eighty three newly designed primers, as well as a number of additional primers that were previous designed by other students, were used during the gap closure process. The list of primer sequences and the set of primers that were used for gaps closure PCR are listed in Table 3.3. After the poor quality regions of the PCR sequences were clipped and these sequences were aligned against the draft genome sequence, 186 of them overlapped with the targeted contig edges of the draft genome assembly. A total of 49 gap sequences were filled by using this PCR and sequencing approach. The size of the gaps ranged from 5 bp to 4500 bp. Gaps regions with the absence of contig adjacency information, PCR sequencing confirmation and reads during the boundaries extension were considered as low complexity gap regions. Two low complexity (G/C rich) gaps were confirmed and closed after optimisation of the normal PCR reaction and sequencing.

3.3.3. Contig boundary extension

The un-scaffolded contigs were inspected and attempts to extend their boundaries were made. Through the examination of the edges of these contigs, some of the missing gap sequences were recovered from either the unused 454 reads or within the draft genome's high coverage regions. The identified unused reads were typically highly similar repeat sequences with some degree of diversity, as small insertions and deletions were found within these repeat sequences. These reads were extracted, but failed to be assembled using Bioedit or Genious. The extracted reads could, however, be clustered into separate groups based on similarity and were then manually assembled into the draft genome based on the alignment similarity between the contigs boundary sequences. Through boundary extension, a total of 12 gaps were closed manually with this approach.

3.3.4. Automated Genome re-assembly

Instead of inserting the new PCR sequences manually into the draft genome, a genome re-assembly was performed using Newbler. By comparing the re-assembled draft genome sequence to the original draft genome sequence in Mauve, it was found that some of the regions, including short repeat regions and tandem repeats regions could be resolved and closed by the assembler, as long as the long read sequences had an overlap within the draft contig edges and spanned over the entire gap regions.

Unexpectedly 9 more gaps were joined during re-assembly by Newbler, as the long reads had resolved some of the conflicting regions and the expected coverage value was provided. On the other hand, repeat sequences that were larger than 1200 bp became difficult to be resolved by the assembler as reads that are longer that 2 kbp were rejected by Newbler during the assembly. Eight tandem small repetitive regions were joined by Newbler as the Minimal overlap length decreased and other tandem repeats were manually joined together.

3.3.5. Location and resolving of the ribosomal DNA fragments

With the local blast alignment function of NCBI, contig 71, 9, 115, 55 and 113 were identified as being part of the ribosomal RNA fragments. The 16S small rDNA subunit was made-up of three contigs: contig 71, 9 and 115. The 16S rDNA subunit was separated into 3 contigs, the regions that were not assembled appeared to have ambiguous bases or small inserts. The large 23S rDNA subunit was present in contig 55 and the 5S rDNA gene was represented as one small fragment in contig 113. Coverage analyses on the rDNA contigs were done, and the coverage ranged from 125x to 215x. To determine the copy numbers of rDNA, a coverage estimation was done, and based on this calculation 6 to 8 copies of the rDNA units were expected within the *P. ananatis* genome.

Neighbouring contigs to the rDNA genes were identified from the contig adjacency information. Further confirmations were done by normal PCR using 16S/23S internal primers and targeted contigs primers. Fourteen contigs were confirmed to bridge directly to the rDNA sequence, where seven contigs were linked to both the 23S and 16S rDNA subunit, respectively (Figure 3.5). Eight copies of the 5S rDNA subunit were also resolved by means of the PCR sequencing results.

Contigs 4, 5, 6, 7, 15, 26, 27, 29, 31, 53, 54, 56 and 85 were identified as those that were linked to the rDNA subunit. The rDNA repeat sequences were resolved with optimised long range PCR. Partial sequences for the ambiguous regions were obtained for each copy of the rDNA sequence. Both the PCR sequences and the rDNA contig sequences were collected, and assembled separately using Genius. The resolved rDNA sequences were approximately 5.5 kbp long. During the re-assembly, the 7 copies of the rDNA sequence were submitted but Newbler was unable to assemble them, as their sequences were larger than the read limitation of 2 kbp in size.

After the PCR gap closure, contig boundary extension and re-assembly of the genome, the 8 super-contigs were scaffolded together. One of the super-contigs could be closed to form a complete circular piece of DNA. The remaining 7 super-contigs were separated by the rDNA sequences. The correct rDNA sequences were manually added to the genome assembly. All 117 contigs from the initial draft genome assembly were therefore assembled together. Homopolymer sequencing errors were corrected through genome annotation. Open reading frames were predicted by Glimmer, while the annotation was predicted by BASys (Delcher *et al*., 2007; Guindon *et al*., 2004). Genes that were larger than 50 aa were further curated for any frame shift errors. The final genome assembly has one completed circular chromosome and a megaplasmid, with the size of 4 386 227 bp and 317 146 bp, respectively. This non-contiguous finished genome (De Maayer *et al*., 2010) was updated in the NCBI database.

3.4. Discussion

The rate at which new genomes are sequenced is substantially higher than the rate at which genomes are finished and released (Gordon *et al*., 1998). This raises the questions of whether it is necessary to finish a draft genome sequence and what benefits such an action would have. Genome assembly programmes have the ability to greatly reduce the time and effort required to assemble the raw sequence data into a draft genome. The draft assembly produced using a genome assembly programme will typically still contain numerous gaps or mis-assembled sequences. Draft genomes are useful and can be used to explore the metabolic diversity of an organism, but a complete genome is usually required for comparative and functional genomics as well as studies investigating genome evolution (Fraser *et al*., 2002). It was, therefore, decided that in spite to the extra time and effort required, it was important that the draft *P. ananatis* LMG 20103 genome should be finished. This genome was one of the first representatives of the genus *Pantoea* to be sequenced and would form part of a number of other projects on the evolution, pathogenicity and ecology of this bacterium.

The types of problems experienced when attempting to finish a genome differs depending on the nature of sequencing technology and the assembly algorithm used afterwards. The typical assembly problems associated with the draft genome assembly of *P. ananatis* LMG 20103 included individual contigs with a consistently high coverage over the entire contig, as a result of repeat sequences that were stacked together (Wicker *et al*., 2006). Large repeat units that could not be resolved were also observed. According to Wicker (2006) as well as Han and Chain (2006), repeat regions that are larger than the read length are typically problematic regions and often remain unresolved. This phenomenon has previously been observed during genome assembly with Newbler, where the extension of a collapsed region always terminated with conflicting or ambiguous sequences (DiGuistini *et al*., 2009; Pop, 2009).

The largest repeat sequence in the genome of *P. ananatis* LMG 20103 was the 23S rDNA subunit sequence, which was assembled into one contig with the size of 3 kbp and a coverage of 215x for this part of the contig. The presence of multiple SNPs within this contig served as a positive indicator for stacked polymorphisms (Arner *et al*., 2006; Nowrousian *et al*., 2010). Another indication was that the edges of these contigs branched-off to a number of contigs. The 16S rDNA region was treated differently by Newbler. Due to the presence of sequence ambiguities between the 7 copies of 16S rDNA subunit these regions were, therefore, also collapsed but the conserved regions were represented by three smaller contigs in the draft assembly.

In addition, it was also found that repeat sequences with additional indels were difficult to assemble by Newbler, and these repeat sequences were left as unused reads after the assembly took place. It appeared as if this earlier version of the assembler lacked the ability to distinguish between the different copies. These repeat sequences could be assembled by the assembler, once the sequences were either divided into different clusters or when a newer version of the Newbler assembler was used.

The presence of non-repeat associated gaps with a low coverage character in the draft assembly was suggestive of a shortage of reads for use by the assembler. The lack of reads might have been due to the mis-assembly of reads or inconsistent coverage of the area. On close inspection of the data and PCR verification, it was concluded that most of the low coverage gap regions (coverage between 12x to 15x) observed in the draft assembly were not necessarily caused by mis-assembly. Some of the contigs were split by Newbler as a result of the assembler's stringency as it considered regions with a coverage below the threshold value to be of poor quality and excluded them during the assembly (Wicker *et al*., 2006).

With the recent development in genome research, mate-paired reads have been heavily relied upon during the *de novo* genome assembly of sequence data generated by next generation sequencers. Contig scaffolding and repeat sequences can be resolved easily with this method and reduce the amount of effort needed to complete the assembly process. However, without the benefit of a paired-end library for the genome assembly, it has become a challenge to complete a *de novo* assembled draft genome. Alternative or more traditional methods had to be considered during finshing of the *P.ananatis* LMG 20103 genome assembly.

One of the available resources for contigs scaffolding was contig adjacency information. This information was collected from the 454 contig graph.txt file or after comparing the assembly with closely related genomes. In the 454 contig graph.txt file, information with regards to the overall coverage of each and every individual contig and suggested links between contigs were provided and assisted with the quick identification of large collapsed repeat sequence e.g. rDNA sequence. A visual representation of the contig orientations in the contig graph also helped to find appropriate strategies to resolve the different assembly problems. This information was also used for the design of primer sets, as it provided information on which contigs to avoid during primer design, the approximation of the gap sizes, the possible branching points for the repeat sequences as well as suggestions for possible insert sequence that may be present between repeatitive elements.

In this study PCR assays and sequencing were used for both gap closure and assembly validation with great success. PCR reactions targeting specific areas were used to validate the order of the contigs, provide the missing sequences for gap closure, and to resolve mis-assembly caused by polymorphisms. For gap closure, primer pairs had to be designed to fall within unique sequence regions, in order for the PCR product to cover the gap-flanking sequences (Garber *et al*., 2009). This ensured that the amplified PCR product covered the targeted gap and the extra gap-flanking sequence helped with re-assembling the genome. With the extended length of the PCR sequencing (600 to 700 bp), the sequenced PCR products provided a backbone for the assembler to resolve small repetitive regions, and assist with the resolution of a number of small tandem repeats and repeat sequences.

The re-assembly with Newbler was performed to ensure that the long single reads were assembled correctly into the draft genome and that many of the gaps were closed. By doing so, the amount of time and effort required to manually add the sequences and merge the contigs together was reduced, as a number of tandem repeats and repeat sequences, along with gap sequences could be resolved using this approach. There were, however, limitations with regards to re-assembly. Reads provided by PCR sequences or other resolved repeat regions were restricted to 2 kbp in size (454 Life Science, 2008). Therefore not all the mis-assembly or gap problems were re-solved, as was the case with the seven copies of rDNA sequences/genes within the *P. ananatis* genome. These gaps that were not resolved by the automated re-assembly were curated manually.

3.5. Conclusions

Finishing a genome assembly is a painstaking task. During this study, repeat sequences and low complexity regions were found to be the most common cause of assembly errors as they resulted in gaps due to the termination of contig extension during assembly. Due to the short read length from the pyrosequencing data, most of the gaps and mis-assembly were the direct result of repeat sequences, as they were stacked together and could not be resolved without the help of additional information.

Improvements to the pipelines used for the finishing of genomes are far from perfect. As the validation of genome assembly can be done with both *in-vitro* and/or *in-silico* approaches, new bioinformatics pipelines and their different analyses such as coverage analysis, SNPs analysis and paired-end read information are capable of pin-pointing the mis-assembly error and the problematic regions can be inspected. Though, undetectable low coverage repeats can still cause assembly errors within the complete genome. While PCR is still the most reliable method for the confirmation of contigs orientation and gap closure.

Without any paired-end reads, conversional PCR assays and contig adjacency information were used as the primary source of information to assist with gap closure and final assembly of the contigs. Depending on the character of each and every individual gap, a range of conventional PCRs as well as long-range and high GC PCRs were required to resolve the gap regions. Visualisation of the draft assembly, by means of a "contig graph" and other genome visualisation tools, assisted greatly in the design of these PCRs and are highly recommended for use in future genome finishing projects.

Using a combination of various *in vitro* and *in silico* approaches as described in this chapter it was possible to assemble all 117 contigs from the initial draft assembly. This was followed by the prediction of open reading frames using Glimmer and annotation of the genome with the help of the BASys pipeline (Delcher *et al*., 2007; Guindon *et al*., 2004). The final genome assembly has one completed circular chromosome of 4 386 227 bp and a megaplasmid with a size of 317 146 bp. Despite the amount of effort and cost, it is believed that the complete genome will be a valuable resource for many subsequent studies to investigate the evolution and biology of this emerging plant pathogen.

References

454 Life Science., 2008. Genome Sequencer Data Analysis Software Manual, Software Version 2.0.00., A Roche company, Branford, CT.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410.

Arner, E., Tammi, M., Tran, A.N., Kindlund, E., and Andersson, B., 2006. DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. BMC bioinformatics 7, 155.

Bonfield, J.K., and Whitwham, A., 2010. Gap5—editing the billion fragment sequence assembly. Bioinformatics 26, 1699–1703.

Chain, P.S.G., Grafham, D.V., Fulton, R.S., FitzGerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., Cole, J.R., Ding, Y., Dugan, S., Field, D., Garrity, G.M., Gibbs, R., Graves, T., Han, C.S., Harrison, S.H., Highlander, S., Hugenholtz, P., Khouri, H.M., Kodira, C.D., Kolker, E., Kyrpides, N.C., Lang, D., Lapidus, A., Malfatti, S.A., Markowitz, V., Metha, T., Nelson, K.E., Parkhill, J., Pitluck, S., Qin, X., Read, T.D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R.L., Sutton, G., Thomson, N.R., Tiedje, J.M., Weinstock, G., Wollam, A., Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, and Detter, J.C., 2009. Genome Project Standards in a New Era of Sequencing. Science 326, 236 –237.

Coutinho, T.A., Preisig, O., Mergaert, J., Cnockaert, M.C., Riedel, K.H., Swings, J., and Wingfield, M.J., 2002. Bacterial blight and dieback of Eucalyptus species, hybrids, and clones in South Africa. Plant disease 86, 20–25.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T., 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Research 14, 1394–1403.

De Maayer, P., Chan, W.Y., Venter, S.N., Toth, I.K., Birch, P.R.J., Joubert, F., and Coutinho, T.A., 2010. Genome Sequence of *Pantoea ananatis* LMG20103, the Causative Agent of Eucalyptus Blight and Dieback. Journal of Bacteriology 192, 2936–2937.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L., 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23, 673.

DiGuistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., Mardis, E., Marran, M.A., Hamelin, R.C., Bohlmann, J., Breuil, C., and Jones, S.J.M., 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10, R94.

DOE Joint Genome Institute. Software Tool. Available at: http://www.jgi.doe.gov/software/. [Accessed April 05, 2011]

Frackman, B.S., Kobs, G., Simpson, D., and Storts, D., 1998. Betaine and DMSO: enhancing agents for PCR. Promega notes 65, 9–12.

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

Garber, M., Zody, M.C., Arachchi, H.M., Berlin, A., Gnerre, S., Green, L.M., Lennon, N., and Nusbaum, C., 2009. Closing gaps in the human genome using sequencing by synthesis. Genome biology 10, R60.

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O., 2004. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Research 33, W557 –W559.

Gordon, D., Abajian, C., and Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome research 8, 195–202.

Han, C.S., and Chain, P., 2006. Finishing repetitive regions automatically with Dupfinisher. In Proceeding of 2006 international conference on Bioinformatics and Computational Biology: 2006. Las Vegas, Nevada, USA. CSREA Press.

Huang, W., and Marth, G., 2008. EagleView: A genome assembly viewer for next-generation sequencing technologies. Genome Research 18, 1538–1543.

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

Mardis, E., McPherson, J., Martienssen, R., Wilson, R.K., and McCombie, W.R., 2002. What is Finished, and Why Does it Matter. Genome Research 12, 669–671.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D., 2010. Tablet—next generation sequence assembly visualization. Bioinformatics 26, 401–402.

Nagarajan, N., Cook, C., Di Bonaventura, M.P., Ge, H., Richards, A., Bishop-Lilly, K.A., DeSalle, R., Read, T.D., and Pop, M., 2010. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. BMC genomics 11, 242.

National Human Genome Research Institute. 2001. Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project. National Institutes of Health. Available at: http://www.genome.gov/10001812 [Accessed June 22, 2011]

National Human Genome Research Institute. 2002. Human Genome Sequence Quality Standards. National Institutes of Health. Available at: http://www.genome.gov/10000923 [Accessed June 22, 2011]

Nowrousian, M., Stajich, J.E., Chu, M., Engh, I., Espagne, E., Halliday, K., Kamerewerd, J., Kempken, F., Knab, B., Kuo, H.-C., Osiewacz, H.D., Pöggeler, S., Read, N.D., Seiler, S., Smith, K.M., Zickler, D., Kück, U., and Freitag, M., 2010. De novo Assembly of a 40 Mb Eukaryotic Genome from Short Sequence Reads: Sordaria macrospora, a Model Organism for Fungal Morphogenesis. PLoS Genetics 6, e1000891.

Phillippy, A.M., Schatz, M.C., and Pop, M., 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome biology 9, R55.

Pop, M., 2009. Genome assembly reborn: recent computational challenges. Briefings in bioinformatics 10, 354–366.

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

Roux, K.H., 1995. Optimization and Troubleshooting in PCR. Cold Spring Harbor Protocols 2009, pdb.ip66.

Wetzel, J., Kingsford, C., and Pop, M., 2011. Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. BMC Bioinformatics 12, 95.

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

Table 3.1: List of primer used for gap PCR

F46	CCTTTCCTGACGATTTGAACGGC
F320	ATGACCAGAAGTACCTTGAACAGG
F330	CGCGGAATCTGTGGTTTTGGC
F300	CCGTGTGGAGGAGTAATGGC
F332	AGGACGGATAAATTTGGTGGCTCG
F308R	GAATGGTACGCCGGAACTGC
F264	AGACATTAGCCAGCAGCTTTATGAGAATGC
R281R	TGGCCAGCAGATAAAGCTCGC
R302	GTGAGTGCAATGAAGGGGCG
R242R	GATGCGATTGTGACCCAACTGGC
R274	GGTACAACATACCGGGCTGCG
R326R	ATCTGCACGCACATAGATCACCC
R329	TGAAGTGGTGGCCAGACTGG
R352	TTGCTTGCTGATGAGTACGTGCG
R353R	CTCAACCAGGCGATCTTCAACG
contig 093	CTTCAGAGCACAACAATTCAG
Contig 002R	CGAATGATAGGCAGGATGGT
Rev 333	CGCAGTCATGCGATTCGCTTTGCC
Rev 320	ACTACCGTGCATTGCAGGCCTTCC
Rev 330	TCTGCTACAGTAATGACGGGTGG
Rev 300	ATATTTCGGGACTTCCCGACACG
332	ATGATGGAGCAGAGATGGAGC
Rev 332	AGCGTCTTCATCTTATCCTCCCCG
Rev 308	AGGTGAAAAGAGAATTTGGAAGGCG
Rev 347	ACGCGAATGAGTCAAGTCCGGTCG
Rev 264	AGCGGTAGAGAAGTAGAAGGGCG
Rev 281	TCCGTTCGTGAAGACAGCCG
302	ACGGCTCGCCCACCGCAATAACC
Rev 302	TACTTATTCGCACCGTTCACGG
Rev 242	TAGTTGGTACTGCGCGTCTGG

Rev 286	TAGTGGCGATACTAATTTAAGGGTCTGAC
Rev 329	TGGTCGGGTCTTGGGGAACTATACC
Rev 352	ACGACCGAATTCGCATTGACC
Rev 353	TCACGTcTGTGTTATTAATACGGG
R286	CCCTAAAACATAGCTTTGAAAGTTGTGGGC
F333	TCGTGAGGGATGTCGTGCAAAGC
F347	TGCAGGGGATCTTTAGAGGCG
contig 002-096F	GATCTACTCACCCTGCGAG
contig 002-096R	CTCAATACCTCAACCTCCC
contig 033-003F	GCGTGGTTCTTAAATTGTCG
contig 033-003R	GATAAAGCCATCAAGGCCGT
contig 003-341F	GAGAAGGTAATCTGGATGGC
contig 003-341R	GGTCACGGCAATACTAACC
contig 005-039F	ATGTCGCTTCAGGATCAATACG
contig 005-039R	TGGTTGTCATCTTCTCCTCTG
contig 344-011F	GGCGGATGGAGATTAACCTG
contig 344-011R	CCTAACAGCAAATATAACCACAG
001-347F	TTTGAGAAGCACAACCACCT
001-347R	CTCAGCGAAGCCAGATAGAG
088-006F	CACCTGTTAAGCCCTTATCG
088-006R	CCCTAGTCTCTGACTATTCG
297-015F	TTCAATACCACCGAACCTACTG
297-015R	TTCGTTACCCATTCAC
015-062F	CGTTAAGCAGAATCCAGATGAC
015-062R	GAAGATGGCTCCTCTGACTG
296-019F	GTAATCTGGCCTACGGTCAG
296-019R	CCGATGAAGATCAGCACCAG
C55F	TTTTACGCTGCGTGTCATGAGC
c55R	AAGGCATTCGGGGGAGGC
c15-113	GGGAAGTCAAAGAGGATTGGG
c26-113	TTATGCCCACGTAACCGC
c27-113	GTTTTCAGTGCCCAAAGGG

c4-113	CGATGGCGTTTGTTGTATCACG
c56-113	GCCCTTTAACGCCGTGTATTAC
c5-71	CGCGGAATCTGTGGTTTTGGC
c6-71	TTCATTATGCGGATGACTGGCGG
c53-71	CAAGTCTTGCTAAGCTTGGCC
c63-71	GCGCGCCTAAATTAGTAACTCGC
N53-71	CGACCCTTACTATCGTCCAC
N63-71	CGCCTCTTTCAAAGACTACGTGGT
c85-71	ACTCGAAAACCCGCATTGAG
c80-115	CAGACACGGGTAGTTTCACG
*o	AACTCAAAGGAATTGACGG
gamma	ACTGCTGCCTCCCGTAGGAG
c17	CCGCCATTGCAAAGACGAGAC
c28	GATTCTTACAGTCTGGCGTCAG
c34	CTGATACACCATCTGGGCTTCG
c36	GCGCAGGTTCGTTGTCTGG
c 52	AGCAAGCGAATGTGTCTG
c 62	CCCTGTAAGCAGCATCCTG
c 75	GGTTGCTCAGAGGCTGACG
c 78	TTGTGTGGGAAGGCTACCG
c 46	CGGGAGTGAAATCGATGGCTGG
c 32	CAGGCATGACAGTCACAACCG
c54-71	CACTGGCGCTTTAGAAGG
c54-113	CGGGTATTAAAGGTGAGCTAGCAG
c85R	CCGAATGAGCTTTTGGGGCG
C7R	CGTCTATTTCGTGCTCACAGTG
C90	CCGATGAAGGCATAACGTAAGG
C31	CACAACGGTAATGCGGGGGAATTTG
C31R	CACATGCAGATTGTCAGCGCTTG
C81R	AACGCTCCGGATGCCTCCTG
C81	GTGCGGGAAAGTAATATGTGCG
C59	GAACATGGCTTTGGTTGGTGG

C60	CCCCGGGTCAAATCAAATTG
C64	CTTGTTGACGCTGTGTAG
C56R	GCTGGGCATATAGGGGAGTTCCG
C5R	CGGCCAGAAGATTCGCAAGACAG
C27R	GGACCGCTATTTAGATGTCAGCACG
C111	ACTGTTAGCGCTGAGAAGGTATCG
C26R	CCCACGCCTCAGCAAATGTTAATGG
C40R	CGCTAAAAACCCGTCTCCGCCAGG
C13	GACAGCACATTAGTCGCGGG
C42	AAGATCCACGCAAGGCCG
C39R	GTCAGTAGGGTGCTCAAGTTCCGG
C47	GAATATGACCCGCAATCCCACAG
C14	GAGATTTATAAACGTCGGGTCGCCAG
C14R	GTGTGTTCCCCGGTGTTGGTAAGG
C69	GCATACGTTGTTAGAGGATGCCG
C30	GCAAGGAGGCGGATGATTCAGTCGG
C64R	GGTTTGACCGGATATCGAGTG
C43	GCCGCTGGATTACTGGACG
C10	CTGCCGGGTAGATCGTTGTGAG
C10R	GGAAGTTGGGTCTGACAATG
C68	GGGCGTTCACATAGTCTTTCATGG
C44	GTACGTTGCAAACCTATCCCTG
C44R	GCCACGTTCCGTTTCCTTCCG
C17R	GCATGGGCTATTTCTCGTCTGACCG
C6R	CAATGGATGTCGCATACAGG
C18R	AGTTTGTGCGTAAGAGCTTCGG
C18	GAAGGCCTGCATAGCGATGAG
C63	CCATCGTCATGTCGAGCTGTTTG
C29	GATATGGCGGCCTCTTATATGG
C33	CCTGAACAGCGGTGGAGAG
C61R	TTGAGCAGTGGCGCATCGTG
C61	TGAGGTACTTCCGTATCCG

C67	TATCGCCTTCTCACCCAGGAAG
C57R	ATCGAAATGGGAGCGAGGTG
C57	ATGAGCCTCTCGTAATGGG
C2R	GCCGATTGCTCATCAACTCATG
C2	CCCTATAGTAGCGCCCCGTTG
C49R	ACGTAAACCCTTACTGGAGG
C46	GCGTTTCTCTGTACCTCCAAAG
C42N	GAGGCTCGCCATGCTCAGG
C85	GCGTTACATCAAGCTTGCGTG
*3	AGTCCCGCAACGAGCGCAAC
*pD	CAGCAGCCGCGGTAATAC
23sF	GTTCCCCCGGTTCGCCTCATAG
23sR	GAAACTTGCCCCGAGATGAG
C93	GACCGCCAGTTATCACAGTAGCCTG
C34N	GCGTGTCAAGGTTCATTGCATG

Table 3.2: Characteristic summary of contig boundary break point

Type of contig boundary	**Number of contig boundaries**
High coverage regions	84
Low coverage regions	66

Table 3.3: Primer combinations for gap PCR

PCR #	Contigs A	Contig A boundary	Contigs B	Contig B boundary	Primer A	Primer B	Additional contigs within gaps region
1	11	3' end	70	3' end	005-062F	005-062R	
2	110	3' end	75	3' end	297-015F	297-015R	
3	48	5' end	10	5' end	c10R	c68	
4	68	5' end	48	3' end	c10R	c68	
5	111	5' end	81	5' end	c111	c81R	contig 107
6	14	5' end	47	5' end	c14	c47	
7	64	3' end	30	5' end	c14R	c69	
8	69	3' end	14	3' end	c14R	c47	
9	113	5' end	15	3' end	c15-113	23sR	contig 105 and 97
10	63	3' end	18	3' end	c18	c63	
11	113	5' end	26	3' end	c26-113	23sR	contig 117
12	26	5' end	66	5' end	c26R	c40R	
13	66	3' end	40	5' end	c26R	c40R	
14	113	5' end	27	3' end	c27-113	23sR	contig 97
15	28	5' end	36	5' end	c28	c36	
16	71	3' end	29	3' end	c29	gamma	contig 107
17	113	5' end	31	3' end	c31	23sR	
18	46	5' end	32	3' end	c32	c46	
19	61	5' end	33	3' end	c33	c61R	
20	113	5' end	4	5' end	c4-113	23sR	contig 97
21	52	5' end	46	3' end	c46	015-062R	
22	49	5' end	78	3' end	c49	clpV1F	
23	113	5' end	54	3' end	c54-113	23sR	contig 117
24	71	3' end	53	5' end	c54-71	gamma	
25	113	5' end	56	5' end	c56-113	23sR	contig 105 and 98
26	62	3' end	110	5' end	c56-113	rev330	
27	71	3' end	5	5' end	c5-71	gamma	contig 106
28	57	5' end	2	3' end	c57R	c2	contig 87
29	64	5' end	60	3' end	c60	c64	contig 65
30	13	3' end	61	3' end	c61	c13	contig 72, 93 and 84
31	78	5' end	62	5' end	c62	c78	contig 77 and 1
32	71	3' end	6	5' end	c6-71	gamma	contig 91
33	6	3' end	18	5' end	c6R	c18R	
34	20	5' end	75	5' end	c75	R326R	contig 86,35,89, 98, 37,95,38

35	7	3' end	96	3' end	c7R	c90	
36	81	3' end	5	3' end	c81	c5R	contig 82, 8 and 50
37	59	3' end	114	5' end	c81R	c59	contig 107
38	71	3' end	85	3' end	c85-71	gamma	contig 91
39	31	5' end	3	3' end	c85R	c31R	
40	96	5' end	90	5' end	c90	c7R	
41	67	3' end	24	5' end	contig033-003F	contig033-003R	
42	41	5' end	109	5' end	contig093	contig002R	
43	109	3' end	2	5' end	contig093	contig002R	
44	45	5' end	12	3' end	contig344-011F	contig344-011R	
45	10	3' end	45	3' end	F347	c10	
46	22	3' end	21	5' end	F46	274	
47	71	3' end	63	5' end	N63-71	gamma	contig 107
48	71	3' end	7	5' end	R242R	gamma	contig 91
49	9	5' end	115	3' end	σ*	N63-71	

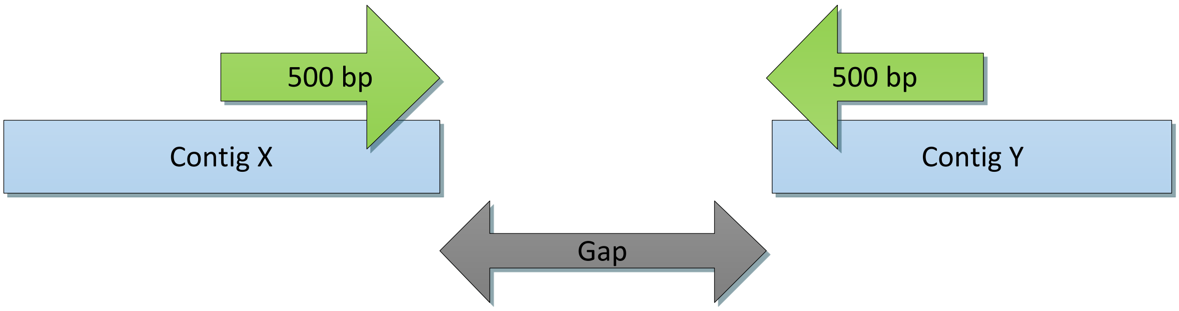
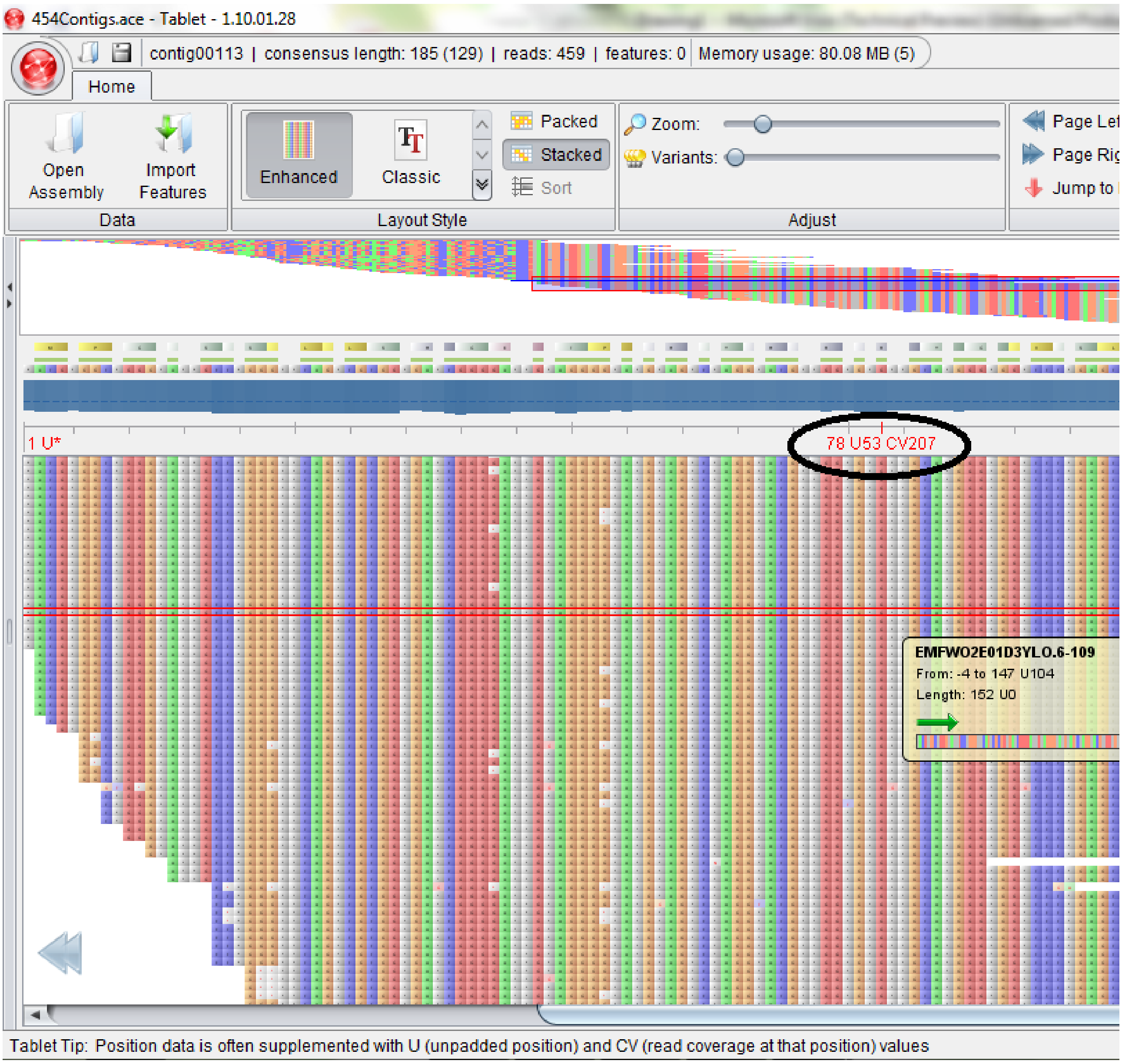


Figure 3.1: An example of selection regions for primer design. Primer design illustration of selection regions (green arrows), which is 500 bp from the edge of the adjacent contigs X and Y.



454Contigs.ace - Tablet - 1.10.01.28

contig00113 | consensus length: 185 (129) | reads: 459 | features: 0 | Memory usage: 80.08 MB (5)

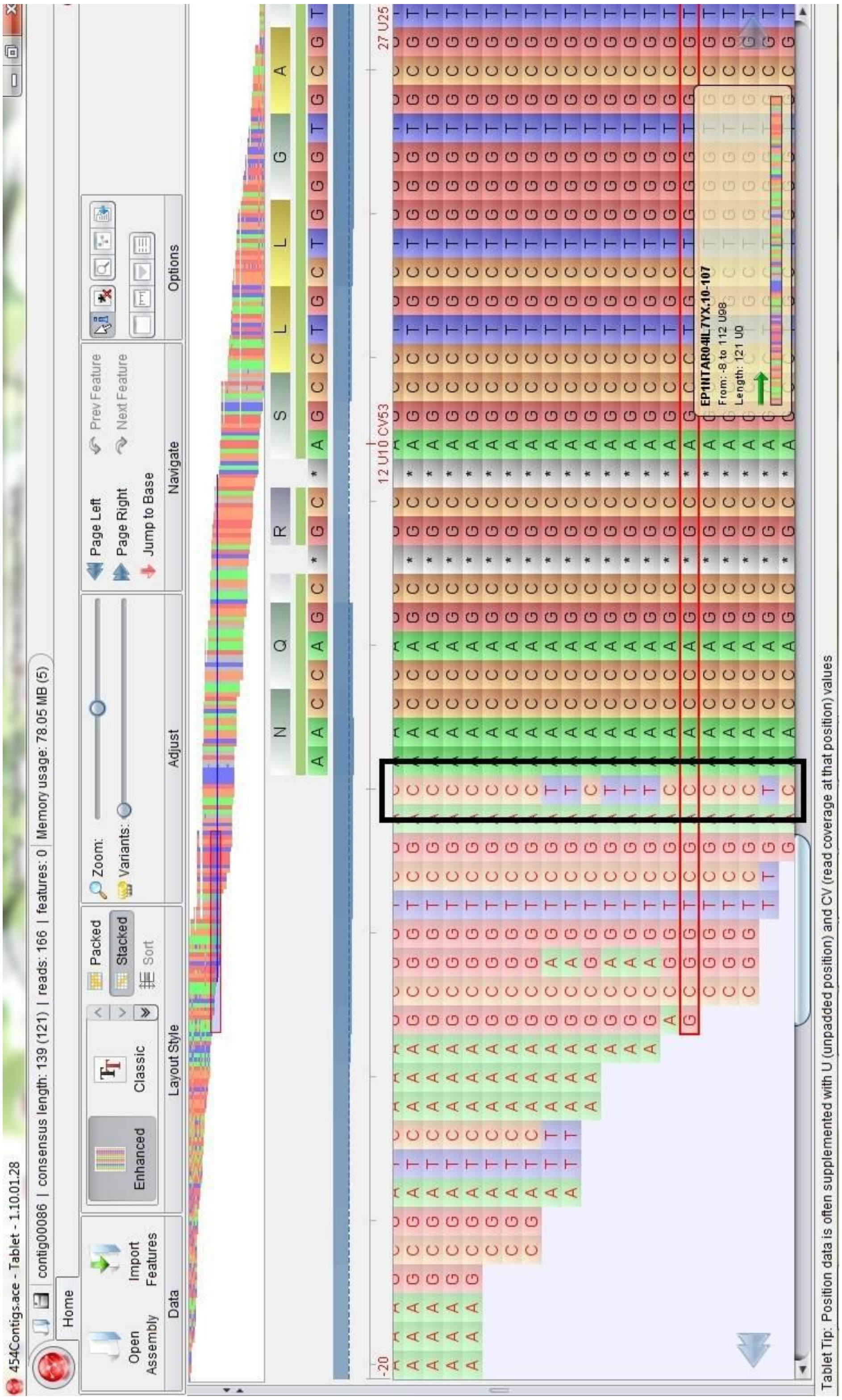
1 U*

78 U53 CV207

EMFWO2E01D3YLO.6-109
From: -4 to 147 U104
Length: 152 U0

Tablet Tip: Position data is often supplemented with U (unpadded position) and CV (read coverage at that position) values

Figure 3.2: An example of a high coverage region (the coverage is as high as 207 in some areas of the 5S ribosomal DNA region).



454Contigs.ace - Tablet - 1.10.01.28

contig00086 | consensus length: 139 (121) | reads: 166 | features: 0 | Memory usage: 78.05 MB (5)

EP1HTAR04IL7YX.10-107
From: -8 to 112 U98
Length: 121 U0

Tablet Tip: Position data is often supplemented with U (unpadded position) and CV (read coverage at that position) values

Figure 3.3: An example of a SNP at the edge of a contig.

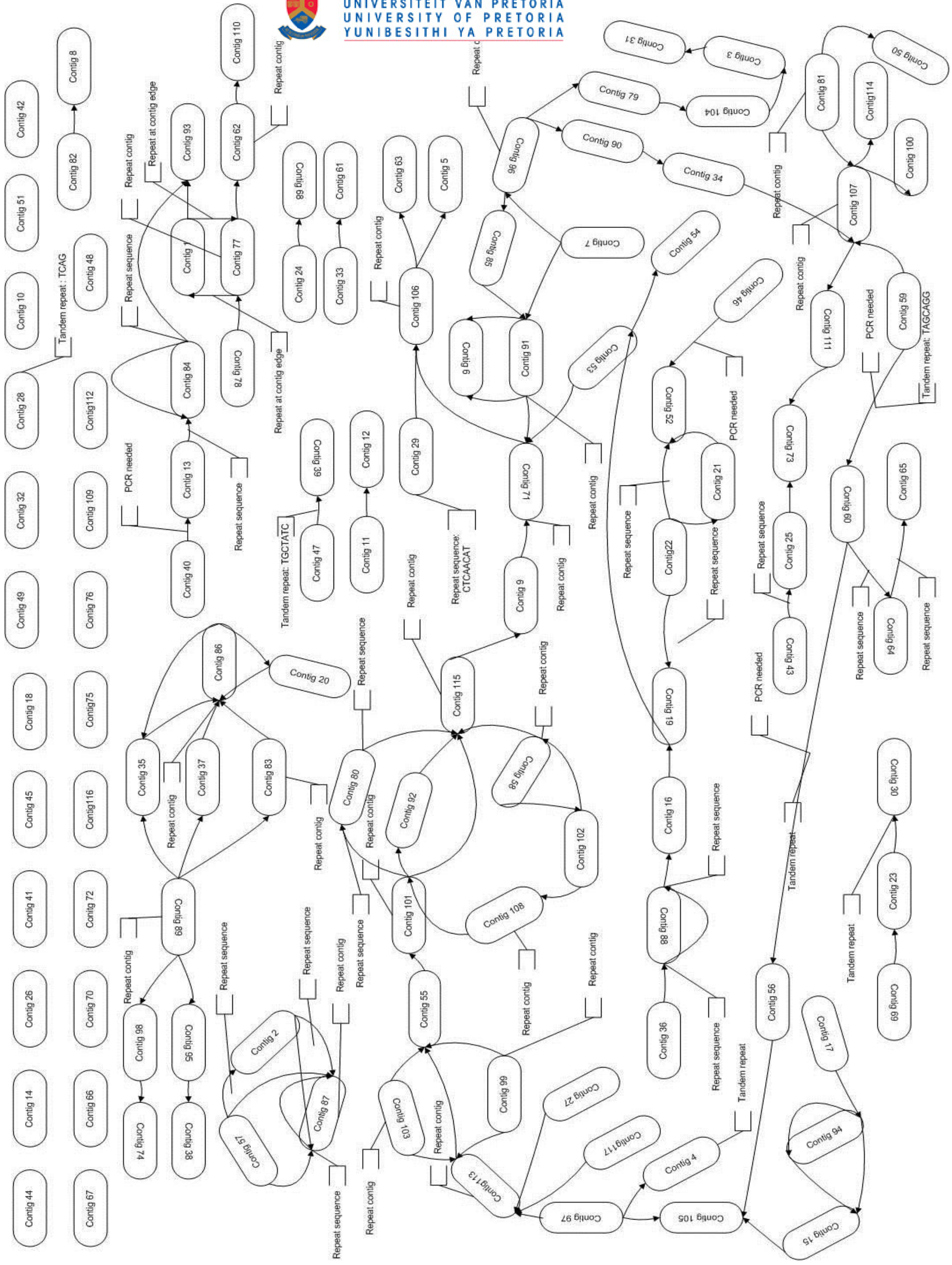


Figure 3.4: Contig graph of *Pantoea ananatis* LMG 20103 draft assembly. All the possible contig orientations are represented, where each node represents an individual contig and the connecting lines/edges represents 454 reads that can bridge between two contigs. Additional information was provided.

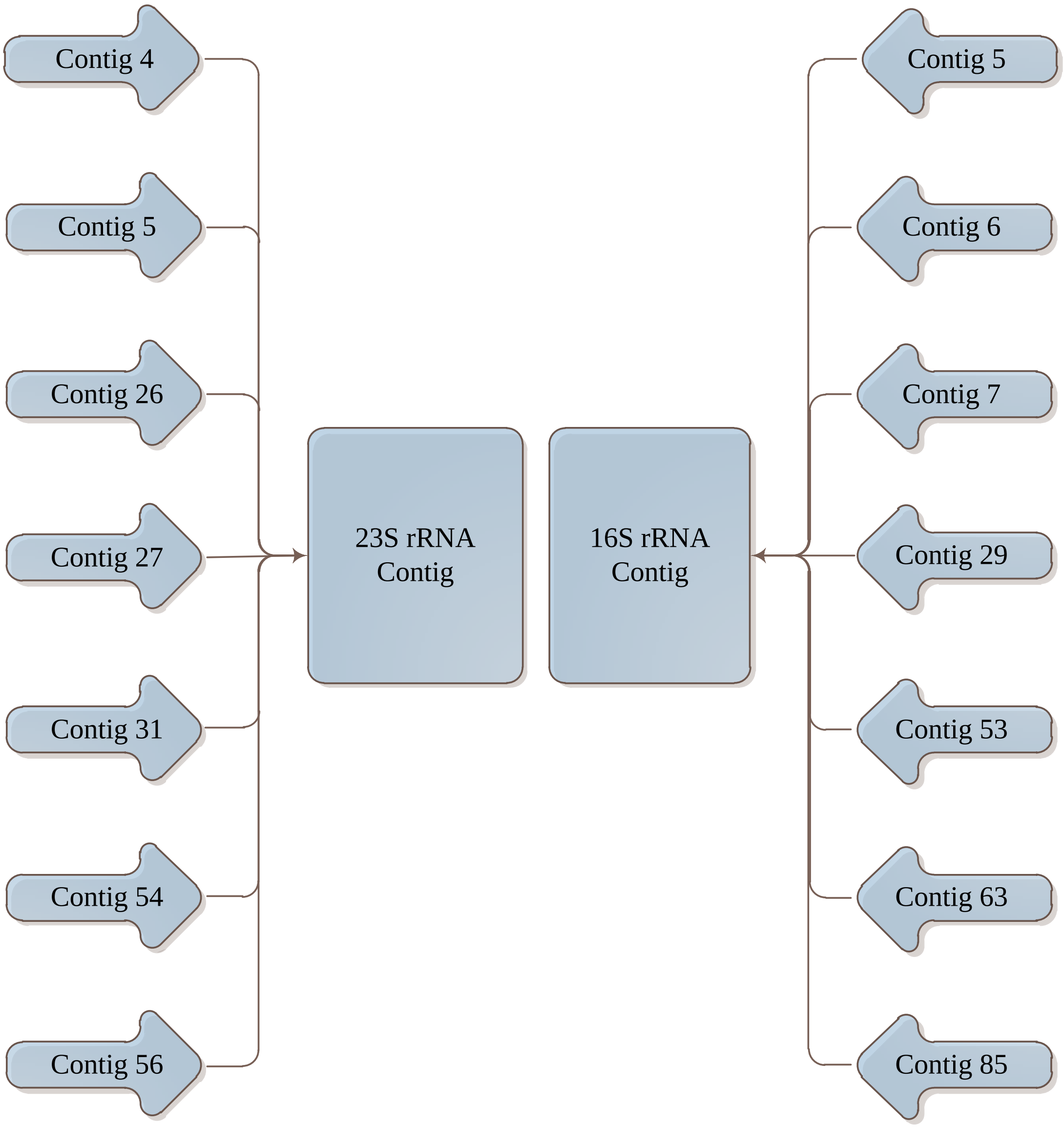


Figure 3.5: The seven branch points of the 23S and 16S rDNA.

Chapter 4

Genomic metabolic pathway reconstruction of *Pantoea ananatis* LMG 20103

Genomic metabolic pathway reconstruction of *Pantoea ananatis* LMG 20103

4.1. Introduction

Understanding the metabolic capabilities of an organism has always been an element in unravelling the biology of an organism. For this reason various tools have been created to assist in recreating the metabolic network of bacteria. In 1995, EcoCyc (http://ecocyc.org/), became the first metabolic profile database to become publicly available, which combined phenotypic and biochemical data. This database consists of a collection of metabolic and transcriptomic reactions, transporter reactions, transcriptional regulation and signalling pathways data of the bacterium *Escherichia coli* K-12 MG1655, and is still under constant study and review (Keseler *et al*, 2009 and 2011). Currently, this database has described 277 pathways, 1922 reactions, 254 transporters and 3422 transcription regulation units. With its advanced and well-studied information, EcoCyc has become the referencing network for numerous metabolic network reconstruction projects for biologists. The number of re-constructed networks for a variety of organisms is expanding rapidly, for example, BsubCyc for *Bacillus subtilis* (http://bsubcyc.org), MouseCyc for the laboratory mouse (http://mousecyc.jax.org) and HumanCyc for *Homo sapiens* (http://humancyc.org) (Evsikov *et al*., 2009; Oh *et al*., 2007; Romero *et al*., 2004; Thiele and Palsson, 2010) have been created.

A reconstructed metabolic network often serves as the foundation for further in depth studies related to the target organism. According to Durot (2009), numerous reconstructed network based models were described, including models to predict growth behaviour. These networks are also used for the prediction of unknown phenotypic expression of the organism, or the formulation of better hypotheses, which could reduce experimental time and effort (Imielinski and Belta, 2008; Risso *et al*., 2008). Apart from these typical examples, a metabolic model can also provide possible explanations or interpretations of experimental results such as metabolic flux measurements, metabolite concentrations or differences in gene expression when the organism is under different stresses or growth stages (Durot *et al*., 2009; Hanisch *et al*., 2002; Risso *et al*., 2008; Zamboni *et al*., 2008). Depending on the quality of the re-construction model, these predictions are also used during metabolic engineering experiments involving the manipulation of the metabolic system to increase or decrease an organism's performance or to optimise a particular metabolic pathway for the synthesis of a certain type of biomass or product (Durot *et al*., 2009). The network predictions are also used to determine the putative interactions between the introduced drug and the infectious organism during drug design studies (Karp *et al*., 2011). The comparisons between the different organisms' metabolic networks have also shown evidence of potential evolutionary pattern/markers to be used for taxonomic purposes (Barona-Gómez *et al*., 2011).

Re-constructing the basic part of the pathway network is easy, as most of the canonical pathways are well described (Thiele and Palsson, 2010). This is typically achieved through a comparative approach using the available databases to identify candidate genes. However, the construction of a unique organism's specific pathways and understanding the different interactions between the pathways as a network are more of a challenge (Ates *et al*., 2011). As the complexity of all the interactions within the whole network increases, other factors need to be taken into consideration. This includes issues such as whether or not a substrate can be imported into or exported out of the cell and the accessibility of additional requirements for the different reactions to take place such as the presence of co-factors (Durot *et al*., 2009; Feist *et al*., 2008; Henry *et al*., 2010). These reactions are also under strictly controlled regulation, and the different regulators and signals for activation or suppression can also impact on further network studies (Feist *et al*., 2008).

The re-construction of the metabolic pathways network of an organism has never been a simple task. Without any universal protocol for the genomic-scale metabolic pathways re-construction, different approaches to resolve particular problems have been developed. These approaches generally share a number of common methodologies, where both bioinformatic and experimental data are merged to evaluate the predicted metabolic network of the organism. According to Thiele and Palsson (2010), the entire process of metabolic pathway reconstruction can be divided into four stages. The first stage is to compose the draft reconstruction of the pathway network. This is followed by a series of refinements of the reconstruction data. Once the refinement is completed, the conversion of the data into a mathematical model or computable format, for example, Flux balance analysis (FVA) can be done for cell network validation (Kauffman *et al*., 2003). Finally the reconstructed pathway data can be put into a format that will allow the dissemination of a complete, comprehensive metabolic pathway for the specific organism being studied.

With the advancement in genome sequencing, detection assays, and the developments in bioinformatics pipelines, genomic-scale metabolic pathway re-construction of a targeted bacterium has become less complex. According to Thiele and Palssson (2010) it takes on average 1 to 7 days for the re-construction of a draft pathway network. At present the complete genomic sequence of the target organism is the basic requirement for network re-construction. The well-known automated pipelines are Pathway Tools, metaSHARK and Model SEED (Aziz *et al*., 2008; Henry *et al*., 2010; Karp *et al*., 2002 and 2010; Paley and Karp, 2006; Pinney *et al*., 2005).

During the initial stage, the draft metabolic pathways that are created rely heavily on the genome annotation, where the candidate genes/proteins with potential metabolic function are already identified and assembled. Expansion of the data set is done through extensive searches in publicly available metabolic pathway database e.g. BioCyc (http://biocyc.org), Kyoto Encyclopedia of Genes and Genomes: KEGG (http://www.genome.jp/kegg), enzyme reaction databases, e.g. BRENDA (http://www.brenda-enzymes.org) and the literature related to the target organism (Chang *et al*., 2009; Durot *et al*., 2009; Kanehisa *et al*., 2006; Karp *et al*., 2005 and 2011; Ogata *et al*., 1999; Scheer *et al*., 2010; Schomburg *et al*., 2002 and 2004; Thiele and Palsson, 2010). Based on the collected information, the bioinformatics pipelines such as Pathway Tools and Model SEED (http://seed-viewer.theseed.org) are able to reconstruct a draft pathway network. These systems use the "co-occurrence relationships" approach, where the metabolicly related reactions are grouped together according to their particular functions (Aziz *et al*., 2008; DeJongh *et al*., 2007; Durot *et al*., 2009; Henry *et al*., 2010; Karp *et al*., 2002 and 2010; Paley and Karp, 2006).

The automated pipelines can save both time and effort on the draft network re-construction, but the collected data used is often organism non-specific. This is mainly due to the fact that the databases used are either collections from other organism specific experimental results or generated through computational analyses based on homology and similarity between gene/protein sequence or protein structure domains. Very often these datasets contain false or incorrect data with regards to the targeted organism, which can lead to false reactions and assumptions in the final pathway network (Thiele and Palsson, 2010).

The initial pathway reconstruction can also show inconsistencies due to new and unfamiliar gene annotations, incomplete genomic sequences or reactions to which a partial EC number has been assigned. Reactions, for which the EC number is absent, will also not be detected by the bioinformatics pipelines during the automated re-construction. These issues will result in missing reactions and incomplete metabolic pathways, which will require additional manual inspection and experimental data to verify the network (Durot *et al*., 2009; Green and Karp, 2005; Thiele and Palsson, 2010). As an example using the "gene-reaction association" approach of PathoLogic, annotated genes within the genome are assigned to metabolic reactions before assembly of the draft network. A poorly annotated or incomplete genome will, therefore, result in numerous missing reactions (Karp *et al*., 2002; Reed *et al*., 2006). To overcome this problem, some reconstruction pipelines such as Model SEED and metaSHARK provide the extra function of an automated gene annotation as part of their package (Henry *et al*., 2010; Pinney *et al*., 2005).

A high quality re-constructed network is required as the inaccurate predictions can have an effect on the design of studies to investigate the biology of the organism. The current automated draft networks still require manual refinement and curation to improve the quality of pathway prediction and data refinement steps are, therefore, still an integral part of any pathway re-construction process. This often involves the manual inspection of each metabolic reaction in terms of their functions, stoichiometry, direction and localisation. Information such as gene-protein-reaction (GPR) associations, intracellular transport, biomass composition, growth medium requirements, growth-associated ATP maintenance reaction (GAM) or demand reactions for mass-balance of the cell can also be used (Durot *et al*., 2009; Feist *et* *al*., 2008; Thiele and Palsson, 2010). Even though manual inspection can ideally remove most of the false positive pathways within the predicted network it will not necessarily be a 100% accurate and experimental data remains vital in both the refinement step of the re-construction and the verification of unknown reactions or pathways.

The objective of this part of the study was to reconstruct a metabolic pathway for *Pantoea ananatis* LMG 20103 in order to expand the knowledge of the metabolic capabilities of this bacterium and to verify and correct the genome annotation currently available. The completed genome sequence was submitted to two different network reconstruction pipelines in order to compare their predictions. In addition, detection of mis-annotated genes within the genome was done by curating the component interactions of the network and by identification of missing reactions which formed part of the essential pathways. Finally to augment the missing reactions within the predicted network, alternative genes were identified based on homology or verified using phenotypic data.

4.2. Methods and Materials

4.2.1. Reconstruction of draft metabolic pathways

The genome sequence of *Pantoea ananatis* LMG 20103 was completely assembled and annotated (De Maayer *et al*., 2010) as discussed in the previous chapter. The initial annotation was performed using the web based automated annotation pipeline BASys (Guindon *et al*., 2004). This annotation was then manually curated, based on further information obtained by submitting the genome to a number of other databases and pipelines. This genome had 4295 annotated genes and formed the basis for the reconstruction of the bacterium's metabolic pathway.

For the reconstruction using Pathway Tools the genome was first submitted to the Microbial Genome Annotation (MaGe) (http://www.genoscope.cns.fr/agc/mage) pipeline (Vallenet *et al*., 2009). From the MaGe platform, the *P. ananatis* genome's data was passed onto the metabolic network reconstruction program: Pathologic, as part of the Pathway tools package. The generated Pathway Genome Database (PGDB) *of P. ananatis* LMG 20103 was stored and made available for download from the MicroCyc database (http://www.genoscope.cns.fr/agc/microcyc) (Karp *et* *al*., 2002). Pathway Tools version 14.0 software with the Tiers 1 + 2 was downloaded from the Bioinformatics Research Group at SRI international with an academic-use licence and installed for Desktop Mode usage. The downloaded PGDB .zip file of *P. ananatis* LMG 20103 was extracted and placed into the PGDBs/biocyc folder in Pathway Tools program directory. The pathway diagrams and pathway information were visualised and studied with the Desktop Mode of Pathway Tools, where the inspection of the individual pathways and reactions could be performed.

For the second draft network reconstruction the Model SEED programme was used. The complete un-annotated genome sequence was first submitted to the RAST annotation server (http://rast.nmpdr.org) as a single fasta file. The annotation of the genome was done automatically using the FIGFam database. The genome was then run through a series of programs for automated gene predictions and gene annotation. The genome annotation and reconstructed pathways that were generated were then verified using the SEED Viewer version 2.0 interface (http://seed-viewer.theseed.org) or the Model SEED Viewer version 1.0 (http://seed-viewer.theseed.org/seedviewer.cgi?page=ModelView) (Aziz *et al*., 2008; DeJongh *et al*., 2007; Henry *et al*., 2010).

4.2.2. Comparison of the pathway reconstructed models

A comparison between the metabolic pathways of *P. ananatis* LMG 20103 as constructed by Model SEED and PGDB was performed. Since the RAST server had re-annotated the genome, the automated RAST annotations were also matched to the initially manually curated genome annotation by matching the start or end gene position and orientation of the genes between the two sets of annotations using a programme written for this purpose in python. This was followed by extracting the detail of all the genes associated with the pathway. In the case of the Model SEED predictions, pathway details such as the subsystem and scenarios involved, the reactions it carried, the substrates and products for the reactions, the EC numbers and the gene names were listed. With the PGDB model, the metabolic pathway name, the EC number and its corresponding gene name as well as the reaction and substrate information were listed for each gene. The manual comparison between the two pathway models was done based on the gene annotation and the EC number assigned by the two pipelines as well as information on reactions, substrate and products and the involvement of genes in specific metabolic pathways.

4.2.3. Refinement of reconstructed metabolic pathways:

4.2.3.1. Identification of missing pathway reactions and their candidate genes

With the Pathway Tools-v14.0 program, the missing reactions or pathway report was generated for all the missing reactions predicted for the *P. ananatis* LMG 20103 genome (Green and Karp, 2005). This report consisted of the name of all the missing reactions as well as the corresponding EC number, chemical reactions, and their associated pathways. This information was then used to search for the possible candidates to fill the missing reactions from the data created by the Model SEED pathway model. Genes that could potentially carry out the missing reactions were located and verified based on the reaction's substrate, reactants, co-factors and EC number.

4.2.3.2. Verification of the metabolic pathway with phenotypic data

Another step taken to refine the metabolic pathways was to check the transport mechanism in order to determine the types of substrate that the organism can take up and utilise. For this purpose, the transporter genes were extracted from the Model SEED database and their reaction and type of substrates were recorded. The utilisation of these putative substrates by *P. ananatis* was determined by searching for their corresponding converting reactions or pathway involvement.

Other, more organism specific experimental data was also incorporated into the prediction. Phenotypic characteristics of *Pantoea* sp. were obtained from a report by Brady (2008). Characteristics of *P. ananatis* LMG 20103 had been determined using API 20E, BIOLOG, API 50CH and Biotype 100. Approximately 194 substrates were tested for any evidence of transport and metabolic reactions. Compounds that showed positive results were recorded and the existence of their corresponding pathways were verified in the proposed network.

4.3. Results

4.3.1. Reconstruction of draft metabolic pathways

According to the pathway genome database (PGDB) summary obtained after the genome of *Pantoea ananatis* LMG 20103 was submitted to MicroScope (MaGe) for metabolic pathways re-construction, the draft metabolic reconstructed process was based on a total of 4501 genes from the genome, which consisted of 4385 protein coding and 116 RNA genes. The gene-reaction match algorithm identified 1277 genes as putative candidates for possible involvement in metabolic process and 60 genes were identified as transporter proteins. Three hundred and twenty-two pathways were predicted, within which 1541 enzymatic reactions and 23 transport reactions were recognised (Table 4.1). A total of 1160 different biochemical compounds were predicted as metabolites for the metabolic network model.

The second metabolic network was re-constructed using the Model SEED pipeline. The genomes of 30 other bacteria were determined to be the closest neighbours of *P. ananatis* LMG 20103 by the RAST server (Aziz *et al*., 2008; Overbeek *et al*., 2005). According to the automated computational annotation, 4454 genes and 92 RNAs were identified, using the FIGFam database. Within these annotated genes, it was predicted that a total of 496 subsystems with 3565 features or reactions and 361 scenarios were recognised for the *P. ananatis* network model (Figure 4.1 and Table 4.2).

According to the Subsystem coverage analysis shown in Figure 1, 2448 (55%) of the predicted genes of the completed genome sequence were assigned to a specific subsystem. This included 2318 known genes and the rest were annotated as hypothetical proteins (130 genes). The other 2006 (45%) predicted genes remained un-assigned to any functions, with almost half of them annotated as hypothetical proteins (925 genes). Only 26 subsystem categories were identified, with the carbohydrate, amino acids and derivatives, protein metabolic, cell wall and capsule, co-factors, vitamins, prosthetic groups, pigments and stress response features forming the major pathway contributors to the predicted network.

4.3.2. Comparison of the pathway reconstructed models

Two different approaches for the classification of reactions and pathways were used by the two reconstruction pipelines: Pathway Tools used a system of classes and subclasses to group the pathways and Model SEED used subsystems and scenarios to group the pathway reactions. In addition, the SEED pipeline used the KEGG network representation of pathways, and Pathway Tools uses the curated diagrams from its own database. The resulting differences in reaction groupings and metabolic pathway diagrams make comparisons between the two predicted models difficult. Using the export table function in SEED, information including the gene name, start and stop positions, the functions and the subsystem involvement were extracted and exported to a Microsoft Excel table for further analysis. With the Pathway Tools model, this information was extracted from the DAT files of the PGDB. The information included the classes.dat, compounds.dat, Enzrxns.dat, genes.dat, pathways.dat, proteins.dat, reactions.dat and substrate.dat files. Once the two sets of annotations were matched based on the positions of genes within the genome, the comparison between the collected information from both PGDB and SEED were performed, and the missing reactions in the Model SEED predictions could be recorded during the process.

During the comparison of the two predicted models, slight variations between the two databases were detected. In some cases the protein name and EC number for a particular reaction appeared to differ, even though the same gene was recognised for that reaction and the final predictions were not affected. In other cases, due to reaction variations, the conversion of some of the substrates that involved a single step reaction in one model was substituted for multiple reaction steps in the other model. It was also noted that a missing reaction might occur in one model but not in the other one, but in some instances the particular reactions were missing from both of the predicted models.

4.3.3.1. Identification of missing pathway reactions and their candidate genes

According to the Pathway gap report, 263 reactions were reported as missing from the PGDB model. Using the EC numbers and the reactions equations of these missing reactions, candidate genes associated with these missing reactions were identified from the Model SEED network. A total of 53 genes were found to fill the missing reactions in the PGDB model (Table 4.3). On further inspection to find the cause of these reconstruction faults, a number of problems were identified. These included incorrect open reading frame predictions, where the required gene was split and incorrectly annotated into a number of individual genes (PANA_4114, PANA_4115 and PANA_4116); and genes indicated as “Hypothetical proteins” as these genes were not recognised by the reconstruction pipeline. In one instance 6 hypothetical genes (PANA_0880, PANA_0881, PANA_0883, PANA_1278, PANA_1282 and PANA_1284) were identified which completed one of the essential pathways. Other problems were mis-annotated genes where 8 specific examples were identified and, lastly, cases where the genes were correctly annotated but not recognised by the reconstruction pipeline.

The comparison resulted in the filling of a number of the missing reactions including 12 metabolic pathways and 9 scenarios that could be completed in the PGDB and the Model SEED network models, respectively. Not all the missing reactions could be completed, as indicated in Table 4.4. Forty three pathways are still near to completion with less than 2 reactions steps missing from the complete pathway. The remaining 31 pathways are considered to be potential false positive pathway predictions as little evidence exists to support their presence. In many cases only 1 or 2 reactions are present for the whole pathway.

4.3.3.2. Summary of pathway network data

4.3.3.2.1. Summary of essential pathways

After the comparison of the two networks and the completion of a number of missing reactions, a total of 257 metabolic pathways were found to be present in *P. ananatis* LMG 20103 (Table 4.5) (this includes the different variations of pathways and super-pathways). All these pathways are considered to be complete and identical between the two network predictions. According to Table 4.5, the nucleotide and nucleoside pathways are complete and the bacterium is capable of *de novo* biosynthesis, salvaging and recycling of both purine and pyrimidine nucleotides and their derivatives. Based on the genomic-scale metabolic pathway network, *P. ananatis* LMG 20103 is also capable of synthesizing all 20 essential and non-essential amino acids. Most of the amino acids have more than one co-existing biosynthesis pathways within the network. The catabolic pathways for the amino acids were also present in the model. The completed carbohydrate degradation pathways for the degradation of D-mannose, D-glycerate, Fructose, Glucose, Glycogen, L-arabinose, L-lyxose, Lactose, Melibiose, Rhamnose, Ribose, Sucrose, Trehalose and Xylose were found. Only complex carbohydrate compounds such as ADP-L-glycero-beta-D-manno-heptose, Cellulose, CMP-KDO, dTDP-L-rhamnose, Glycogen, Sorbitol, Trehalose and UDP-D-xylose could be synthesised according to the pathway model.

4.3.3.2.2. The transport mechanisms and membrane Transporters of *Pantoea ananatis*

The import and export of compounds by the cell can provide some clues as to what type of reactions or pathways *P. ananatis* is capable of performing. According to the data obtained from the re-constructed pathway predictions, most of the identified compounds are transported via ABC transporters, antiporters, symporters, ion-coupled transporters, permeases, specific transporters, Phosphoenolpyruvate (PEP) group translocator or will simply diffused across the cell membrane. Some of the identified transporters are able to transport a particular class of substrates and a range of substrates can, therefore, be associated with these transporters. Approximately 150 compounds were suggested to be transported across the cell membrane or between the different compartments within the cell according to the Model SEED prediction of transport proteins. These compounds include different nucleotides, monosaccharides, oligosaccharides, polyols, amino acids, peptides, co-factors, minerals and organic ions. Only 96 of the 150 compounds were, however, found to feature as substrates for the reactions predicted in the metabolic network model.

4.3.3.2.3. Verification of the metabolic pathway with phenotypic data

The phenotypic characteristics of *P. ananatis* LMG 20103 (Brady, 2008) showed that of the 196 different compounds tested only 94 of these substrates could be utilized by this strain. Another 12 substrates showed uncertain reactions. Most of these substrates will require carbohydrate or the secondary metabolite transport systems to enter the cell. When comparing this list with the list of recognised transporters, only 38 of the 94 compounds could be linked to available transporters in the network model (Table 4.6).

4.4. Discussion

The final quality of a re-constructed metabolic pathway is dependent on the actual curation of the pathways. Curation usually relies on current on-going projects which study metabolic networks and the availability of the newly verified data. The different pipelines might use different functions and approaches and the resulting reconstructed network might differ from those of other pipelines. Model SEED is, for example, orientated towards prokaryotic organisms, as its annotation is done based on data from the "Project to Annotate 1000 Genomes" (Overbeek *et al*., 2005). On the other hand, Pathway Tools has data based on a larger range of organisms, including prokaryotic and other more complex eukaryotic organisms, available for network reconstruction. Due to the differences between the approaches of the two pipelines, a comparison study of the two predicted networks was undertaken.

The number of complete pathways in the Pathway Tools and Model SEED network predictions were 75% and 64%, respectively (Table 4.1 and 4.2). Ideally, similar predictions for the two pipelines would be expected as they both used the same initial genome sequence data for the network re-construction. Automated re-constructions are, however, very generic in their approach and false positive pathways and reactions have always been a major contributor to the incomplete pathways of the predicted network (Paley and Karp, 2002). To reduce the amount of false positive predictions, Pathway Tools is designed to recognise the candidate pathways and then to filter out the false positive pathways based on the assumption that a pathway will only be present if most of the genes associated with the pathway were detected (Paley and Karp, 2002). Using this approach, a subset of false positive pathways was removed from the Pathway Tools prediction, while other incomplete pathways with supporting evidence of the presence of unique enzymes remained as part of the predicted network. Due to this approach the number of false positive pathways were less when compared to the Model SEED prediction.

One of the common goals of network comparison studies is to identify possible candidate genes to complete some of the incomplete pathways. This is done by comparing the draft network of one organism to the network of a closely related organism that has a well curated annotated genome (Osterman and Overbeek, 2003). During this process, mis-annotated genes and new undiscovered pathways can often be recognised. In the present study a slight variation of this approach was applied as two different network predictions based on the same genome sequence were used. This approach was ideally suited to pinpoint candidate genes and investigate the main causes for missing genes in the predicted network.

As most of the recognised candidate genes were correctly annotated (Table 4.1), one of the limitations of the Pathway Tools pipeline may be the use of the "linking enzyme to reactions" method (Paley and Karp, 2002). In Pathway Tools, the network prediction is done by using a name-matching algorithm that links the required proteins to pathways based on the annotation provided by the input Genbank file. The inability of the programme to recognise the qualifier name/gene name from the input Genbank file results in missing reactions and incomplete pathways (Paley and Karp, 2002). Therefore, the cause of missing reactions in the Pathway Tools prediction for *P. ananatis* LMG 20103 was mainly due to the use of new or uncommon gene names during annotation.

The inconsistent use of terminology is a major problem in biological research. This is not an exception when it comes to pathway re-construction. Synonyms of gene and protein names and abbreviation of the biochemical name, and metabolites in the network can result in conflicts and human errors during curation and comparison of the networks. Most of the pipelines are using databases that can detect a range of synonyms for the different gene/protein names and metabolites (Caspi *et al*., 2006; Green and Karp., 2005; Ogata *et al*., 1999). However, it is still dependant on the regular update of these databases, since the data generated from the genomic research is growing exponentially (Feist *et al*., 2008; Karp et al., 2010; Reed *et al*., 2006; Vieira *et al*., 2011).

Mis-annotated and the identification of proteins as hypothetical in the input file are also problematic in the re-construction steps, as they are not recognised by the Pathway Tools pipeline. Annotating a gene as a hypothetical protein is often due to low alignment scores to existing proteins in the publicly available reference databases. Reed (2006) and Ates (2011) also suggested that genes with mis-assigned gene name and functions can lead to gaps or missing reactions in the network. It is, however, very difficult to address this problem because without any experimental data the correct annotation and actual function of a gene are difficult to determine (Thiele and Palsson, 2010). Incorrect open reading frame predictions have also been one of the sources for mis-annotation and the identification of proteins as hypothetical. Hypothetical proteins are also problematic in Model SEED. From the data presented in Figure 1, it can be observed that within the incomplete scenarios in the Model SEED prediction 50% are made-up of hypothetical proteins. This data further suggests that these incomplete pathways might not be false positive pathways but rather undiscovered pathways or annotation errors (Ates *et al*., 2011; Feist *et al*., 2008).

Although it is believed that the *P. ananatis* LMG 20103 genome was complete there might still be missing reactions and incomplete pathways if specific plasmid sequences were lacking. As the size of plasmid/mega plasmids increase and their contents also becomes more abundant, their involvement in the metabolic pathway network may be vital. In the case of *P. ananatis* LMG 20103 genome a 3.4 Mb sized plasmid was identified which carried part of the thiamine metabolism pathways and other secretion systems. It is, therefore, important to include the associated plasmid sequences during network reconstruction.

Through comparison of the two predicted pathway networks, missing reactions that were caused by computational error, annotation error or missing sequences were identified and corrected. However, the remaining missing reactions (Table 4.4) and incomplete pathways were unexpected, as both computational pipelines have automatic gap filling functions, but these missing reactions remained un-resolved in both models. It is believed that some of the missing reactions could be carried by analogous proteins for which experimental data will be required before it can be included into the network (Barona-Gómez *et al.*, 2011; Hiratsuka *et al.*, 2008; Sandoval-Calderon *et al.*, 2009).

One of the benefits of using Model SEED is that it provides the advantage of spotting missing sets of reactions, as the pathways are divided into subsystems and scenarios. The final prediction is based on completion of these individual subsystem and scenarios, before assessing whether the entire pathway is complete (Henry *et al*., 2010). The assessment in Pathway Tools is done using evidence scores and by judging the number of unique enzymes present. This becomes rather difficult with small pathways consisting of only a few reactions (Paley and Karp, 2002). Overall, with the poor understanding of the whole network, any assessment not based on experimental data is rather unreliable, as vital information might be neglected.

During metabolic pathway reconstruction, phenotypic data is used as indirect evidence for the verification of the presence of transporter and possible metabolic pathways, by determining whether the compounds are actually being transported and utilised by the organism (Thiele and Palsson, 2010). According to the phenotypic data collected (Table 4.6), most of the compounds utilized were carbohydrates and their corresponding transporter proteins were present in the network. Other forms of transportation of compounds, including diffusion or ion channels were also suggested from the network but there is no experimental data to support this finding.

According to Table 4.6, more than half of the substrates that are predicted to be transported into the cell were expected to be to be utilised as well. Only a small portion of these substrates were confirmed to be utilized based on the phenotypic data collected. This could be due to the fact that some of the transporters are non-specific and the prediction will include all the possible substrates even though it may not be transported. On the other hand the negative phenotypic results cannot serve as verification for false positive pathways and reactions. As the function of the cell changes depending on the external and internal environment, different pathways and reactions may be activated (Thiele and Palsson, 2010). According to Winn (2006), *Pantoea* sp. was described as an organism that is unable to utilise lysine, arginine and ornithine. However, according to the re-constructed model, various biosynthesis and degradation pathways for these amino acids are present, which further suggests that characterisation of an organism on phenotypic data may not be completely reliable.

For further curation of the pathway network, a well-constructed network model is required to conduct a flux balance analysis (FBA) (Thiele and Palsson, 2010). With the missing reactions in the current *P. ananatis* network model (Table 4.5) a flux balance analysis might be difficult. These missing reactions can cause undesirable results for example, dead-end metabolites, as particular compounds will accumulate inside the cell without further conversion, utilisation or act on transportation as based on the current predictions. Further experimental data is, therefore, required to close these gaps and locate the missing reactions.

4.5. Conclusions

Today, reconstructing a genomic metabolic pathways network is faster and less complicated than in the past, as automated reconstruction can be performed with bioinformatic pipelines which utilise both curated experimental and computational generated databases. The reconstructed networks may, however, still contain missing reactions, incomplete pathways and false positive pathways. The comparison between the two predicted metabolic network of *P. ananatis* LMG 20103 from Pathway Tools and Model SEED, helped to identify the cause of some of the re-construction problems. They were either caused by mis-annotation of the genome, missing sequences or algorithm errors in the re-construction pipeline. Through the identification of these errors the comparison assisted in the correction of mis-annotated genes and verification and improvement of the current genome annotation.

Refinement of the re-constructed network and filling in missing reactions are necessary for improvement of the quality of the network and to avoid problems during downstream use of the data, as it might result in imbalanced flux balance analyses and misleading future experimental design. During refinement of the network, genome sequence data, phenotypic data and biochemical data are required as evidence for the verification of pathways and filling the missing reactions by means of candidate genes. Experimental data is, however, lacking as some of these experiments might be difficult to conduct. For these reasons refinement of the network is still considered to be the most time-consuming and labour intensive step in the network reconstruction process.

Reconstruction of the metabolic network of *P. ananatis* LMG 20103 has provided a foundation for future experiments. From the metabolic network it can be concluded that *P. ananatis* is an independent organism that is capable of carrying-out most of the *de novo* biosynthesis reactions required for survival. As an opportunistic pathogenic organism, *P. ananatis* also has the ability to procure numerous nutrients via transporters. The co-existence of various alternative pathways for the utilisation of certain metabolites and the presence of multiple genes to catalyse the same reaction helps to explain the ubiquitous nature of the bacterium and its flexibility in adapting to the different niches it can occupy.

References

Ates, Ö., Oner, E.T., and Arga, K.Y., 2011. Genome-scale reconstruction of metabolic network for a halophilic extremophile, Chromohalobacter salexigens DSM 3043. BMC Systems Biology 5, 12.

Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O., 2008. The RAST Server: rapid annotations using subsystems technology. BMC genomics 9, 75.

Barona-Gómez, F., Cruz-Morales, P., and Noda-García, L., 2011. What can genome-scale metabolic network reconstructions do for prokaryotic systematics? Antonie van Leeuwenhoek 101, 35–43.

Brady, C.L., 2008. Taxonomic evaluation of the genus *Pantoea* based on a multigene approach, PhD (Microbiology) thesis, University of Pretoria, Pretoria.

Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P., and Karp, P.D., 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic acids research 34, D511–D516.

Chang, A., Scheer, M., Grote, A., Schomburg, I., and Schomburg, D., 2009. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic acids research 37, D588–D592.

DeJongh, M., Formsma, K., Boillot, P., Gould, J., Rycenga, M., and Best, A., 2007. Toward the automated generation of genome-scale metabolic networks in the SEED. BMC bioinformatics 8, 139.

De Maayer, P., Chan, W.Y., Venter, S.N., Toth, I.K., Birch, P.R.J., Joubert, F., and Coutinho, T.A., 2010. Genome Sequence of *Pantoea ananatis* LMG20103, the Causative Agent of Eucalyptus Blight and Dieback. Journal of Bacteriology 192, 2936–2937.

Durot, M., Bourguignon, P.Y., and Schachter, V., 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS microbiology reviews 33, 164–190.

Evsikov, A.V., Dolan, M.E., Genrich, M.P., Patek, E., and Bult, C.J., 2009. MouseCyc: a curated biochemical pathways database for the laboratory mouse. Genome biology 10, R84.

Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., and Palsson, B.Ø., 2008. Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology 7, 129–143.

Green, M., and Karp, P., 2005. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. Nucleic acids research 33, 4035–4039.

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O., 2004. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Research 33, W557 –W559.

Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T., 2002. Co-clustering of biological networks and gene expression data. Bioinformatics 18, S145 –S154.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L., 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotech 28, 977–982.

Hiratsuka, T., Furihata, K., Ishikawa, J., Yamashita, H., Itoh, N., Seto, H., and Dairi, T., 2008. An Alternative Menaquinone Biosynthetic Pathway Operating in Microorganisms. Science 321, 1670–1673.

Imielinski, M., and Belta, C., 2008. Exploiting the pathway structure of metabolism to reveal high-order epistasis. BMC systems biology 2, 40.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic acids research 34, D354–D357.

Karp, P.D., Keseler, I.M., Altman, T., Caspi, R., Fulcher, C.A., Subhraveti, P., Kothari, A., Krummenacker, M., Latendresse, M., Lee, T., Paley, S.., Shearer, A.G., and Trupp, M., 2011. BioCyc: Microbial Genomes and Cellular Networks. Microbe 6, 176–182.

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., and López-Bigas, N., 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic acids research 33, 6083–6089.

Karp, P.D., Paley, S., and Romero, P., 2002. The pathway tools software. Bioinformatics 18, S225–S232.

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I.M., and Caspi, R., 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in bioinformatics 11, 40–79.

Kauffman, K.J., Prakash, P., and Edwards, J.S., 2003. Advances in flux balance analysis. Current Opinion in Biotechnology 14, 491–496.

Keseler, I.M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R.P., Johnson, D.A., Krummenacker, M., Nolan, L.M., Paley, S., Paulsen, I.T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A.G., and Karp, P.D., 2009. EcoCyc: a comprehensive view of Escherichia coli biology. Nucleic acids research 37, D464–D470.

Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A.G., Mackie, A., Paulsen, I., Gunsalus, R.P., and Karp, P.D., 2011. EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic acids research 39, D583–D590.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 27, 29–34.

Oh, Y.-K., Palsson, B.O., Park, S.M., Schilling, C.H., and Mahadevan, R., 2007. Genome-scale Reconstruction of Metabolic Network in Bacillus subtilis Based on High-throughput Phenotyping and Gene Essentiality Data. Journal of Biological Chemistry 282, 28791–28799.

Osterman, A., and Overbeek, R., 2003. Missing genes in metabolic pathways: a comparative genomics approach. Current Opinion in Chemical Biology 7, 238–251.

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

Paley, S.M., and Karp, P.D., 2002. Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. Bioinformatics 18, 715–724.

Paley, S.M., and Karp, P.D., 2006. The pathway tools cellular overview diagram and omics viewer. Nucleic acids research 34, 3771–3778.

Pinney, J.W., Shirley, M.W., McConkey, G.A., and Westhead, D.R., 2005. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. Nucleic acids research 33, 1399–1409.

Reed, J.L., Famili, I., Thiele, I., and Palsson, B.O., 2006. Towards multidimensional genome annotation. Nature Reviews Genetics 7, 130–141.

Risso, C., Van Dien, S.J., Orloff, A., Lovley, D.R., and Coppi, M.V., 2008. Elucidation of an Alternate Isoleucine Biosynthesis Pathway in Geobacter sulfurreducens. Journal of Bacteriology 190, 2266–2274.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., and Karp, P.D., 2004. Computational prediction of human metabolic pathways from the complete human genome. Genome biology 6, R2.

Sandoval-Calderon, M., Geiger, O., Guan, Z., Barona-Gomez, F., and Sohlenkamp, C., 2009. A Eukaryote-like Cardiolipin Synthase Is Present in Streptomyces coelicolor and in Most Actinobacteria. Journal of Biological Chemistry 284, 17383–17390.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., and Schomburg, D., 2010. BRENDA, the enzyme information system in 2011. Nucleic Acids Research.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D., 2004. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Research 32, 431D–433.

Schomburg, I., Chang, A., and Schomburg, D., 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Research 30, 47 –49.

Thiele, I., and Palsson, B.Ø., 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5, 93–121.

Vallenet, D., Engelen, S., Mornico, D., Cruveiller, S., Fleury, L., Lajus, A., Rouy, Z., Roche, D., Salvignol, G., Scarpelli, C., and Medigue, C., 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. Database 2009, bap021–bap021.

Vieira, G., Sabarly, V., Bourguignon, P.Y., Durot, M., Le Fèvre, F., Mornico, D., Vallenet, D., Bouvet, O., Denamur, E., Schachter, V., and Médigue, C., 2011. Core and panmetabolism in Escherichia coli. Journal of bacteriology 193, 1461–1472.

Winn, W.C., and Koneman, E.W., 2006. Koneman’s color atlas and textbook of diagnostic microbiology. Lippincott Williams & Wilkins.

Zamboni, N., Kümmel, A., and Heinemann, M., 2008. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. BMC bioinformatics 9, 199.

Table 4.1: Summary of Pathway tools metabolic pathway network model predictions for *P. ananatis* LMG 20103

	Original PT prediction	**After comparison and filling missing reactions**
Total number of pathways predicted	322	
Number of complete pathways	245	257
Number of incomplete pathways	87	75
Total number of reactions	1039	
Number of completed reactions	776 (75%)	829 (80%)
Number of missing reactions	263 (25%)	210 (20%)

Table 4.2: Summary of Model SEED pathway network model predictions for *P. ananatis* LMG 20103

	Original SEED prediction	**After comparison and filling missing reactions**
Total number of scenarios	361	
Number of completed scenarios	231 (64%)	240 (66%)
Number of incomplete scenarios	130 (36%)	121 (34%)

Table 4.3: List of candidate genes proposed to be associated with missing reactions in the PGDB model

FadB	PANA_0201	Fatty oxidation complex alpha subunit [Includes: Enoyl-CoA hydratase
TyrB	PANA_0267	Aromatic-amino-acid aminotransferase
HprA	PANA_0570	Glycerate dehydrogenase
MtnN	PANA_0783	MTA/SAH nucleosidase
FabZ	PANA_0803	(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase
PANA0880	PANA_0871	Sugar Aldolase
PANA0881	PANA_0872	Protein UTR4
PANA0883	PANA_0874	Putative translation initiation factor eIF-2B
FolD	PANA_1060	FolD bifunctional protein[Include:Methylenetetrahydrofolate dehydrogenase]
SdhC	PANA_1171	Succinate dehydrogenase cytochrome b556 subunit
SdhD	PANA_1172	Succinate dehydrogenase hydrophobic membrane anchor protein
SdhA	PANA_1173	Succinate dehydrogenase flavoprotein subunit
SdhB	PANA_1174	Succinate dehydrogenase iron- sulfur protein
PANA1278	PANA_1264	Urocanate hydratase
PANA1282	PANA_1268	Atrazine Chlorohydrolase
PANA1284	PANA_1270	N-Formylglutamate Amidohydrolase
FabA	PANA_1390	3-hydroxydecanoyl-[acyl-carrier- protein] dehydratase
MocC	PANA_1626	Rhizopine catabolism protein MocC
PfkA	PANA_1860	6-phosphofructokinase
PntB	PANA_1994	NAD(P) transhydrogenase subunit beta
PntA	PANA_1995	NAD(P) transhydrogenase subunit alpha
FabI	PANA_2035	Enoyl-[acyl-carrier-protein] reductase [NADH]
BtuR	PANA_2060	Cob(I)yrinic acid a,c-diamide adenosyltransferase
PncA	PANA_2106	Pyrazinamidase/nicotinamidase
MsbB	PANA_2195	Lipid A biosynthesis
HisC	PANA_2470	Histidinol-phosphate aminotransferase
Udk	PANA_2511	Uridine kinase
FruK	PANA_2559	1-phosphofructokinase
NrdA	PANA_2617	Ribonucleoside-diphosphate reductase 1 alpha subunit
NrdB	PANA_2618	Ribonucleoside-diphosphate reductase 1 beta subunit
CcmD	PANA_2703	Heme exporter protein D
GlpQ	PANA_2710	Glycerophosphoryl diester phosphodiesterase
MmuM	PANA_2986	Homocysteine S-methyltransferase
NrdI	PANA_2991	Protein NrdI
NrdE	PANA_2992	Ribonucleoside-diphosphate reductase 2 alpha subunit
NrdF	PANA_2993	Ribonucleoside-diphosphate reductase 2 beta subunit
DeoD	PANA_3299	Purine nucleoside phosphorylase
HldE	PANA_3380	Bifunctional protein HldE [Includes: D-beta-D-heptose 7-phosphate kinase
UgpQ	PANA_3713	Glycerophosphoryl diester phosphodiesterase
IolC	PANA_3736	IolC Protein

IolD	PANA_3737	Probable malonic semialdehyde oxidative decarboxylase
IolE	PANA_3739	Inosose dehydratase
UgpQ	PANA_3781	Glycerophosphoryl diester phosphodiesterase
Mqo	PANA_3809	Malate:quinone oxidoreductase
SthA	PANA_3840	Soluble pyridine nucleotide transhydrogenase
PfkA	PANA_3871	6-phosphofructokinase
CoaBC	PANA_3907	Coenzyme A biosynthesis bifunctional protein CoaBC
MocC	PANA_4051	Rhizopine catabolism protein MocC
Mqo	PANA_4112	Probable malate:quinone oxidoreductase
DcuS	PANA_4114	Sensor protein DcuS
DcuR	PANA_4115	Transcriptional regulatory protein DcuR
ApbE	PANA_4116	Thiamine biosynthesis lipoprotein ApbE
YiaE	PANA_4154	Putative 2-hydroxyacid dehydrogenase HI1556

Table 4.4: List of incomplete pathways associated with *Pantoea ananatis* LMG 20103

Types of compounds	**Types of function**	**Pathway name**
Amino acids	**Biosynthesis and metabolism**	Cysteine biosynthesis/homocysteine degradation
		L-glutamine biosynthesis (tRNA-dependent)
		Serine racemization
	Degradation	Homocysteine degradation
Carbohydrate	**Biosynthesis and metabolism**	GDP-D-rhamnose biosynthesis
		GDP-glucose biosynthesis
		GDP-mannose biosynthesis
	Degradation and recycling	D-allose degradation
		Galactose degradation (Leloir pathway)
Fatty acid and lipid	**Biosynthesis and metabolism**	Cis-vaccenate biosynthesis
		Palmitate biosynthesis (bacteria and plants)
Amines and Polyamines	**Biosynthesis**	Ectoine biosynthesis
	Degradation	Putrescine degradation
Cell structure	**Biosynthesis and metabolism**	Enterobacterial common antigen biosynthesis
Co-factors, Prosthetic groups, Electron carriers	**Biosynthesis and metabolism**	6-hydroxymethyl-dihydropterin diphosphate biosynthesis
		Folate polyglutamylation

		Folate transformations
		FormylTHF biosynthesis
		Tetrahydrofolate biosynthesis
		Thiamin biosynthesis
		Thioredoxin pathway
		Ubiquinone-8 biosynthesis (prokaryotic)
Non-specific	**Generation of precusors metabolites and energy**	Acetoin biosynthesis
		Aerobic respiration - electron donor II
		Entner-Doudoroff pathway
		Mixed acid fermentation
		Respiration (anaerobic)
Aminoacyl-tRNA charging	**Biosynthesis**	Asparagine biosynthesis (tRNA-dependent)
		L-glutamine biosynthesis (tRNA-dependent)
Aromatic compound	**Biosynthesis**	2,3-dihydroxybenzoate biosynthesis
	Degradation	Parathion degradation
		Quinate degradation
		Shikimate degradation
Other	**Biosynthesis**	Mannosylglycerate biosynthesis
		Autoinducer AI-2 biosynthesis
Secondary metabolite	**Biosynthesis**	*myo*-inositol biosynthesis
		Neurosporene biosynthesis
		Zeaxanthin-beta-D-diglucoside biosynthesis
	Degradation	1,6-anhydro-N-acetylmuramic acid recycling
		Beta-D-glucuronide and D-glucuronate degradation
1C compounds	**Degradation**	Formaldehyde oxidation (tetrahydrofolate pathway)

		Reductive monocarboxylic acid cycle
Carboxylates	**Degradation**	L-ascorbate degradation, anaerobic
Inorganic nutrients	**Degradation**	Nitrate reduction

Table 4.5: List of complete pathways associated with *Pantoea ananatis* LMG 20103

Types of compounds	Types of function	Pathway name
Nucleotides and nucleosides	**Biosynthesis and metabolism**	5-aminoimidazole ribonucleotide biosynthesis
		Adenosine nucleotides de novo biosynthesis
		Guanosine nucleotides de novo biosynthesis
		Inosine-5’-phopshate biosynthesis
		Pyrimidine deoxyribonucleotides de novo biosynthesis
		Pyrimidine ribonucleotides interconversion
		Salvage pathways of adenine, hypoxanthine, and their nucleosides
		Salvage pathways of guanine, xanthine, and their nucleosides
		Salvage pathways of purine and pyrimidine nucleotides
		Salvage pathways of pyrimidine deoxyribonucleotides
		Salvage pathways of pyrimidine ribonucleotides
		Uridine-5’-phosphate biosynthesis
	Degradation	Degradation of pyrimidine ribonucleosides
		Degradation of purine ribonucleosides
		Purine deoxyribonucleosides degradation
		Pyrimidine deoxyribonucleosides degradation
Amino acids	**Biosynthesis and metabolism**	Alanine biosynthesis
		Arginine biosynthesis
		Asparagine biosynthesis
		Beta-alanine biosynthesis
		Cysteine biosynthesis
		Glutamate biosynthesis
		Aspartate biosynthesis (Glutamate degradation)
		Glutamine biosynthesis
		Glycine biosynthesis
		Histidine biosynthesis
		Homoserine biosynthesis
		Isoleucine biosynthesis (from threonine)
		Leucine biosynthesis
		Lysine biosynthesis
		Methionine biosynthesis

		Ornithine biosynthesis
		Phenylalanine biosynthesis
		Proline biosynthesis
		S-adenosyl-L-methionine cycle
		Serine biosynthesis
		Threonine biosynthesis from homoserine
		Tryptophan biosynthesis
		Tyrosine biosynthesis
		Valine biosynthesis
	Degradation	2-ketoglutarate dehydrogenase complex
		Arginine degradation
		Asparagine degradation
		Aspartate degradation
		Glutamate degradation
		Glutamine degradation
		Glycine cleavage complex
		L-cysteine degradation
		L-serine degradation
		Lysine degradation
		Methionine degradation
		Ornithine degradation (proline biosynthesis)
		Proline degradation
		Taurine degradation
		Threonine degradation
		Tryptophan degradation
Carbohydrate	**Biosynthesis and metabolism**	ADP-L-glycero-beta-D-manno-heptose biosynthesis
		Cellulose biosynthesis
		CMP-KDO biosynthesis
		dTDP-L-rhamnose biosynthesis
		Gluconeogenesis
		Glycogen biosynthesis (from ADP-D-Glucose)
		Glycogen degradation
		Sorbitol biosynthesis
		Trehalose biosynthesis
		UDP-D-xylose biosynthesis
	Degradation and recycling	2-O-alpha-mannosyl-D-glycerate degradation
		D-mannose degradation

		Fructose degradation
		Glucose and glucose-1-phosphate degradation
		Glycogen degradation
		L-arabinose degradation
		L-lyxose degradation
		Lactose degradation
		Melibiose degradation
		Rhamnose degradation
		Ribose degradation
		Sucrose degradation
		Trehalose degradation
		Xylose degradation
Fatty acid and lipid	**Biosynthesis and metabolism**	Acyl-CoA hydrolysis
		Biotin-carboxyl carrier protein assembly
		Cardiolipin biosynthesis
		CDP-diacylglycerol biosynthesis
		Cis-dodecenoyl biosynthesis
		Cyclopropane fatty acid (CFA) biosynthesis
		Fatty acid activation
		Fatty acid biosynthesis initiation
		Fatty acid elongation – saturated
		KDO transfer to lipid IV_A
		KDO_2-lipid A biosynthesis I
		Lipid-A-precursor biosynthesis
		Palmitoleate biosynthesis
		Phosphatidylethanolamine biosynthesis
		Phosphatidylglycerol biosynthesis
	Degradation	Fatty acid beta-oxidation
		Oleate beta-oxidation
Amines and Polyamines	**Biosynthesis**	Aminopropylcadaverine biosynthesis
		Arginine degradation
		Choline degradation
		Glycine betaine biosynthesis (gram-negative bacteria)
		Putrescine biosynthesis
		Spermidine biosynthesis
		Trypanothione biosynthesis

		UDP-N-acetyl-D-glucosamine biosynthesis
	Degradation	4-aminobutyrate degradation
		Choline degradation
		N-acetylglucosamine degradation
Cell structure	**Biosynthesis and metabolism**	Peptidoglycan biosynthesis
		dTDP-L-rhamnose biosynthesis
		KDO transfer to lipid IV_A
		Lipid-A-precursor biosynthesis
		O-antigen biosynthesis (combine of dTDP-L-rhanose biosynthesis and UDP-N-acetyl-D-glucosamine biosynthesis
		UDP-N-acetyl-D-glucosamine biosynthesis
		UDP-N-acetylmuramoyl-peptapeptide biosynthesis
	Degradation	Phosphate utilization in cell wall regeneration
Co-factors, Prosthetic groups, Electron carriers	**Biosynthesis and metabolism**	Acyl carrier protein metabolism
		Adenosylcobalamin salvage from cobalamin
		Aminopropanol biosynthesis
		Biotin biosynthesis
		Coenzyme A biosynthesis
		di-trans,poly-cis-undecaprenyl phosphate biosynthesis
		Flavin biosynthesis (bacteria)
		Geranyldiphosphate biosynthesis
		Geranylgeranyldiphosphate biosynthesis
		Glutathion biosynthesis
		Glutathion redox reactions
		Heme biosynthesis from uroporphyrinogen-III
		Tetrapyrrole biosynthesis
		Heptaprenyl diphosphate biosynthesis
		Hexaprenyl diphosphate biosynthesis
		Lipoate biosynthesis and incorporation
		Lipoate salvage and modification
		Menaquinone-8 biosynthesis
		Methylerythritol phosphate pathway

		Molybdopterin guanine dinucleotide biosynthesis
		NAD biosynthesis (from aspartate)
		NAD salvage pathway
		Octaprenyl diphosphate biosynthesis
		Pantothernate biosynthesis
		Pyridoxal 5’-phosphate biosynthesis
		Pyridoxal 5’-phosphate salvage pathway
		S-adenosylmethionine biosynthesis
		Siroheme biosynthesis
		Trans-trans-farnesyl diphosphate biosynthesis
Non-specific	**Generation of precursors metabolites and energy**	(R)-acetoin biosynthesis
		2-ketoglutarate dehydrogenase complex
		Acetyl-CoA biosynthesis (from pyruvate)
		Glycolysis
		Homolactic fermentation
		NADH to cytochrome *bd* oxidase electron transfer
		NADH to cytochrome *bo* oxidase electron transfer
		NADH to nitrate electron transfer
		NADH to trimethylamine N-oxide electron transfer
		Pentose phosphate pathway (non-oxidative branch)
		Pentose phosphate pathway (oxidative branch)
		Pentose phosphate pathway (partial)
		Pyruvate fermentation to acetate
		Pyruvate fermentation ot ethanol
		Succinate to cytochrome *bd* oxidase electron transfer
		Succinate to cytochrome *bo* oxidase electron transfer
		Sulfide oxidation I (sulfide-quinone reductase)
		TCA cycle
Aminoacyl-tRNA charging	**Biosynthesis**	tRNA charging pathway
Aromatic compound	**Biosynthesis**	3-dehydroquinate biosynthesis

		4-hydroxybenzoate biosynthesis (bacteria and fungi)
		Chorismate biosynthesis
	Degradation	Methyl parathion degradation
		Paraoxon degradation
		Protocatechuate degradation
Metabolic regulators	**Biosynthesis**	ppGpp biosynthesis
Other	**Biosynthesis**	PRPP biosynthesis
	Degradation	Seed germination protein turnover
		Wound-induced proteolysis
Secondary metabolite	**Biosynthesis**	Zeaxanthin biosynthesis
	Degradation	D-arabitol degradation
		D-galactarate degradation
		D-galacturonate degradation
		D-glucarate degradation
		DIMBOA-glucoside degradation
		L-galactonate degradation
		L-idonate degradation
		Mannitol degradation
		N-acetylglucosamine degradation
		Sorbitol degradation
Alcohols	**Degradation**	Ethanol degradation
		Ethylene glycol degradation
		Glycerol degradation
		Oxidative ethanol degradation
Aldehydes	**Degradation**	L-lactaldehyde degradation (aerobic)
		Methylglyoxal degradation
1C compounds	**Degradation**	CO_2 fixation into oxaloacetate
		Formaldehyde oxidation (glutathion-dependent)
		Formaldehyde oxidation (thiol-dependent)
Carboxylates	**Degradation**	3-oxoadipate degradation
		Acetate conversion to acetyl-CoA

		Acetate formation from acetyl-CoA
		Acetyl-CoA biosynthesis (from pyruvate)
		Glycolate and glyoxylate degradation
Inorganic nutrients	**Degradation**	Sulfate activation for sulfonation
		Sulfide oxidation (sulfidequinone reductase)
		Thiosulfate desproportionation (rhodanese)
		Two-component alkanesulfonate monooxygenase
	Degradation	Phosphate acquisition
Acid resistance	**Detoxification**	Arginine dependent acid resistance
Aresenate detoxification	**Detoxification**	Aresenate detoxification (glutaredoxin)
Methylglyoxal Detoxification	**Detoxification**	Methylglyoxal degradation
Superoxide radicals Detoxification	**Detoxification**	Superoxide radicals degradation

Table 4.6: List of recognised transporters

Enzyme Number	Transporter Name	Substrate	Pathway evidence	Phenotypic evidence
Undetermined	2-dehydro-3-deoxy-D-gluconate transport via proton symport, reversible	2-keto-3-deoxygluconate	Y	
6.3.3.2	5-Formyltetrahydrofolate cyclo-ligase (ADP-forming)	Folate	Y	
Undetermined	Allantoin transport in via proton symport	Allantoin		
Undetermined	Ammonia transport via diffusion	Ammonia	Y	
Undetermined	Arbutin transport via PEP:Pyr PTS	Arbutin	Y	Y
Undetermined	Arginine/ornithine antiporter	Arginine and ornithine	Y	
TC-2.A.3.1,2.A.3.1	Aromatic amino acid transport protein AroP: PHEt6	L-Phenylalanine	Y	
Undetermined	Arsenate transporter	Arsenate	Y	
Undetermined	Arsenite transporter via uniport	Arsenite	Y	
Undetermined	Aspartate-H+ symport	Aspartate	Y	
.6.3.1	ATP synthase (four protons for one ATP)	Phosphate	Y	
2.7.6.3	ATP:2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine			
2.7.1.22	ATP:N-ribosylnicotinamide 5&#39;-phosphotransferase	Nicotinamide ribonucleotide	Y	
3.A.1.12	Betaine-ABC transport	Betaine	Y	
Undetermined	Branched chain amino acid:H+ symporter (Isoleucine)	L-Isoleucine	Y	
Undetermined	Branched chain amino acid:H+ symporter (Leucine)	L-Leucine	Y	
Undetermined	Butanesulfonate transport via ABC system	Butanesulfonate		

TC-3.A.3,3.A.3	Cadmium transport out via ABC system	Cadmium		
TC-2.A.19,2.A.19	Calcium transport in/out via proton antiporter	Calcium		
Undetermined	Cellobiose transport via PEP:Pyr PTS	Cellobiose	Y	Y
Undetermined	Cob(1)alamin transport via ABC system	Cob(1)alamin	Y	
Undetermined	Cobalamin uptake in via ABC transport	Cobalamin	Y	
Undetermined	Cobalt transport in/out via permease (no H+)	Cobalt		
TC-3.A.3,3.A.3	Copper export via ATPase	Copper	Y	
Undetermined	Copper transport via ABC system	Copper	Y	
Undetermined	cysteate transport via ABC system	Cysteate		
Undetermined	Cytidine ion-coupled transport	Cytidine	Y	
Undetermined	Cytosine transport in via proton symport	Cytosine	Y	
TC-2.A.3.1,2.A.3.1	D-alanine transport in via proton symport	D-Alanine	Y	Y
Undetermined	Deoxycytidine ion-coupled transport	Deoxycytidine	Y	
Undetermined	Deoxyguanosine transport in via proton symport	Deoxyguanosine	Y	
Undetermined	Deoxyinosine transport in via proton symport	Deoxyinosine	Y	
Undetermined	Deoxyribose transport via ABC system	Thyminose		
Undetermined	Deoxyuridine transport in via proton symport	Deoxyuridine	Y	
Undetermined	D-fructose transport via PEP:Pyr PTS	D-Fructose	Y	Y
Undetermined	D-Galactose ABC transport	D-Galactose	Y	Y

Undetermined	D-galactose transport in via proton symport	D-Galactose	Y	
Undetermined	D-gluconate transport via proton symport	D-Gluconate	Y	
Undetermined	D-glucose transport in via proton symport	D-Glucose	Y	Y
Undetermined	D-Glucose-ABC transport	D-Glucose	Y	Y
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-asp)	ala-asp		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-gln)	ala-gln		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-glu)	ala-glu	Y	
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-gly)	ala-gly		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-his)	ala-his		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-leu)	ala-leu		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (ala-thr)	ala-thr		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (c)	ala-thr		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (cgly)	cys-gly		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (gly-asp)	gly-asp		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (gly-gln)	gly-gln		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (gly-glu)	gly-glu		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (gly-met)	gly-met		
TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (gly-pro-L)	gly-pro-L		

TC-3.A.1.5,3.A.1.5	Dipeptide transport via ABC system (met-ala)	met-ala		
Undetermined	D-mannose transport via PEP:Pyr PTS	D-Mannose	Y	Y
Undetermined	D-Methionine-ABC transport	D-Methionine		
TC-3.A.1.2,3.A.1.2	D-ribose transport out via ABC system	D-Ribose	Y	Y
TC-2.A.3.1,2.A.3.1	D-serine transport in/out via proton symport	D-Serine	Y	
Undetermined	D-sorbitol transport via PEP:Pyr PTS	D-Sorbitol	Y	Y
Undetermined	D-Xylose-ABC transport	D-Xylose	Y	Y
Undetermined	Ethanesulfonate transport via ABC system	Ethanesulfonate		
TC-2.A.3.5,2.A.3.5	Ethanolamine transport in/out via proton symport	Aminoethanol		
6.2.1.3	Fatty-acid--CoA ligase (hexadecenoate), peroxisomal	Fatty-acid (hexadecenoate)	Y	
6.2.1.3	Fatty-acid--CoA ligase (octadecenoate)	Fatty-acid (octadecenoate)		
6.2.1.3	Fatty-acid--CoA ligase (tetradecanoate)	Long-chain-fatty-acid		
Undetermined	Fe(III)dicitrate-ABC transport	Fe(III)dicitrate		
Undetermined	Ferrous iron transport periplasmic protein EfeO: FE3t4	Ferrous iron	Y	
Undetermined	Formate transport in via proton symport	Formate		Y
1.2.1.2	Formate:NAD+ oxidoreductase	NAD		Y
.3.99.	Fumarate reductase	Fumarate	Y	
TC-2.A.3.1,2.A.3.1	Galactarate transport in via proton symport	Galactarate	Y	Y
TC-2.A.1.14,2.A.1.14	Galacturonate transport in via proton symport	Galacturonate	Y	Y
TC-	Glucarate transport in via proton	Glucarate	Y	Y

2.A.3.1,2.A.3.1	symport			
Undetermined	Glucose dehydrogenase (ubiquinone-8 as acceptor)	Glucose	Y	
Undetermined	Glucose-phosphotransferase (PTS) system	Glucose	Y	
TC-2.A.1.14,2.A.1.14	Glucuronate transport in via proton symport	Glucuronate	Y	Y
Undetermined	Glycerol transport in/out via diffusion reversible	Glycerol		Y
3.A.1.5	Gly-Cys ABC transporters	Gly-Cys		
3.A.1.5	Gly-Leu ABC transporters	Gly-Leu		
3.A.1.5	Gly-Phe ABC transporters	Gly-Phe		
3.A.1.5	Gly-Try ABC transporters	Gly-Try		
Undetermined	Hexanesulfonate transport via ABC system	Hexanesulfonate		
.18.99.	Hydrogenase (ubiquinone-8: 2 protons)	Hydrogen	Y	
Undetermined	Hypoxanthine ion-coupled transport	Hypoxanthine	Y	
Undetermined	Inositol transport in via proton symport	Inositol	Y	Y
Undetermined	Iron (II) transport via ABC system	Iron (II)	Y	
Undetermined	Iron (III) dicitrate transport via ABC system	Iron (III)		
Undetermined	Isethionate transport via ABC system	Isethionate		
Undetermined	L-arabinose transport via proton symport	L-Arabinose	Y	Y
3.A.1.2	L-Arabinose-ABC transport	L-Arabinose	Y	Y
Undetermined	L-asparate transport via proton symport (3 H)	L-asparate		
Undetermined	L-Aspartate-ABC transport	L-Aspartate	Y	Y
Undetermined	Lead (Pb+2) ABC transporter	Lead		
Undetermined	L-Glutamate-ABC transport	L-Glutamate	Y	Y
Undetermined	L-Isoleucine-ABC transport	L-Isoleucine	Y	

Undetermined	L-lactate reversible transport via proton symport	L-lactate	Y	
Undetermined	L-Leucine-ABC transport	L-Leucine	Y	
Undetermined	L-Lysine-ABC transport	L-Lysine	Y	
Undetermined	L-Methionine ABC transport	L-Methionine	Y	
Undetermined	L-methionine R-oxide transport via ABC system	L-methionine R-oxide	Y	
Undetermined	L-methionine S-oxide transport via ABC system	L-Methionine S-oxide	Y	
1.8.4.13,1.8.4.14	L-methionine:oxidized-thioredoxin S-oxidoreductase	L-Methionine	Y	
6.2.1.3	Long-chain-fatty-acid--CoA ligase: FACOAL140(ISO)	Long-chain-fatty-acid		
6.2.1.3	Long-chain-fatty-acid--CoA ligase: FACOAL150(anteiso)	Long-chain-fatty-acid		
6.2.1.3	Long-chain-fatty-acid--CoA ligase: FACOAL150(ISO)	Long-chain-fatty-acid		
6.2.1.3	Long-chain-fatty-acid--CoA ligase: FACOAL160(ISO)	Long-chain-fatty-acid		
6.2.1.3	Long-chain-fatty-acid--CoA ligase: FACOAL170(anteiso)	Long-chain-fatty-acid		
6.2.1.3	Long-chain-fatty-acid--CoA ligase: FACOAL170(ISO)	Long-chain-fatty-acid		
Undetermined	L-Ornithine-ABC transport	L-Ornithine	Y	
TC-2.A.3.1,2.A.3.1	Low-affinity inorganic phosphate transporter: TYRt6	L-Tyrosine	Y	
Undetermined	L-proline transport in via proton symport	L-proline	Y	Y
Undetermined	L-Proline/Glycine betaine transporter ProP	L-Proline	Y	Y
Undetermined	L-threonine transporter: THRt2	L-Threonine	Y	
TC-2.A.3.1,2.A.3.1	L-tryptophan transport in via proton symport	L-tryptophan	Y	Y
Undetermined	L-valine transport in via proton symport	L-valine	Y	
Undetermined	L-Valine-ABC transport	L-Valine	Y	

TC-1.A.35,1.A.35	Magnesium transport in/out via permease (no H+)	Magnesium		
Undetermined	Malate-H+/Na+-lactate antiporter	Malate-H+ and Na+-lactate	Y	Y
Undetermined	Maltohexaose transport via ABC system	Maltohexaose		
Undetermined	Maltose transport via PEP:Pyr PTS	Maltose	Y	Y
3.A.1.1	Maltose-ABC transport	Maltose	Y	Y
Undetermined	Maltotriose transport via ABC system	Maltotriose	Y	Y
TC-2.A.55,2.A.55	Manganese transport in via proton symport	Manganese		
3.A.1.15	Manganese-ABC transport	Manganese		
Undetermined	Mannitol transport via PEP:Pyr PTS	Mannitol	Y	Y
Undetermined	Mercury (Hg+2) ABC transporter	Mercury		
Undetermined	Methanesulfonate transport via ABC system	Methanesulfonate		
TC-3.A.1.8,3.A.1.8	Molybdate transport via ABC system	Molybdate		
Undetermined	MOPS transport via ABC system	MOPS		
Undetermined	Na+ lactate / malate H+ antiporter	D-Lactate	Y	Y
Undetermined	Na+/malate symporter	Na+ and malate	Y	
Undetermined	Na+:proline symport	L-Proline	Y	
Undetermined	N-Acetyl-D-glucosamine transport via PEP:Pyr PTS	N-Acetyl-D-glucosamine	Y	
.6.1.	NAD(P) transhydrogenase	NAD(P)	Y	
.6.1.1,1.6.1.	NADPH:NAD+ oxidoreductase (B-specific)	NAD	Y	
1.8.1.8	NADPH:protein-disulfide oxidoreductase	NADPH	Y	

Undetermined	Nitrate reductase (Menaquinol-8)	Nitrate	Y	
.7.99.	Nitrate reductase (Ubiquinol-8)	Nitrate	Y	
Undetermined	Nitrate transport in via proton symport	Nitrate	Y	
Undetermined	Nitrite transport in via proton symport	Nitrite	Y	
Undetermined	NMN permease	Nicotinamide ribonucleotide	Y	
Undetermined	Nucleoside permease NupC: ADNt2r	Adenosine	Y	
Undetermined	Nucleoside permease NupC: DADNt2	Deoxyadenosine	Y	
Undetermined	Nucleoside permease NupC: INSt2r	Inosine	Y	Y
TC-3.A.1.7,3.A.1.7	Orthophosphate-ABC transport	Phosphate	Y	
Undetermined	Peptidoglycan subunit synthesis	Peptidoglycan polymer	Y	
Undetermined	Phosphate ABC transporter permease protein	Phosphate	Y	
Undetermined	Potassium ABC transporter	Potassium		
TC-2.A.37,2.A.37	Potassium transport out via proton antiport	Potassium		
Undetermined	Proton sodium antiport	Sodium		
Undetermined	Salicin transport via PEP:Pyr PTS	Salicin	Y	Y
TC-2.A.42,2.A.42	Serine transporter: SERt2	L-Serine	Y	Y
Undetermined	sn-Glycerol ABC transport	sn-Glycerol	Y	
TC-2.A.20,2.A.20	Sodium-dependent phosphate transporter: PIt6	Phosphate	Y	
TC-2.A.58,2.A.58	Sodium-dependent phosphate transporter: PIt8	Phosphate	Y	
1.3.99.1	Succinate dehydrogenase (irreversible)	Succinate	Y	Y

Undetermined	Succintate transport via proton symport (3 H)	Succinate	Y	Y
Undetermined	Sucrose transport via PEP:Pyr PTS	Sucrose	Y	Y
TC-2.A.53,2.A.53	Sulfate transport in via proton symport	Sulfate	Y	
Undetermined	Sulfate-ABC transport	Sulfate	Y	
Undetermined	Sulfoacetate transport via ABC system	Sulfoacetate		
Undetermined	Thiamin-ABC transport	Thiamin	Y	
Undetermined	Thiamine transport in via proton symport	Thiamin	Y	
.6.4.5,1.8.1.	Thioredoxin reductase: TRDR	Thioredoxin	Y	
Undetermined	Thiosulfate-ABC transport	Thiosulfate	Y	
Undetermined	Thymidine ion-coupled transport	Thymidine	Y	Y
Undetermined	Transporter, LysE family: LYSt3r	L-Lysine	Y	
TC-2.A.40,2.A.40	Uracil ion-coupled transport	Uracil	Y	
Undetermined	Urea transport via facilitate diffusion	Urea	Y	
Undetermined	Uridine ion-coupled transport	Uridine	Y	
Undetermined	Vitamin ABC transport	Vitamin	Y	
TC-2.A.40,2.A.40	Xanthine ion-coupled transport	Xanthine	Y	
TC-3.A.1.15,3.A.1.15	Zinc-ABC transport	Zinc		

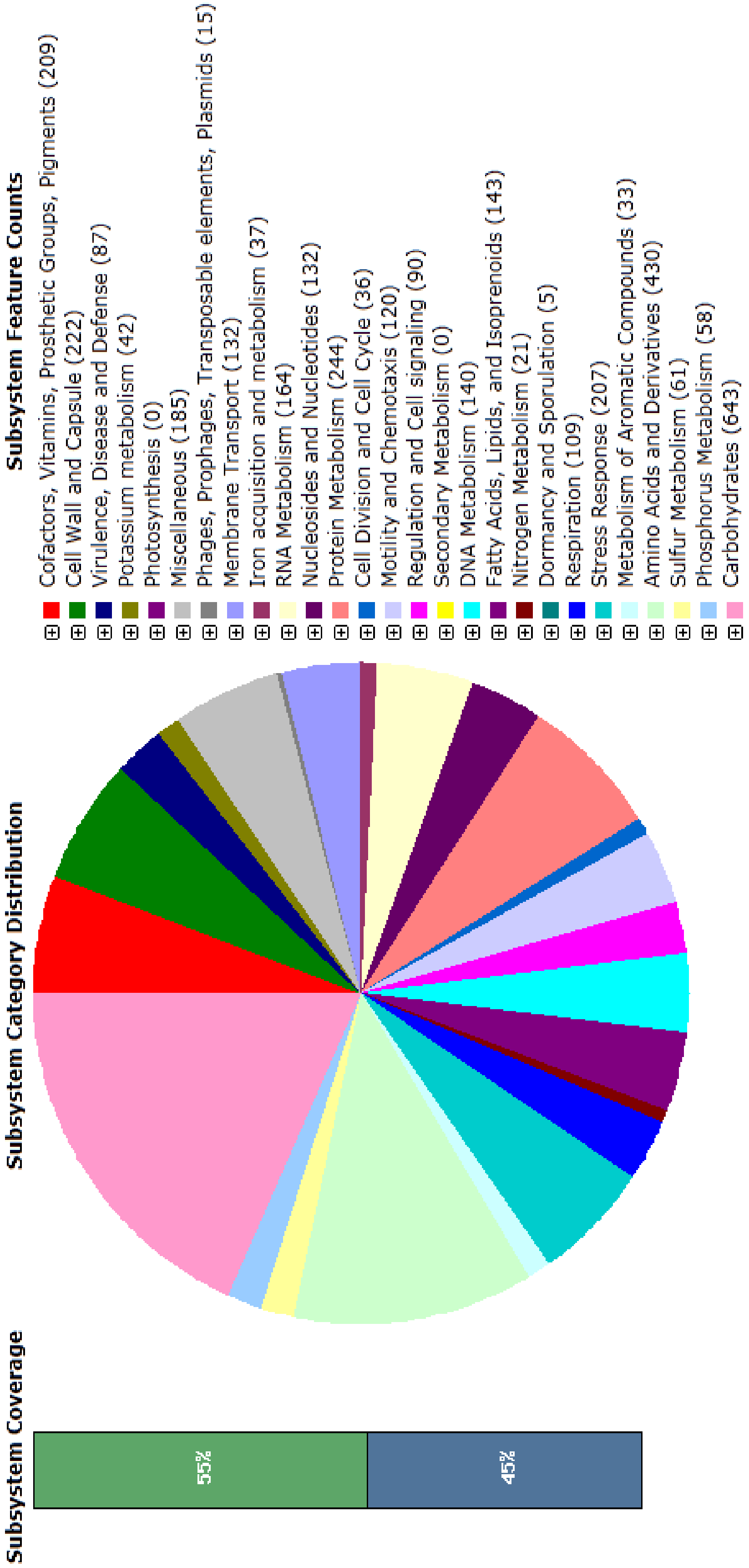


Figure 4.2 : Analysis of the subsystem distribution and coverage of *Pantoea ananatis LMG 20103* as predicted by the RAST server.

Chapter 5

Conclusion

Conclusion

Since the first genome was sequenced in 1979 by Roger Staden, more complex organisms, e.g. bacteria and human genomes, have been subjected to the same approach. The shotgun sequencing method was used and the genomes assembled by overlapping the sequence fragments based on sequence similarity. With the reduced cost and time required by next generation sequencing technologies (NGS), obtaining a genome sequence has become an essential resource for studies related to the evolution, ecology and biology of an organism.

Pantoea ananatis has been isolated from a wide-range of hosts and is well known for its pathogenic, epiphytic and endophytic behaviour. In 1998, *P. ananatis* was responsible for an outbreak of bacterial blight and dieback of *Eucalyptus* in a South African nursery. At that time very little was known about this pathogen. A better understanding of the pathogenicity and ecology of the bacterium is required in order to develop strategies to control the disease in an attempt to avoid further outbreaks and financial losses.

In order to study the evolution and biology of *P. ananatis*, it was realised that only a complete genome sequence would provide the solid foundation required for the envisaged studies to understand the pathogenicity and ecology of this bacterium. For this purpose the genomes of *P. ananatis* strain LMG 20103 (a strain pathogenic to *Eucalyptus*) and LMG 2665 (Type strain pathogenic to pineapple) were sequenced with the 454 Roche GS 20 and Illumina sequence analyser NGS technologies, respectively. The genome assemblies were done with either Newbler or Velvet after various attempts to optimize the assembly parameters. Numerous assembly problems could not be resolved using the assembler programmes on their own and manual curation to complete the genomes was required. The main focus, therefore, shifted towards the complete assembly and annotation of the *Eucalyptus* pathogen, *P. ananatis* LMG 20103. A combination of both *in-vitro*, and *in-silico* approaches was required to scaffold contigs, resolve regions with repeat sequences and fill in gap sequences in order to complete the assembly.

One of the benefits of having a complete genome is that it can be used for reconstructing the organism's metabolic pathways and network. As more genome sequences are becoming available and the demand for integrated metabolic data increases, programmes for the automated reconstruction of metabolic pathways are constantly being improved to ensure the reliability and accuracy of the pathway predictions. Apart from using this data to comprehend the metabolic capabilities of the organism, it was also realised that the metabolic network data could help to evaluate and improve the initial genome annotation. Using the complete genome sequence of *P. ananatis* LMG 20103, the metabolic pathways and network were, therefore, reconstructed using both the Pathway Tools and Model SEED pipelines. Incomplete pathways and differences between the predicted networks were noted and improved through further manual curation. Although some improvements could be made to the predicted network and annotations, further experimental data is still required to improve the accuracy of the draft metabolic network.

Based on the findings of the study and the experience gained during completion of the *P. ananatis* LMG 20103 genome, a number of conclusions and observations were made which can be summarized as follows:

- Completing the genome assembly requires extensive manual curation of the genome assembly and annotation, and additional experimental data to verify the problematic regions within the assembly.
- Genome assemblers are designed to handle data from specific NGS technologies and optimisation of assembly parameters is often required. Comparison of draft genome sequences between closely related species or different assembly versions of the same draft can assist with genome sequence scaffolding and assembly.
- The termination of contig extension during assembly (gaps) and mis-assemblies by the Newbler assembler are mostly due to low coverage of regions or the presence of repetitive sequences including numerous copies of the ribosomal DNA.
- Contig graphs and adjacency information are useful for genome sequence scaffolding.

- SNP identification and coverage analyses are useful to detect mis-assemblies and repetitive sequences.
- Well-designed conventional PCRs are the most reliable approach for gap closure and validation of the draft assembly.
- Using the genome sequence, a genomic-scale metabolic network can be reconstructed with the advanced bioinformatic pipelines such as Pathway Tools and Model SEED.
- Missing metabolic reactions and incomplete pathways in the draft metabolic network are mainly caused by missing genome sequences, incorrect gene annotations, or bioinformatic errors during the automated network reconstruction.
- Comparison of metabolic networks created by different pipelines and approaches can assist with the identification of candidate genes to fill the missing pathway reactions.
- With the limited experimental data for pathways validation, curation of the reconstructed network is still essential to generate a reliable network for future experiments.
- *Pantoea ananatis* is capable of *de novo* synthesis of most of the essential compounds required for growth. In addition it is also equipped with numerous transporter proteins for the uptake of available resources from the surrounding environment.

During this study a complete genome assembly and draft metabolic network could be constructed for *P. ananatis* LMG 20103. The genome consisted of one completed circular chromosome of 4 386 227 bp and a megaplasmid with a size of 317 146 bp. Despite the amount of effort and cost, it is believed that the complete genome and draft metabolic network will be valuable resources for many subsequent studies to investigate the evolution and biology of this emerging plant pathogen.