Towards unravelling the genome of *Avibacterium paragallinarum*

by

Yolandi Roodt

Towards unravelling the genome of *Avibacterium paragallinarum*

by

Yolandi Roodt

Submitted in fulfillment of the requirements
for the degree

Philosophiae Doctor

In the Faculty of Natural and Agricultural Sciences
Department of Microbial, Biochemical and Food Biotechnology
University of the Free State
Bloemfontein
South Africa

2009

Promoter: Prof. J. Albertyn
Co-promoter: Prof. R. R. Bragg

Acknowledgements

I would like to express my gratitude to the following people and institutions:

My promoters Prof. Albertyn and Prof. Bragg for their time and advice.

Prof. Herik Christensen from the Department of Veterinary Disease Biology, Center for Applied Bioinformatics, Denmark for the A-1 (0083) raw contig set.

To The National Research Foundation and the Strategic Academic Cluster for financial support.

All the angels on my way:

Pieter Muller, don't know where I'd be without your help.
Sanet for her friendship and encouragement.
Jinx and Dez for their friendship.
Michel for his friendship and all his ideas and advice with regards to this project.

My mother for enabling me to commend my studies, for her love and everlasting support. Oom Dirk, thanks for always showing interest in my studies and work.

To my sister, An-Sophie, and brother, JD, thanks for your love and support, and for trying to understand even though it was very difficult at times.

Above all, *Our Heavenly Father*, for all the blessings He bestowed on me, and without Him I would not have come this far.

Well you know those times
When you feel like there's a sign there on your back
Says I don't mind if ya kick me Seems like everybody has
Things go from bad to worse You'd think they can't get worse than that
And then they do

You step off the straight and narrow And you don't know where you are
Use the needle of your compass To sew up your broken heart
Ask directions from a genie In a bottle of Jim Beam And she lies to you
That's when you learn the truth

If you're going through hell
Keep on going, don't slow down
If you're scared, don't show it
You might get out
Before the devil even knows you're there

Well I been deep down in that darkness I been down to my last match
Felt a hundred different demons Breathing fire down my back
And I knew that if I stumbled
I'd fall right into the trap that they were laying

But the good news
Is there's angels everywhere out on the street
Holding out a hand to pull you back up on your feet
The one's that you've been dragginig for so long
You're on your knees You maight as well be praying
Guess what I'm saying

If your going through hell
Keep on going, don't slow down
If you're scared don't show it
You might get out
Before the devil even knows you're there

If you're going through hell
Keep on moving, face that fire
Walk right through it
You might get out
Before the devil even knows you're there

Rodney Atkins – If you’re going through hell

This thesis is dedicated to my

Mother,

My sister,

My brother and

In loving memory of my Father

Table of contents

List of Abbreviations

att	Attachment site
attB	Bacterial attachment site
attC	Integron gene cassette attachment site
attP	Phage attachment site
Asp	Aspartic acid
ATP	Adenosine Tri-Phosphate
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
CCD	Charge-couple device
cm	Centimetre
˚C	Degrees Celsius
Da	Dalton
DNA	Deoxyribonucleic acid
dNTP's	Nucleotides
EDTA	Ethylene diaminetetraacetic acid
emPCR	Emulsion based polymerase chain reaction
FAD	Flavin adenine dinucleotide
HA	Haemagglutinin
IA	Iminosuccinic acid
IC	Infectious Coryza
IPTG	Isopropylthio-β-D-galactoside
IS	Insertion sequence
JCVI	J. Craig Venter Institute
kb	Kilobase
kDa	Kilodalton
LB	Luria-Bertani
M	Molar
$MgCl_2$	Magnesium chloride
mg/µl	Milligram per microlitre
mg/l	Milligram per litre
min	Minutes

min/kb	Minutes per kilobasepair
ml	Millilitre
mM	Millimolar
m/v	Mass per volume
Mb	Megabases
NA	Nicotinic acid
NAaD	Nicotinate adenine dinucleotide
NaCl	Sodium chloride
NAD	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide dehydrogenase
NADP	Nicotinamide adenine dinucleotide phosphate
Nam	Nicotinamide
NaMN	Nicotinic acid mononucleotide
NaOH	Sodium hydroxide
NCBI	National Centre for Bioinformatics
NEB	New England Biolabs® Inc
NMN	Nicotinate mononucleotide
NR	Nicotinamide riboside
µg	Microgram
µg/ml	Microgram per millilitre
µl	Microlitre
µM	Micromolar
µm	Micron meter
ORF	Open reading frame
PCR	Polymerase chain reaction
Φ	Phage
pI	Isoelectric point
PTP	Picotitre plates
Qa	Quinolinic acid
rDNA	Ribosomal deoxyribonucleic acid
RNam	Ribosyl nicotinamide
sec	Seconds

TAE	2-Amino-2-(hydroxymethyl)- 1,3-propandiol, ethylene diamine tetraacetic acid, glacial acetic acid
Tris-HCl	2-Amino-2-(hydroxymethyl)-1,3-propandiol, hydrochloric acid
UV	Ultraviolet
V	Volts
V/cm	Volts per centimeter
v/v	Volume per volume
w/v	Weight per volume
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside

List of Figures

Chapter 1 - Introduction to present study

Chapter 2 - Literature review

Chapter 3 - Identification of putative genes involved in NAD^+ pathways within *Avibacterium paragallinarum*

Chapter 4 - Whole genome sequencing of *Avibacterium paragallinarum*

Chapter 5 – Genetic diversity through integration motifs within *Avibacterium paragallinarum*

List of Tables

Chapter 1
Introduction to present study

Avibacterium paragallinarum is the causative agent of infectious coryza (Kume *et al.*, 1978). This bacterium was formerly known as *Haemophilus paragallinarum*, but in 2005 it was reclassified as *Avibacterium paragallinarum* alongside *Pasteurella gallinarum*, *Pasteurella avium* and *Pasteurella volantium* as *Avibacterium gallinarum* gen. nov., comb. nov., *Avibacterium paragallinarum* comb. nov., *Avibacterium avium* comb. nov. and *Avibacterium volantium* comb. nov. The reclassification of this closely related group was based on the results of 16S rDNA sequencing (Blackall *et al.*, 2005). In order to avoid confusion current terminology will be used throughout the text alongside the unique abbreviation of *Av. paragallinarum*.

Infectious coryza (IC) causes vast economical losses for chicken farmers not only in Southern Africa but also in many other parts of the world. This disease is of economic importance wherever chickens are raised since chicken farmers experience an increased culling rate in broiler chickens as well as a 10-40% reduction in egg production in laying as well as breeding hens (Zhang *et al.*, 2003).

Since the late seventies there have been reports of vaccination failures as chickens display clinical signs of IC even though they were recently vaccinated (Kume *et al.*, 1978). Furthermore, during the eighties it was established that poultry flocks were no longer adequately protected through available vaccines due to the alteration in the incidence of the different serovars (Bragg *et al.*, 1996). In addition, cross-protection is not a precise occurrence in *Av. paragallinarum* and thus playing an important role in vaccination failures (Blackall, 1995).

Nicotinamide adenine dinucleotide ($NAD^{+)}$ independent strains of *Av. paragallinarum* have been reported since the early nineties (Mouahid *et al.*, 1992). Almost instantly this capability was linked to a native plasmid from *Av. paragallinarum* through the use of crude plasmid extraction followed by transformation of NAD^+-dependent strains. However, up to date no sequencing confirmation of a plasmid obtained from NAD^+-independent strains has confirmed that this trait is plasmid borne (Bragg *et al.*, 1993).

Blackall (1988) suggested that *Av. paragallinarum* is unlikely to contain any native plasmids. However, in 2003 Terry and colleagues isolated and analysed a plasmid, labelled as p250, from *Av. paragallinarum* which encoded genes involved in haemocin production. This haemocin operon produces a protein which is capable of killing a range of other gram-negative bacteria. Analysis of the complete sequence of this plasmid did not reveal any putative gene(s) responsible for NAD^+-independence.

The initial part of this study was focussed on the elucidation of the NAD^+-independence phenomena within *Av. paragallinarum*. In a M.Sc study by van Zyl (2003) it was proposed that a plasmid entity is responsible for NAD^+-independence in this bacterium. This claim was further investigated by attempting to re-isolate this plasmid. As reported by numerous authors plasmid isolations from this bacterium are problematic. Furthermore, this part of the study also indicated that this bacterium has the capability to integrate native plasmid(s).

Numerous factors, including the above mentioned, served as the main driving force behind the whole-genome sequencing project of *Av. paragallinarum* in order to provide the necessary tools to comprehend virulence and pathogenicity within this bacterium.

The purpose of this study was:

i. to prove the incorporation of plasmid p250 within the genome of *Av. Paragallinarum*;
ii. to understand NAD^+ pathways within *Av. Paragallinarum*;
iii. and to report on the on-going whole-genome sequencing project of *Av. paragallinarum*.

The next chapter will provide a comprehensive background regarding integrative elements, NAD^+ pathways as set out for bacteria in general alongside those indicated for the *Pasteurellaceae* and finally pyrosequencing in order to provide background for the research conducted in this study.

Literature cited

Blackall, P. J. 1995. Vaccines against Infectious Coryza. *World Poultry Sci. J.* **51**:17-26.

Blackall, P. J. 1988. Antimicrobial Drug Resistance and the Occurrence of Plasmids in *Haemophilus paragallinarum*. *Avian Dis.* **32**:742-747.

Bragg, R. R., L. Coetzee and J. A. Verschoor. 1996. Changes in the incidences of the different serovars of *Haemophilus paragallinarum* in South Africa: A possible explanation for vaccination failures. *Onderstepoort J. Vet. Res.* **63**:217-226.

Bragg, R. R., L. Coetzee and J. A. Verschoor. 1993. Plasmid-encoded NAD^+ independence in some South African isolates of *Haemophilus paragallinarum*. *Onderstepoort J. Vet. Res.* **60**:147-152.

Kume, K., A. Sawata and Y. Nakase. 1978. *Haemophilus* Infections in chickens. 1. Characterization of *Haemophilus paragallinarum* isolated from chickens affected with coryza. *Jpn. J. Vet. Sci.* **40**:65-75.

Mouahid, M., M. Bisgaard, A. J. Morley, R. Mutters and W. Mannheim. 1992. Occurrence of V-factor (NAD^+) independent strains of *Haemophilus paragallinarum*. *Vet. Microbiol.* **31**:363-368.

Terry, T. D., Y. M. Zalucki, S. L. Walsh, P. J. Blackall and M. P. Jennings. 2003. Genetic analysis of a plasmid encoding haemocin production in *Haemophilus paragallinarum*. *Microbiology.* **149**:3177-3184.

Van Zyl, A. E. 2003. Characterization of a plasmid conferring NAD^+ independence in *Haemophilus paragallinarum*. M.Sc. thesis. University of the Free State, Bloemfontein, South Africa.

Zhang, P. J., M. Miao, H. Sun, Y. Gong and P. J. Blackall. 2003. Infectious Coryza due to *Haemophilus paragallinarum* serovar B in China. *Aust. Vet. J.* **81**:96-97.

Chapter 2
Literature review

2.1. Infectious Coryza

Infectious coryza (IC) is an upper respiratory tract disease of chickens and is of economical importance wherever chickens are raised (Blackall *et al.*, 1990a). This disease is of great economic importance because it causes a decrease in egg production and an increase in slow-growing chickens (Chen *et al.*, 1996).

IC is generally presented in an acute form but it may be portrayed as a chronic disease of chickens (Eaves *et al.*, 1989). Unusual clinical signs have been reported but the most common signs are nasal discharge, facial oedema, lacrimation and diarrhoea. Growth retardation in young stock and a reduction in egg production occur due to a decreased feed and water consumption (Blackall, 1999). IC can be transmitted by means of contaminated drinking water, is airborne over short distances and it has been suggested that replacement stock is a major source of infection (Blackall *et al.*, 1990a). It is also known that this disease spreads rapidly within a flock (Kume *et al.*, 1983).

In India IC is the second most important bacterial disease associated with mortality of chickens and in Thailand it was the most common cause of death in chickens younger than two months and in layers older than six months (Blackall, 1999). In South Africa it is regarded as one of the most serious diseases of layers (Buys, 1982) and vaccination failures have been reported as early as the mid 1980's (Bragg, 2005).

2.2. *Avibacterium paragallinarum*

Avibacterium paragallinarum is a micro-aerophillic coccobacilli gram-negative bacterium and grows in a 5-10% CO_2 concentration (Kume *et al.*, 1978). This organism is slow-growing and fastidious in nature (Blackall, 1999) and in 24 h cultures it appears as short rods 1-3 µm in length and 0.4-0.8 µm in width with a tendency for filament formation. This bacterium undergoes degeneration within 48-60 h presenting fragments and indefinite forms. It is non-motile, catalase negative and some strains are encapsulated (Yamamoto, 1984). *Avibacterium paragallinarum* was formerly known as *Haemophilus paragallinarum* but based on 16S rDNA sequencing it was classified within a separate genus: *Avibacterium* gen. nov., along with *Pasteurella gallinarum*, *Pasteurella avium* and *Pasteurella volantium* (Blackall *et al.*, 2005).

2.2.1. Serological classification.

The main classification system of *Av. paragallinarum* is serologically based; the approach varies with the serovars and the intended use which may either be to detect infection or to test vaccination responses (Blackall, 1999). Two classification systems that are applied to *Av. paragallinarum* are the Page and the Kume classification.

Page first performed serological differentiation of *Av. paragallinarum* in 1962. This was done through the use of the plate agglutination method and three different serotypes were detected which were termed A, B and C (Page, 1962). Kume and co-workers (1983) on the other hand, based a scheme on haemagglutination antigens through which three different serogroups, consisting of seven different serovars were detected. The serogroups were designated I, II and III and the different serovars were termed HA-1 to HA-7. Serogroup I consisted of serovars HA-1 to HA-3, serogroup II contained

serovars HA-4 to HA-6 and serovar HA-7 belonged to serogroup III (Kume *et al.*, 1983). It is interesting to note that Kume's serogroups I, II and III correspond to Page serovars A, C and B respectively. This subsequently led towards the proposal to alter the Kume scheme nomenclature concluding with the nine currently recognized serovars (Table 1) (Blackall *et al.*, 1990b).

Table 2.1. Nine currently recognized serovars of *Avibacterium paragallinarum*.

Currently Recognized Serovars	Kume	Page
A-1	I; HA-1	A
A-2	I; HA-2	A
A-3	I; HA-3	A
A-4	I; HA-8*	*
B-1	III; HA-7	B
C-1	II; HA-4	C
C-2	II; HA-5	C
C-3	II; HA-6	C
C-4	II; HA-9**	**

* Eaves identified a new serovar, HA-8, within Kume serogroup I in 1989 (Eaves *et al.*, 1989).

** Blackall further identified a new serovar, HA-9, within Kume serogroup II in 1990 (Blackall *et al.*, 1990b).

2.2.2. Innate plasmids of *Avibacterium paragallinarum*.

Native plasmids of *Av. paragallinarum* have eluded the research world for just over a decade. In 1988 a total of 75 *Av. paragallinarum* isolates were screened for plasmids in order to obtain the source of occurring streptomycin resistance. Five different extraction techniques were applied and no plasmids were obtained. It was suggested that this resistance may not be plasmid mediated and that it might be as a result of reduced antibiotic uptake, but more importantly it was suggested that a plasmid may integrate within the

genome of *Av. paragallinarum* as was seen with *Haemophilus influenzae* (Blackall, 1988).

Av. paragallinarum was always considered to be dependent on NAD^+ but as early as 1932 NAD^+-independent bacterial isolates have been identified that were thought to be *Av. paragallinarum* in all aspects except for their observed NAD^+-independence (McGaughy, 1932 as cited by Bragg *et al.*, 1993). It took more or less fifty years to confirm that NAD^+-independence was indeed a trait belonging to *Av. paragallinarum*. This trait was attributed to a plasmid when Bragg and colleagues transformed NAD^+-dependent *Av. paragallinarum* strains with crude plasmid extract to produce NAD^+ independent *Av. paragallinarum* strains (Bragg *et al.*, 1993).

In 2003 the first native plasmid was described for *Av. paragallinarum* (Figure 2.1) (Terry *et al.*, 2003). This plasmid conferred haemocin production, processing, export and immunity. Haemocin is capable of killing a range of other gram-negative bacteria. Alongside the haemocin operon the plasmid includes a replication gene (*repB*) and a putative integrase gene (*Int*). The haemocin locus was also detected on the chromosome of some strains suggesting that at some stage the p250 plasmid has either been excised from the chromosome or integrated within it (Terry *et al.*, 2003).

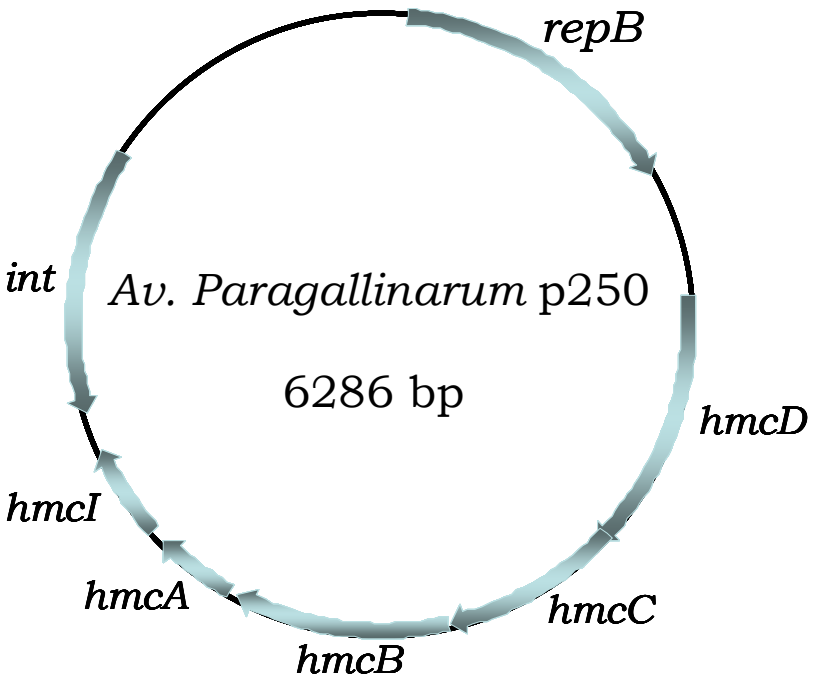


Figure 2.1: *Avibacterium paragallinarum* p250 plasmid. This figure illustrates the haemocin operon (*hmc*), the putative integrase (*int*) and the putative replication (*repB*) genes respectively (Terry *et al.*, 2003).

Hsu *et al.* (2007), described a plasmid, designated pYMH5, isolated from *Av. paragallinarum* which allowed streptomycin, sulfonamide, kanamycin and neomycin resistance (Figure 2.2). This was the first multi-drug resistance plasmid that was reported for *Av. paragallinarum*. This plasmid showed high identity to a broad-range host vector isolated from *Haemophilus ducreyi*. The pYMH5 plasmid was transformed within *Escherichia coli* DH5α and the streptomycin-, sulfonamide-, neomycin-, and kanamycin-resistance genes contained within it were shown to be all functional within *E. coli* (Hsu *et al.*, 2007).

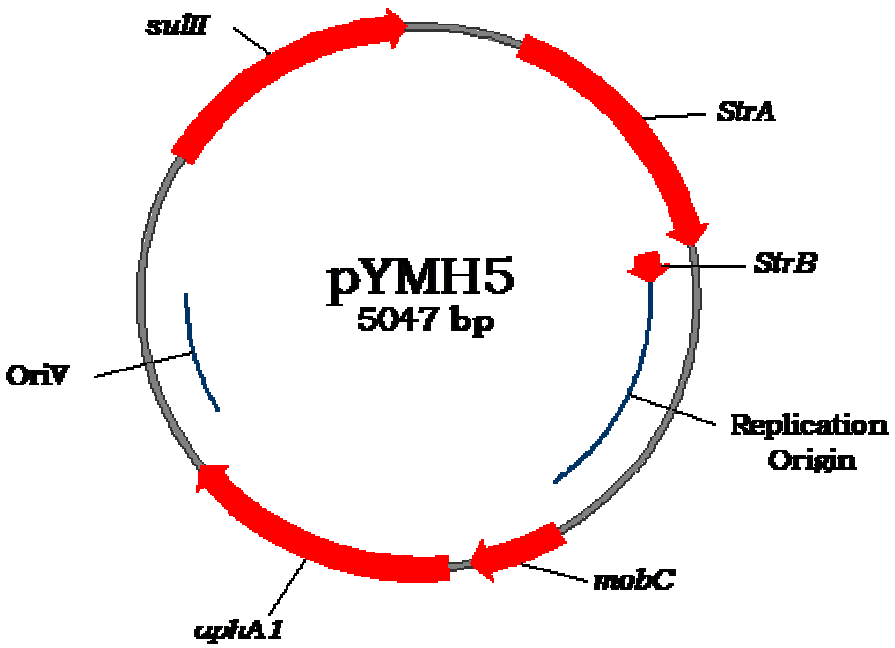


Figure 2.2: *Avibacterium paragallinarum* pYMH5 multi-drug resistance plasmid. This plasmid contains four open reading frames with most coding for antibiotic resistance. *StrA* encodes an aminoglycoside phosphotransferase as well as *StrB* (incomplete); *mobC* encodes a mobilization protein and *aphA1* encodes an aminoglycoside (3') phosphotransferase (Hsu *et al.*, 2007).

Alongside pYMH5, Hsu and colleagues (2007) described a partial *Av. paragallinarum* plasmid pA14. Two DNA fragments were sequenced namely a 1 kb-fragment which contained a putative truncated MglA protein (encoding an ABC-type monosaccharide transport system) and a 3 kb-fragment which encoded a putative RNase II protein.

Hsu and co-workers (2007) used 20 individual *Av. paragallinarum* strains (Page serovars A and C only) in their study. Of these 20 strains, seven strains contained no plasmids, 12 strains contained two plasmids and only one strain contained one plasmid. In accordance with what Terry and co-workers stated, four of the seven strains which did not contain any plasmids did however show haemocin activity which is likely encoded by chromosomally located genes (Hsu *et al.*, 2007; Terry *et al.*, 2003).

2.3. Integrative elements

Antibiotics have been in clinical use for more than the past six decades and there have been a significant appearance and distribution of antibiotic resistance amongst bacteria in the last fifteen years (Mazel, 2006; Bennet, 2008). It seems that bacteria have long been preparing for the antibiotic age through the acquisition of appropriate genetic tools, including transposition and site-specific recombination to quickly and effectively overcome the challenge at hand (Mazel, 2006).

2.3.1. Site-specific recombination

Site-specific recombination of the integrase type plays a fundamental role in: the life cycles of temperate bacteriophages, bacteria and yeast; the integration and the excision of phage genomes into and from host chromosomes; the stable partitioning of plasmids, phage or bacterial genomes and the copy number control of yeast plasmids via replicative amplification (Voziyanov *et al.*, 1999). The tyrosine family of recombinases encode proteins which use a conserved catalytic tyrosine residue to break and covalently attach to their target sequence at a specific phosphodiester bond. Two inverted binding sites are contained within the target sequence which binds a molecule of the recombinase. Cleavage can take place in a dimeric structure consisting of two recombinase molecules bound to one target sequence, but recombination mostly takes place in a synaptic structure consisting of two target sites each bound by two molecules of recombinase (Shaikh and Sadowski, 2000).

2.3.1.1. Integrons

Integrons are assembly platforms capable of incorporating exogenous open reading frames *via* site-specific recombination and converting them into functional genes thus ensuring their correct expression. In effect they are natural cloning and expression systems that incorporate open reading frames and convert them into functional genes (Rowe-Magnus and Mazel, 2001). Figure 2.3 illustrates the three key elements integrons are composed of; (1) an integrase encoding gene belonging to the tyrosine-recombinase family (*IntI*); (2) a primary recombination/attachment target (*att*-site), which is specific to the integrase (*attI*) and (3) a strong promoter transcribing the integrase (Mazel, 2006; Birch, 2006).

All integron-inserted gene cassettes, which are one or more promoter less genes, share specific structural characteristics and generally contain the gene(s) and an imperfect repeat at the 3' end of the gene called the *attC*/59-base element. The *attC* sites are a diverse family of nucleotides that vary from 57 to 141 bp in length and they function as recognition sites for the site-specific integrase. The sequence similarities of these sites are generally restricted to their borders where the one side is termed the R" sequence (RYYYAAC) and the other the R' sequence (G*TTRRRY). Y indicates pyrimidines and R purines, recombination occurs between the G and the T bases (indicated by the asterisks) in the R' sequence (Mazel, 2006; Birch, 2006). All integrons in turn can be divided into mobile integrons (resistance integrons) or superintegrons.

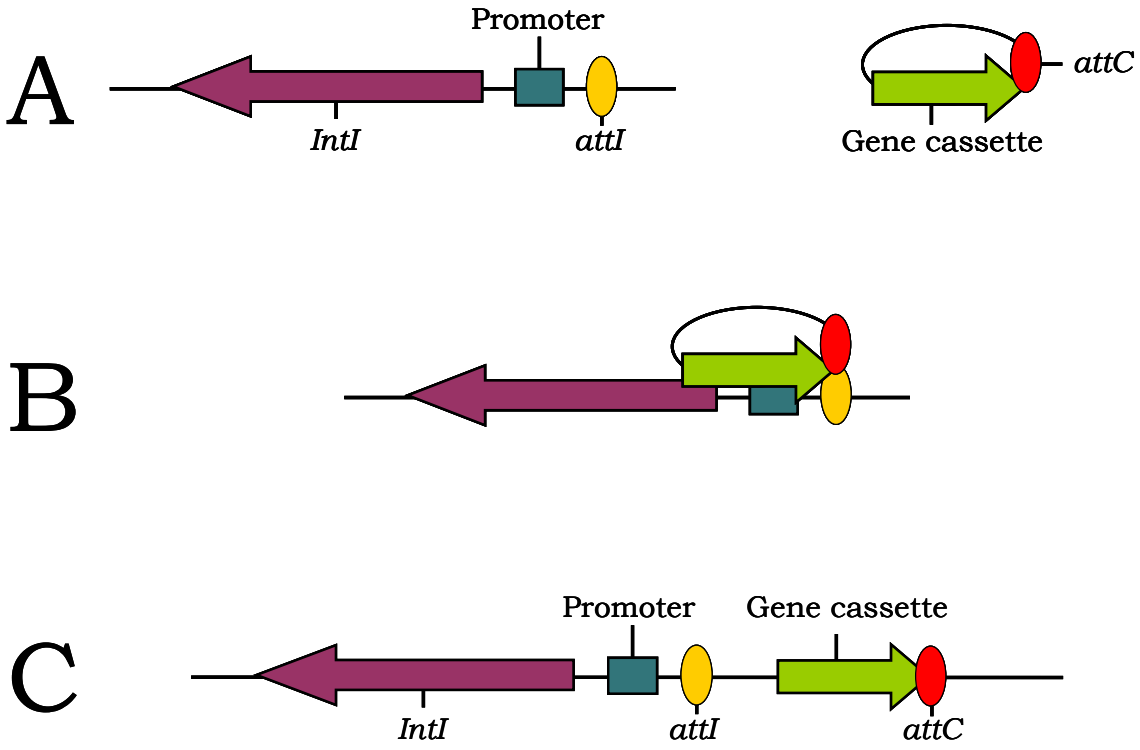


Figure 2.3. The functional integron platform. This diagram illustrates in A the integron platform. The gene that encodes an integrase, *Int*, of the tyrosine-recombinase family and the primary recombination sequence, *attI*, can be seen and alongside that the promoterless gene cassette. In B the integrase mediates recombination between the *attI*-site and a secondary target called the *attC*-site/59 base element. Insertion of the gene cassette takes place at the *attI*-site downstream of a resident promoter internal to the *intI* gene, which will drive expression of the encoded proteins. The functional integron platform, C, are defective for self-transposition but are often found associated with transposons and/or conjugative plasmids that can serve as vehicles for inter- and intra-species transmission of genetic material (Adapted from: Rowe-Magnus and Mazel, 2001).

There are approximately a hundred distinct integron classes known of which five classes are notorious for the distribution of antibiotic-resistance genes (Tetu and Holmes, 2008). These are all physically linked to mobile DNA fragments such as insertion sequences, transposons and conjugative plasmids. Insertion sequences are small, phenotypically cryptic segments of DNA that have a simple organization and are capable of insertion at multiple

sites in a target DNA molecule and are usually less than 2.5 kb in size. A transposon is a mobile DNA element that relocates within the genome of its host and a conjugative plasmid is a plasmid that relocates from one cell to another during conjugation (Mazel, 2006). The *IntI* integrases are not similar to other members of the tyrosine-recombinase family seeing that they can also recombine nucleotide sequences of low similarity. Through the use of integrons, bacteria can accumulate various exogenous genes to establish a wide variety of antimicrobial-resistance. Various integrons harbouring up to eight different resistance cassettes have been characterized. Mobile integron resistance cassettes have been identified in diverse gram-negative species but they are not only restricted to gram-negative bacteria since they have also been found in gram-positive bacteria including the staphylococci; corynebacteria; aerococci and brevibacteria (Mazel, 2006).

The key features that identify and distinguish superintegrons include a specific integrase related to those of mobile integrons, a large number of gene cassettes that are associated with the integron, a high degree of identity which is observed between the *attC* sites of these cassettes and finally, the structure does not seem to be mobile as the integron is located on the chromosome and is not associated with mobile DNA elements. A number of integrons have been characterized in various bacterial species that do not share all of the characteristics of a typical superintegron (Mazel, 2006).

Integrases, encoded by integrons, mediate recombination involving two sites namely the specific *attI* site and the associated gene cassette *attC* site. Five recombination reactions have been documented namely: *attI* × *attC*; *attC* × *attC*; *attI* × *attI*; *attI* × non-specific GTT-containing sequences and *attC* × non-specific GTT-containing sequences. Recombination events between an *attI*-site and an *attC*-site is the most efficient recombination event followed by recombination between two *attC*-sites and the least efficient recombination

event involves two *attI*-sites. Superintegron *attC*-sites are highly homogeneous and species-specific whereas those in mobile integrons are highly variable in length and sequence composition (Mazel, 2006).

2.3.1.2. Phage recombination

Phage integrases carry out recombination between the attachment site on the phage, *attP*, and the attachment site on the bacterial genome, *attB* (Groth *et al.*, 2000). The best studied integrase of the tyrosine recombinases family is the λ Int protein which promotes integration and excision of the phage genome from that of the host (Esposito and Scocca, 1997). The Cre-recombinase system of the P1 bacteriophage efficiently catalyzes recombination between two of its consensus DNA recognition sites (*loxP*-sites) in any cellular environment and any DNA source (Nagy, 2000). Some phages encode integrases belonging to the serine family of site-specific recombinases with a sought after trait of irreversible integration. Thus, once integration took place the transgene cannot be excised by the recombinase, additional factors are required for the reverse reaction (Groth *et al.*, 2000).

2.3.2. Transposable elements

Transposable elements are discrete DNA segments that incorporate into non-homologous DNA during reactions catalyzed by transposon-encoded proteins called transposases. The mechanism includes end cleavage followed by DNA strand transfer resulting with either DNA replication or DNA repair. Transposable elements promote a variety of DNA rearrangements as well as simple insertions. Intra-molecular transposition results in either the deletion or the inversion of adjoining DNA segments (Bernales *et al.*, 1999).

2.3.2.1. Transposons

Essentially, transposons are jumping gene systems which incorporate a gene of interest within the transposon element. They are present in a variety of forms that are recognized through their structure, genetic relatedness and mechanism of transposition. These elements can carry a variety of genes and they have the ability to move both inter- and intra-molecularly. In effect they can relocate from one site in a DNA molecule to another site or from one DNA molecule to another DNA molecule. The mechanisms by which these elements relocate generally do not require any DNA homology between the element and the insertion site. While some elements tend to show a preference for a particular nucleotide sequence most of them insert into new sites more or less at random. Transposons differ from insertion sequence elements in that it encodes at least one function that alters the phenotypic appearance of the cell whereas insertion sequences do not necessarily alter the phenotypic appearance of the cell. Resistance transposons confer resistance traits thereby altering the phenotypic appearance of the cell (Bennet, 2008).

2.3.2.2. Insertion sequence elements

Insertion sequences (IS) are the smallest autonomously mobile elements and are generally smaller than 2.5 kb in size. A clear cut definition does not exist and the description of IS remains broad but the features that group them together include their simple organization, the ability to insert within the target DNA at multiple sites and they are phenotypically cryptic in nature (Rousseau *et al.*, 2004; Mahillon and Chandler, 1998).

IS elements act preferentially in a *cis* action and in some cases this is due to the instability of the transposase, while in other cases it depends on the abundance of the transposase or on the association of its N-terminal sequence with DNA (Bernales *et al.*, 1999). Recently it has been shown that most

genomes possess hundreds or thousands of repeated elements capable of recombining through homologous (the IS element and the target DNA share an identical series of nucleotides) or illegitimate (the IS element and the target DNA share only a few identical nucleotides) recombination. These repeats seem to be part of regulatory elements, duplicated genes or IS. These are constantly created through recombination, horizontal transfer or transposition, deletions through recombination or through the accumulation of point mutations (Rocha, 2004). Repetitive sequences within a genome serve as hot-spots for recombination because they provide points that permit secure recombination and transposition without severe damaging effects. Additionally they vary in position and number, thereby supplying added variability (Tobes and Pareja, 2006).

2.4. NAD^+ utilization and production

Nicotinamide adenine dinucleotide (NAD^+; Figure 2.4) plays an important role in a variety of catabolic and anabolic reactions and it is of considerable importance in cellular metabolism (Foster and Moat, 1980). NAD^+ serves as co-factor for countless enzymatically catalyzed redox reactions (Rodionov *et al.*, 2008a) functioning as hydride acceptors or donors (Begley *et al.*, 2001). NAD^+ shuttle between its oxidized form, NAD^+, and the reduced form, NADH, but the total concentration remains constant within the cells (Lin, 2007). NAD^+ contains two relatively high energy bonds namely the N-glycosidic bond, involving nicotinamide, and the pyrophosphate bond (Lin, 2007). The biological half-life of NAD^+ in aerobically grown bacterial cells is approximately 90 minutes and NAD^+ turnover and pyridine cycles have been recognized (Merdanovic *et al.*, 2005). NAD^+ is produced either through biosynthetic or recycling pathways and different combinations of these metabolic routes result in a widespread mixture of the NAD^+ biosynthetic machinery in various species (Rodionov *et al.*, 2008b).

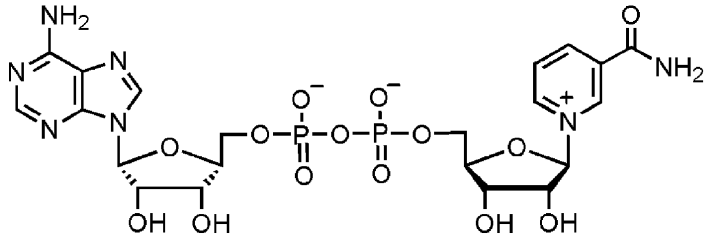


Figure 2.4. The chemical structure of Nicotinamide adenine dinucleotide (NAD^+) (Lin, 2007).

2.4.1. Universal NAD^+ recovery pathways

Many pathogenic bacteria are unable to produce pyridines, the heterocyclic aromatic organic compound with the formula C_5H_5N, and therefore are entirely dependent on recovering these from their hosts (Begley *et al.*, 2001; Rodionov *et al.*, 2008b). There are three generally regarded recovery pathways for NAD^+ recycling and some features repeat within these pathways e.g. enzymes, precursors or intermediates (Rodionov *et al.*, 2008b).

In the NAD^+ recovery pathway I, Figure 2.5, nicotinic acid (NA) and nicotinamide (Nam) is taken up by the niacin transporter, niaP from there the Nam is converted to NA through the nicotinamide deaminase enzyme, pncA. Independently from the previous conversion nicotinate phosphoribosyltransferase pncB converts NA to nicotinic acid mononucleotide (NaMN). Universal biosynthesis enzymes nicotinate mononucleotide adenylyltransferase, nadD and NAD^+ synthetase, nadE sequentially converts NaMN to nicotinate adenine dinucleotide (NaAD) and then to NAD^+ (Rodionov *et al.*, 2008b).

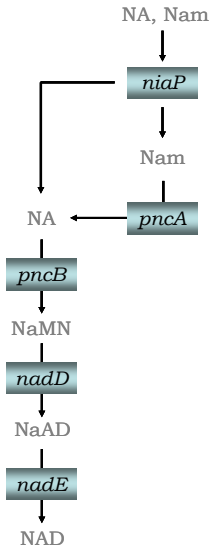


Figure 2.5. NAD^+ recovery pathway I. Nicotinic acid (NA) and nicotinamide (Nam) is transported by niaP to pncA and pncB where pncA converts Nam to NA and pncB converts NA to nicotinic acid mononucleotide (NaMN). NaMN is then converted to nicotinate adenine dinucleotide (NaAD) through nadD and then converted to NAD^+ by nadE.*

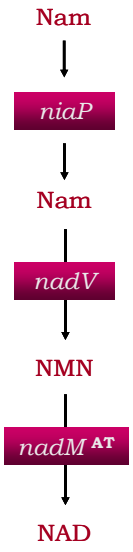


Figure 2.6. NAD^+ recovery pathway II. Nam is transported by niaP to nicotinamide phosphoribosyltransferase (nadV) which converts it to NMN. The final conversion within this pathway is when $nadM^{AT}$ converts NMN to NAD^+.*

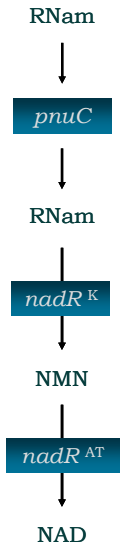


Figure 2.7. NAD^+ recovery pathway III. Exogenous ribosyl nicotinamide (RNam) is transported to nadR through RNam transporter pnuC and then converted to NAD^+.*

* Illustrations adapted from Rodionov *et al.* 2008b. The blocks within the above figures represent enzymes present within the various pathways, arrows going through the blocks indicate metabolic enzymes which carry out conversions, and arrows that do not cross enzymes indicate enzymes that only transport precursors.

Recovery pathway II is illustrated in Figure 2.6. This pathway shares niaP with recovery pathway I, but in this case it only transports Nam. Nam is then converted to NMN through nadV (nicotinamide phosphoribosyltransferase). The AT-domain of nicotinamide mononucleotide adenylyltransferase ($nadM^{AT}$) converts NMN to NAD^+ (Rodionov *et al.*, 2008b).

During recovery pathway III (Figure 2.7) exogenous ribosyl nicotinamide (RNam) precursor is delivered by the RNam transporter pnuC. This is followed by two successive reactions catalyzed by two distinct domains of nicotinamide mononucleotide adenylyltransferase ($nadR^{AT}$) and nicotinamide riboside kinase ($nadR^{K}$) (Rodionov *et al.*, 2008b).

Figure 2.8. An additional phosphorylation reaction of NAD^+. This is performed by the enzyme NAD^+ kinase (nadF) which converts NAD^+ to NADP (Rodionov *et al.*, 2008b).

Figure 2.8 illustrates how NAD^+ is converted by the NAD^+ kinase (nadF) to nicotinamide adenine dinucleotide phosphate (NADP) (Gerlach and Reidl, 2006). NADP alongside NAD^+ is an important role player in a variety of catabolic and anabolic reactions and it is of considerable importance in cellular metabolism (Foster and Moat, 1980).

2.4.2. *De novo* synthesis of NAD^+

The biosynthesis of NAD^+ in bacteria follows a series of reactions (Figure 2.9) that include the oxidation of aspartic acid (Asp) to iminosuccinic acid (IA) followed by a condensation reaction with dihydroxyacetone phosphate producing quinolinic acid (Qa). Qa in turn is phosphorylated and

decarboxylated resulting in NaMN. NaAD is formed by the adenylation of NaMN which is followed by amide formation completing the production of NAD^+ (Begley *et al.*, 2001; Rodionov *et al.*, 2008b).

There is a variety of enzymes responsible for the biosynthetic production of NAD^+. Aspartate oxydase (nadB) utilizes loosely adjoining flavin adenine dinucleotide (FAD) as co-factor and employs oxygen or fumarate as co-substrates converting Asp to IA (Begley *et al.*, 2001; Rodionov *et al.*, 2008b). This gene has been cloned from numerous organisms namely *Escherichia coli*, *Bacillus subtilis*, *Salmonella* and *Pseudomonas aeruginosa* but only *E. coli nadB* have been over-expressed (Begley *et al.*, 2001).

Quinolinate synthase (nadA) is responsible for the condensation of IA and dihydroxyacetonephosphate to Qa (Begley *et al.*, 2001; Rodionov *et al.*, 2008b). This enzyme has been cloned and over-expressed from both *E. coli* and *Salmonella* (Begley *et al.*, 2001; Rodionov *et al.*, 2008b).

Quinolinic acid phosphoribosyltransferase (nadC) catalyzes the conversion of Qa to NaMN (Begley *et al.*, 2001; Rodionov *et al.*, 2008b). This enzyme has both been cloned and over-expressed in *E. coli* and *Salmonella*. The most comprehensive structural studies have been carried out on the enzyme originating from *Mycobacterium tuberculosis*. Structures from different steps during the conversion of Qa to NaMN have been reported (Begley *et al.*, 2001).

Figure 2.9. The biosynthesis pathway for NAD^+ synthesis. Asp is converted to IA through the enzyme nadB. Thereafter the enzyme nadA converts IA to Qa. Qa is converted to NaMN through the enzyme nadC. The enzyme nadD converts NaMN to NaAD, and the final conversion occur through the enzyme nadE which converts NaAD to NAD^+ (Rodionov *et al.*, 2008b).

Nicotinate mononucleotide adenylyltransferase (nadD) is accountable for the alteration of NaMN to NaAD (Begley *et al.*, 2001; Rodionov *et al.*, 2008b). This enzyme is also capable of catalyzing the adenylation of NMN. Orthologs of this enzyme have been cloned, over-expressed, purified and kinetically characterized from several bacterial species namely, *E. coli*; *Helicobacter pylori*; *Staphylococcus aureus*; *Fusobacterium nucleatum* and *Synechocystis* sp. All these enzymes showed a distinct preference for NaMN over NMN (Begley *et al.*, 2001).

The decisive enzyme in this pathway is the NAD^+ synthetase (nadE). This enzyme is responsible for NAD^+ production from NaAD. The NAD^+ synthetase

from both *E. coli* and *B. subtilis* have been cloned and over expressed (Begley *et al.*, 2001; Rodionov *et al.*, 2008b).

2.4.3. Common themes amongst the NAD^+ pathways

Some features repeat within the NAD^+ recovery pathways that may either be enzymes, precursors or intermediates. The enzymes nadD and nadE are shared between the *de novo* biosynthesis pathway and the recovery pathway I. The precursor Nam serves as a precursor for both the recovery pathway I and II. Recovery pathway II and III share both the same precursor, NMN, and the same conversion reaction even though it is catalyzed by two different enzymes namely: $nadM^{AT}$ and $nadR^{AT}$ (Rodionov *et al.*, 2008b). The enzymes pncB and nadC from the NAD^+ recovery pathway I and the NAD^+ biosynthesis pathway respectively catalyze a very similar chemical alteration that lead to the formation of NaMN. The specificity of both these enzymes for their substrates is very high, but the only common theme between these enzymes is the α/β barrel fold alongside catalytic residues and mechanistic features. A prominent attribute of pncB that differentiates it from all other phosphoribosyltransferase, including nadC, is that this enzyme is stimulated by ATP (Adenosine Tri-Phosphate) (Begley *et al.*, 2001).

2.4.4. The regulation of NAD^+ synthesis

The biosynthesis pathway is regulated by a combination of feedback control and regulated gene expression of the contributing enzymes. A tri-functional repressor protein, nadR, forms a complex with NAD^+ and ATP. This complex binds to the NAD^+ box sequence and it either represses or allows the expression of *nadA* and *nadB*. This repressor protein also plays a critical role in the utilization of NAD^+ precursors from the growth medium (Begley *et al.*, 2001). Under non-starvation conditions nadR is bound to its co-repressor,

NAD^+ and this enables the complex to bind to DNA and repress transcription of *nadA* and *nadB*. In turn, in the presence of a decreased concentration of NAD^+, nadR associates with ATP and this complex does not bind DNA and as a result, transcription can take place. The recovery pathways are regulated through co-factor degradation and the tri-functional repressor protein, nadR. Here nadR regulates the expression of *pncB* and it controls the uptake and adenylation of NMN (Begley *et al.*, 2001). Again under non-starvation conditions nadR is bound with its co-repressor, NAD^+ and this enables the complex to bind to DNA and this time repress transcription of *pncB*. The same occurs when nadR and ATP associates and transcription repression is lifted (Merdanovic *et al.*, 2005).

2.4.5. NAD^+ pathways within the *Pasteurellaceae*

Certain members of the *Pasteurellaceae* require exogenous NAD^+ for growth because they are unable to synthesize and recycle their own. Their basic NAD^+ pathway consists of an uptake system with negligible resynthesis activity and for that reason these organisms rely on extracellular NAD^+ supply (Gerlach and Reidl, 2006). NAD^+ requirement within the *Pasteurellaceae* is referred to as factor V and it is initiated from a lack of both *de novo* biosynthetic pathways for NAD^+ as well as most of the pyridine nucleotide cycle pathways. NAD^+ requirement can be satisfied by NAD^+, NMN and/or NR (Morton *et al.*, 2008). Nam may also serve as a substrate for certain members of the *Pasteurellaceae*, for example *Haemophilus aphrophilus*; *H. ducreyi*; *Actinobacillus pleuropneumoniae* serotypes and some *Pasteurella* spp., however *H. influenzae* cannot utilize NAm as substrate (Gerlach and Reidl, 2006).

Figure 2.10 demonstrates the NAD^+ pathways within the family *Pasteurellaceae*. NAD^+ dependence is maintained through the uptake of nicotinamide riboside (NR) by the enzyme pnuC (NR permease) and followed by

the conversion of NR to NMN by means of nadR, which subsequently converts NMN to NAD^+. This conversion of NMN to NAD^+ through nadR is shared by both the NAD^+ dependent and independent pathways. It seems that the only difference between these two pathways is the phosphoribosyl pyrophosphate transferase (nadV) which alters Nam to NMN (Gerlach and Reidl, 2006).

The inferred amino acid sequence of nadV shows significant similarity with putative gene products derived from genomes of *Actinobacillus actinomycetemcomitans* and *Pasteurella multocida*. For certain members of the *Pasteurellaceae* it was reported that NAD^+ independence was a transferable trait. This trait was found on a *Haemophilus ducreyi* plasmid, and tandem repeats of this plasmid are integrated within the genome of *H. ducreyi* strain 35000HP. This suggests that *nadV* might be carried within a chromosomally located putative phage element and might therefore be transmissible via horizontal transfer. Plasmid mediated NAD^+-independence was also observed for *Haemophilus parainfluenzae* and *Avibacterium paragallinarum*. The wild-type NAD^+ requirement of *H. influenzae* could be uplifted through transformation with plasmids harbouring the *nadV* gene (Gerlach and Reidl, 2006).

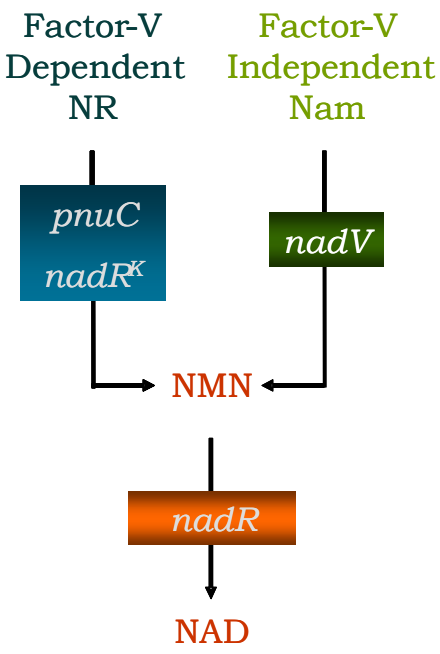


Figure 2.10. NAD^+ pathways within the *Pasteurellaceae*. Within this family the V-factor dependence route, in blue, consist of nicotinamide riboside (NR) uptake through pnuC, NR permease, followed by the conversion to NMN by nadR and subsequently to NAD^+ due to the same enzyme. On the other hand the V-factor independence pathway, in green, starts with the conversion of Nam to NMN through the enzyme nadV, phosphoribosyl pyrophosphate transferase, and thereafter converted to NAD^+ through the enzyme nadR. The phases in orange are shared by these two pathways. (Adapted from Gerlach and Reidl, 2006.)

2.4.6. The regulation of NAD^+ synthesis within the *Pasteurellaceae*

Nicotinamide riboside (NR) uptake within the V-factor dependent route is a regulated process because cell entry is carrier dependent as well as saturable. NadR is a chimeric protein that contains both regulatory and enzymatic activity. The admission of NR within a cell is regulated where NadR directly influence the transport activity of PnuC thereby stimulating NR uptake if cellular NAD^+ levels are low. NR uptake and NAD^+ synthesis is regulated through feedback inhibition. If NadR does not phosphorylate NR to form NMN, no NR would be imported through PnuC and in conjunction if cellular NAD^+ levels are not low enough, NadR will not phosphorylate NR to form NMN.

NadR does not function as a transcriptional regulator and there is no direct interaction between PnuC and NadR (Gerlach and Reidl, 2006).

It seems that the regulation of nadV is not yet well understood. Gerlach and Reidl (2006) proposed that the function of nadV is possibly regulated through feedback inhibition.

2.5. Next generation sequencing

The human genome sequencing project was the focus point for the development of high-throughput, high-capacity DNA sequences. Since the completion of this project, the approach towards genome sequencing has been refocused from bacterial artificial chromosomes (BAC's) to whole genome sequencing (Mardis, 2008).

The methodology of whole genome sequencing includes the shearing of genomic DNA into numerous individual size classes. Thereafter these stretches of genomic DNA are then located within plasmids and fosmids. Through oversampling the ends of these subclones, paired-end reads are created which in turn provide linking information for whole genome assembly algorithms (Mardis, 2008).

2.5.1. Next-Generation Sequencing Methods

Several instruments have been developed recently to achieve next-generation sequencing. These instruments allow highly streamlined sample preparation steps prior to DNA sequencing, saving significant amounts of time and requiring minimal linked equipment in comparison to the highly automated, multistep pipelines necessary for clone-based high-throughput sequencing. Each of these programmes represents a complex interaction of

enzymology, chemistry, high-resolution optics, hardware and software engineering. Each technology pursues, within its own right, the amplification of single strands of a fragment library to perform sequencing reactions on the amplified strands. The fragment libraries are obtained by annealing platform-specific linkers to blunt-ended fragments generated directly from a genome or source DNA of interest. Due to the presence of adapter sequences these molecules can then selectively be amplified by polymerase chain reaction (PCR) and as a result, no bacterial cloning steps are required to amplify the genomic fragment in a bacterial intermediate as is necessary in traditional sequencing approaches. The running time of these machinery is longer and the yields of sequencing reads (bases per instrument run) is significantly higher. Up-and-coming single molecule sequencers, Helicos Heliscope™ and Pacific Biosciences SMRT, do not require *any* amplification of DNA fragments prior to sequencing (Mardis, 2008).

2.5.1.1. Applied Biosystems SOLiD™ Sequencer

This platform uses an adapter-ligated fragment library and an emulsion PCR approach with small magnetic beads to amplify the fragments for sequencing. In order to initiate the initial sequencing cycle, all four labelled reversible terminators, primers and DNA polymerase enzyme are added within the flow cell. This is followed by laser excitation and the capture of the image of emitting fluorescence from each cluster on the flow cell, the identity of the first base for each cluster is identified. Thereafter the blocked 3' terminus and the fluorophore from each incorporated base are removed. This process is repeated to form cycles of sequencing in which the sequence of bases within a given fragment is determined a single base at a time (Mardis, 2008).

2.5.1.2. Illumina Genome Analyzer

This takes place on an oligo-derivatized surface of a flow cell. A flow cell is an eight channel sealed glass microfabricated device that allows bridge amplification of fragments on its surface. The same library can be used in each of the eight channels or different libraries adjacent to each other or combinations of libraries. DNA polymerase is used to produce multiple DNA copies that represent the single molecule that initiated the amplification. The DNA polymerase along with the four nucleotides is added simultaneously to the flow channels. The nucleotides carry a base-unique fluorescent label and the 3'-OH group is chemically blocked in such a way that each incorporation result in a unique event. Each base incorporation is followed by a imaging step during which each flow cell lane is imaged in three one hundred tile segments by the instrument optics at a cluster density per tile of 30 000. The 3' blocking group is chemically removed following each imaging step in order to prepare the strand for the next incorporation by DNA polymerase. These are continued to produce 20-35 distinct bases. A base-calling algorithm assigns sequences and associate quality values to each read and a quality checking pipeline evaluates the Illumina data from each run thereby removing poor-quality sequences (Mardis, 2008).

2.5.1.3. Roche/454 GS-20; FLX and Titanium Pyrosequencer

These systems make use of an alternative sequencing technology known as pyrosequencing. Pyrosequencing allows for the incorporation of a nucleotide by DNA polymerase which results in the release of pyrophosphate, which initiates a series of downstream reactions that ultimately produce light through the use of luciferase. The amount of light produced is proportional to the number of nucleotides incorporated. In short this process starts off with the nebulization of the genomic DNA where the DNA is sheared into fragments after which they are denatured into single stranded molecules. This is

followed by a process called emulsion PCR. Emulsion PCR occurs within water-in-oil microreactors. The single stranded fragments are attached to capture beads and in combination with PCR reagents they are emulsified within the water-in-oil microreactors. Clonal amplification takes place within each microreactor. Following clonal amplification the microreactors are broken and the amplified beads are packaged onto picotitre plates (PTP) in such a way that only one clonally amplified bead fit into each well. The PTP-plates are situated opposite a charge-couple device (CCD) camera that records the light emitted from each bead. Enzyme containing beads that catalyze the downstream reaction steps are then added to each well. In effect the PTP act as a flow cell into which each pure nucleotide solution is introduced in a stepwise fashion. Sequencing is performed simultaneously in all the wells. There is a continuous overflow of reagents containing buffers and nucleotides over the PTP. The first thing the system does is look for the key sequence **TACG** employed by the A and B adaptors during library preparation. If this sequence is not found and if it is not present in the specific order, the well will not be read. This sequence also performs the calibration of the base-calling software. The enzyme beads contain sulfurylase and luciferase. The nucleotides are provided individually by flowing them over the PTP. When a nucleotide is complimentary and incorporated a pyrophosphate molecule is released. The pyrophosphate molecule is converted to ATP by sulfurylase using adenosine phosphosulfate. Luciferase hydrolyzes the ATP using luciferon producing oxy–luciferon and light. The light is detected by the CCD camera attached at the base of the PTP.

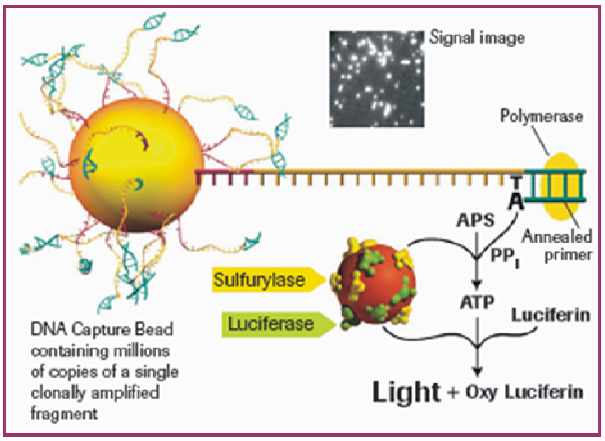


Figure 2.11. 454 Sequencing summary. Here each bead contains identical strands of clonally amplified DNA. When a complimentary nucleotide is incorporated a pyrophosphate molecule is released. This leads to the onset of various enzymatic reactions ending with the emission of light from each well which is recorded by the CCD camera attached to the base of the PTP (Roche Diagnostics, 2008).

The light intensity emitted from the PTP indicates incorporation and the intensity is directly proportional to the number of nucleotides incorporated, but calibrated base calling cannot interpret long stretches of the same nucleotide (a homopolymer run). Figure 2.11 illustrates the process taken place within each well of the PTP plate (Mardis, 2008; Roche Diagnostics 2008; Rothberg and Leamon, 2008).

2.5.2. Next-Generation Sequencing Analysis

A variety of software tools are available for analyzing next-generation sequencing data. This falls into certain categories including; alignment of sequence reads to a reference sequence, base-calling and/or polymorphism detection, *de novo* assembly and genome browsing and annotation. An increasing number of alignment tools have been developed specifically for the rapid alignment of large sets of short reads, while allowing for mismatches and/or gaps. Some of these tools take advantage of well-established alignment

algorithms but there also have been noteworthy advances in developing new algorithms specifically tailored for short reads. SOAP is an example of a software package for efficient gapped or ungapped alignment which uses a memory-intensive seed and look-up table algorithm to accelerate alignment while allowing iterative trimming of the 3' end of reads which usually have a higher error rate. Other approaches use accelerating processes which includes 'bit encoding' which in turn allows sequence data to be compressed into a more computationally manageably efficient format (Shendure and Ji, 2008).

Alignment software is increasingly taking into account the estimated quality of the underlying data in generating read-placements, as is the case with MAQ, an alignment and variation discovery tool that works with data from either Solexa or SOLiD, and SHRiMP which includes a novel "colour-space to letter-space" Smith-Waterman algorithm compatible with two base-encoded sequence data from the SOLiD platform (Shendure and Ji, 2008).

As with alignment algorithms relatively lower accuracies and quantity of data associated with the new technologies make *de novo* assembly into a challenging problem. Several assembly tools have recently been adapted or independently developed for generating assemblies from short reads, and several algorithms have been developed to take advantage of these (Shendure and Ji, 2008). Assemblers tuned for homogeneous sequence data is not suitable for hybrid data. Assembly tools like Celera assembler is modified software to assemble hybrid reads, Sanger-sequencing and pyrosequencing reads, from ABI 3730 and Roche 454 FLX instruments (Miller *et al.*, 2008).

Next-generation sequencing methods each have drawbacks that make *de novo* assembly difficult. Roche 454 reads are shorter than Sanger reads 500bp vs. >650bp, but they are less than one-tenth the cost per base. Roche 454 is also known to add or remove bases around homopolymer runs. Roche 454 reads are also more prone to high rates of small insertions and/or deletions, which

results in frame-shifts leading to problems arising in the identification of open reading frames more than in the case of single nucleotide errors. The Illumina sequencing costs are less than one-hundredth the cost of Sanger sequencing but read lengths result in 36 bp which prevents the assembly of even shorter repeats (Reinhardt *et al.*, 2008).

Reference assisted assembly becomes an inadequate tool when a degree of discrepancy in gene content exist. This is the case of the three fully sequenced pathovars of *Pseudomonas syringae* where approximately 75% of the genes are shared between any two of the three fully sequenced pathovars and as a result neither one can aid in a complete assembly of the other. Horizontally transferred fragments of a genome may assemble unsuccessfully when using a previously sequenced closely related reference isolate and in the end it is essential that these horizontally transferred fragments should be assembled suitably because they often provide genes essential for adaptive phenotypes that allow the bacteria to persist and exploit a variety of environments and hosts (Reinhardt *et al.*, 2008).

Projector 2 is a software program that runs on a UNIX platform. Linux Fedora Core 1 is used as the operating system. A locally installed copy of the BLAST software and an Apache web server with PHP is required. This program consists of 14 sub-programs written in Pascal and compiled by Free-Pascal version 1.0.10. All these programs are linked by a shell script. The web interface is made up of three parts namely the uploading of DNA sequences, selecting the desired settings and the status page. The status page presents the mapping results after a successful run, the status of a run can be inspected at any given time during the run. Projector 2 allows the majority of gaps to be closed within a genome sequence assembly using the minimum resources. The user is provided with pre-processed genomes, automated primer design on reliable and non-repetitive sequences and finally, the possibility of using multiple fragments of a template genome. Furthermore,

primers are designed which can be used in multiplex PCR approaches for closing gaps between different template sequences. As Projector 2 allows obtaining linkage information for template DNA sequences for which no ORF information is known, it can also use genomic DNA sequences that are less coding-dense (van Hijum *et al.*, 2005).

With the use of next generation sequencing techniques the genome of *Avibacterium paragallinarum* was sequenced in order to provide an improved understanding of the genetic composition of this bacterium.

2.6. Literature cited

Begley, T. P., C. Kinsland, R. A. Mehl, A. Osterman, P. Dorrestein. 2001. The Biosynthesis of Nicotinamide Adenine Dinucleotide in Bacteria. *Vitam. Horm.* **61**:103-119.

Bennet, P. M. 2008. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br. J. Pharmacol.* **153**:S347-S357.

Bernales, I., M. V. Mendiola and F de la Cruz. 1999. Intramolecular transposition of insertion sequence IS*91* results in second-site simple insertions. *Mol. Microbiol.* **33**:223-234.

Birch, E. A. 2006. Bacterial and Bacteriophage Genetics. 5th ed. Springer. *Springer science & Business media, Inc.,* New York, New York.

Blackall, P. J. 1988. Antimicrobial Drug Resistance and the Occurrence of Plasmids in *Haemophilus paragallinarum*. *Avian Dis.* **32**:742-747.

Blackall, P. J. 1999. Infectious Coryza: Overview of the Disease and New Diagnostic Options. *Clin. Microbiol. Rev.* **12**:627-632.

Blackall, P. J., C. J. Morrow, A. McInnes, L. E. Eaves and D. G. Rogers. 1990a. Epidemiologic studies on infectious coryza. Outbreaks in northern New South Wales, Australia, using serotyping, biotyping, and chromosomal DNA restriction endonuclease analysis. *Avian Dis.* **34**:267-276.

Blackall, P. J., L. E. Eaves and D. G. Rogers. 1990b. Proposal of a new serovar and altered nomenclature for *Haemophilus paragallinarum* in the Kume hemagglutinin scheme. *J. Clin. Microbiol.* **28**:1185-1187.

Blackall, P. J., H. Christensen, T. Beckenham, L. L. Blackall and M. Bisgaard. 2005. Reclassification of *Pasteurella gallinarum*, *Haemophilus paragallinarum*, *Pasteurella avium* and *Pasteurella volantium* as *Avibacterium gallinarum* gen. nov., comb. nov., *Avibacterium paragallinarum* comb. nov., *Avibacterium avium* comb. nov. and *Avibacterium volantium* comb. nov. *Int. J. Syst. Evol. Microbiol.* **55**:353-362.

Bragg, R. R., L. Coetzee and J. A. Verschoor. 1993. Plasmid-encoded NAD^+ independence in some South African isolates of *Haemophilus paragallinarum*. *Onderstepoort J. Vet. Res.* **60**:147-152.

Bragg, R. R. 2005. Effects of virulence of different serovars of *Haemophilus paragallinarum* on perceived vaccine efficacy. *Onderstepoort J. Vet. Res.* **72**:1-6.

Buys, R. 1982. Die bereiding van 'n bakterien teen *Haemophilus paragallinarum* besmetting, geskik vir Suid Afrikaanse omstandighede. M.Med. Vet (Bact.) thesis. Faculty of Veterinary Science, University of Pretoria.

Chen, X., J. K. Miflin, P. Zhanag and P. J. Blackall. 1996. Development and Application of DNA Probes and PCR Tests for *Haemophilus paragallinarum*. *Avian Dis.* **40**:398-407.

Eaves, L. E., D. G. Rogers and P. J. Blackall. 1989. Comparison of Hemagglutinin and Agglutinin Schemes for the Serological Classification of *Haemophilus paragallinarum* and Proposal of a New Hemagglutinin Serovar. *J. Clin. Microbiol.* **27**:1510-1513.

Esposito, D., J. J. Scocca. 1997. The integrase family of tyrosine recombinases: evolution of a conserved active site domain. *Nucleic Acids Res.* **25**:3605-3614.

Foster, J. W., A. G. Moat. 1980. Nicotinamide Adenine Dinucleotide Biosynthesis and Pyridine Nucleotide Cycle Metabolism in Microbial Systems. *Microbiol. Rev.* **44**:83-105.

Gerlach, G., J. Reidl. 2006. NAD^+ utilization in *Pasteurellaceae*: Simplification of a Complex Pathway. *J. Bacteriol.* **188**:6719-6727.

Groth, A. C., E. C. Olivares, B. Thyagarajan and M. P. Calos. 2000. A phage integrase directs efficient site-specific integration in human cells. *Proc. Natl. Acad. Sci. U.S.A.* **97**:5995-6000.

Hsu, Y.M., H. K. Shieh, W. H. Chen, T. Y. Sun and J. H. Shiang. 2007. Antimicrobial susceptibility, plasmid profiles and haemocin activities of *Avibacterium paragallinarum* strains. *Vet. Microbiol.* **124**:209-218.

Kume, K., A. Sawata, T. Nakai and M. Matsumoto. 1983. Serological Classification of *Haemophilus paragallinarum* with a Hemagglutinin System. *J. Clin. Microbiol.* **17**:958-964.

Kume, K., A. Sawata and Y. Nakase. 1978. *Haemophilus* Infections in Chickens. 1. Characterization of *Haemophilus paragallinarum* Isolated from Chickens Affected with Coryza. *Jap. J. vet. Sci.* **40**:65-73.

Lin, H. 2007. Nicotinamide adenine dinucleotide: beyond a redox coenzyme. *Org. Biomol. Chem.* **5**:2541-2554.

Mahillon, J. and M. Chandler. 1998. Insertion Sequences. *Microbiol. Mol. Biol. Rev.* **62**:725-774.

Mardis, E. R. 2008. Next-Generation Sequencing Methods. *Annu Rev Genom Human Genet.* **9**:387-402.

Mazel, D. 2006. Integrons: agents of bacterial evolution. *Nat Rev Microbiol.* **4**:608-620.

Merdanovic, M., E. Sauer and J. Reidl. 2005. Coupling of NAD^+ Biosynthesis and Nicotinamide Ribosyl Transport: Characterization of NadR Ribonucleotide Kinase Mutants of *Haemophilus influenzae*. *J. Bacteriol.* **187**:4410-4420

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

Morton, D. J., T. M. VanWagoner, T. W. Seale, P. W. Whitby and T. L. Stull. 2008. Catalase as a source of both X- and V-factor for *Haemophilus influenzae*. *FEMS Microbiol. Lett.* **279**(2008):157-161.

Nagy, A. 2000. Cre Recombinase: the Universal Reagent for Genome Tailoring. *Genesis*. **26**:99-109.

Page, L. A. 1962. *Haemophilus* infections in chickens. I. Characteristics of 12 *Haemophilus* isolates recovered from diseased chickens. *Am. J. Vet. Res.* **23**:85-95.

Reinhardt, J. A., D. A. Baltrus, M. T. Nishimura, W. R. Jeck, C. D. Jones and J. L. Dangl. 2008. Efficient *de novo* assembly of bacterial genomes using low coverage short read sequencing. *Genome Res.* November, 2008.

Rocha, E. P. C. 2004. Order and disorder in bacterial genomes. *Curr. Opin. Microbiol.* **7**:519-527.

Rodionov, D. A., L. Xiaoqing, I. A. Rodionova, C. Yang, L. Sorci, E. Dervyn, D. Martynowski, H. Zhang, M. S. Gelfand and A. L. Osterman. 2008a. Transcriptional regulation of NAD^+ metabolism in bacteria : genomic reconstruction of *NiaR* (*YrxA*) regulon. *Nucleic Acids Res.* **36**:2032-2046.

Rodionov, D. A., J. De Ingeniis, C. Mancini, F. Cimadamore, H. Zhang, A. L. Osterman and N. Raffaelli. 2008b. Transcriptional regulation of NAD^+ metabolism in bacteria:*NrtR* family of Nudix-related regulators. *Nucleic Acids Res.* **36**:2047-2059.

Roche Diagnostics. 2008. Roche Applied Science Indianapolis, Indiana. www.roche-applied-science.com

Rousseau, P., E. Gueguen, G. Duval-Valentin and M. Chandler. 2004. The helix-turn-helix motif of bacterial insertion sequence IS*911* transposase is required for DNA binding. *Nucleic Acids Res.* **32**:1335-1344.

Rothberg, J. M., J. H. Leamon. 2008. The development and impact of 454 sequencing. *Nat. Biotechnol.* **26**:1117-1124.

Rowe-Magnus, D. A. and D. Mazel. 2001. Integrons: natural tools for bacterial genome evolution. *Curr. Opin. Microbiol.* **4**:565-569.

Shendure, J. and H. Ji. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**:1135-1145.

Shaikh, A. C. and P. D. Sadowski. 2000. Chimeras of the Flp and Cre Recombinases: Tests of the Mode of Cleavage by Flp and Cre. *J. Mol. Biol.* **302**:27-48.

Terry, T. D., Y. M. Zalucki, S. L. Walsh, P. J. Blackall and M. P. Jennings. 2003. Genetic analysis of a plasmid encoding haemocin production in *Haemophilus paragallinarum*. *Microbiology.* **149**:3177-3184.

Tetu, S. G. and A. J. Holmes. 2008. A Family of Insertion Sequences That Impacts Integrons by Specific Targeting of Gene Cassette Recombination sites, the IS*IIII-attC* Group. *J. Bacteriol.* **190**:4959-4970.

Tobes, R. and E. Pareja. 2006. Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics.* **7**:62.

Van Hijum, S. A. F. T., A. L. Zomer, O. P. Kuipers and J. Kok. 2005. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Res.* **33**:560-566.

Voziyanov, Y., S. Pathania and M Jayaram. 1999. A general model for site-specific recombination by the integrase family recombinases. *Nucleic Acids Res.* **27**:930-941.

Yamamoto, R. 1984. Infectious coryza. In : Diseases of poultry, 9th ed. M. S. Hofstad, H. J. Barnes, B.W. Calnek, W. M. Reid, and H. E. Yonder, Jr., eds. *Iowa State University press*, Ames, Iowa. P. 178-186.

Chapter 3

Identification of putative genes involved in NAD^+ pathways within *Avibacterium paragallinarum*

3.1. Abstract

Nicotinamide adenine dinucleotide (NAD^+) plays an important role as co-factor in innumerable oxidation-reduction reactions. Certain members of the *Pasteurellaceae* require exogenous NAD^+ for growth since they are unable to synthesize and recycle their own. Their basic NAD^+ pathway consists of an uptake system with negligible resynthesis activity and, for that reason; these organisms rely on extracellular NAD^+ supply. The NAD^+-dependent pathway for *Av. paragallinarum* was identified with the use of genomic information obtained from the whole genome sequencing project for NAD^+-dependent *Av. paragallinarum* C-2 Modesto and A-1 0083. Genes identified within this pathway includes *nadR*, *pnuC* and *ppnK*. Various members of the *Pasteurellaceae* are NAD^+-independent due to the presence of a *nadV* gene either chromosomally or plasmid situated. This however is not the case for *Av. paragallinarum* NAD^+-independent strain 1750. An *nadC* and a partial *nadE* gene located on the genomic DNA were identified for this strain. The presence of these genes does not complete the NAD^+-independent pathway for *Av. paragallinarum* but it does seem to indicate that *Av. paragallinarum* follows the classical route for NAD^+ production as set out for bacteria in general.

3.2. Introduction

Nicotinamide adenine dinucleotide (NAD^+) plays an important role as co-factor in innumerable oxidation-reduction reactions (Lima *et al.*, 2009). NAD^+ functions as a redox co-factor by shuttling between NAD^+, the oxidized form, and NADH, the reduced form, but in effect the total concentration remains constant. In addition to being a redox coenzyme, NAD^+ functions as co-substrate for enzymatic reactions through enzymes with a unique chemistry and significant biological functions. These enzymes include NAD^+-dependent DNA ligases and NAD^+ consuming ADP-ribosyltransferases (Lin, 2007).

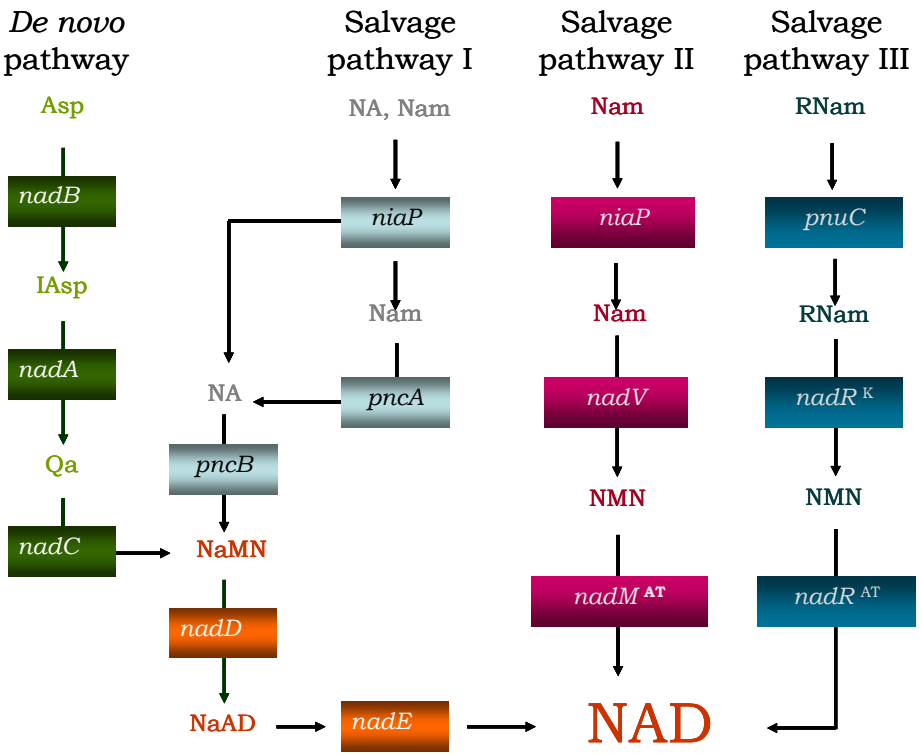


Figure 3.1. An overview of the NAD^+ biosynthesis pathways. The *de novo* pathway together with all three salvage pathways are illustrated along with the genes and their accociated proteins responsible for conversions or transportations (Rodionov *et al.*, 2008 a,b).

Figure 3.1 illustrates an overview of the NAD^+ biosynthesis pathways in bacteria. Three different building blocks can be used as precursors for the biosynthesis of NAD^+, namely nicotinic acid (NA) and nicotinamide (Nam) during the salvage pathways; and aspartic acid (Asp) during the *de novo* pathway (Lima *et al.*, 2009; Rodionov *et al.*, 2008 a,b). Three salvage pathways exist where precursors and intermediates are shared between these pathways but each has mostly different genes producing the required enzymes in order to perform the necessary conversions. During salvage pathway I (Figure 3.1) the genes along with their corresponding enzymes that comprise this pathway include *niaP* (niacin transporter); *pncA* (nicotinamide deaminase); *pncB* (nicotinate phosphoribosyltransferase); *nadD* (nicotinate mononucleotide adenylyltransferase) and *nadE* (NAD^+ synthetase). These proteins sequentially take NA and Nam thereby converting it to NMN (nicotinate mononucleotide) and thereafter to NaAD (nicotinate adenine dinucleotide) to finally produce NAD^+. Salvage pathway II consist of *niaP* (niacin transporter); *nadV* (nicotinamide phosphoribosyltransferase) and $nadM^{AT}$ (adenylyltransferase domain of nicotinamide mononucleotide adenylyltransferase). In this pathway NAD^+ is produced through the sequential conversion of Nam to NMN and finally to NAD^+. Salvage pathway III consist of the enzymes *pnuC* (ribosyl nicotinamide transporter); $nadR^K$ (Nicotinamide riboside kinase) and $nadR^{AT}$ (adenylyltransferase domain of nicotinamide mononucleotide adenylyltransferase). In this pathway RNam (ribosyl nicotinamide) is sequentially converted to NMN and then to NAD^+ (Rodionov *et al.*, 2008 a,b).

The *de novo* synthesis pathway comprises *nadA* (quinolinate synthase); *nadB* (L-aspartate oxidase); *nadC* (quinolinic acid phosphoribosyltransferase); *nadD* (nicotinate mononucleotide) and *nadE* (NAD^+ synthetase). These enzymes sequentially produce NAD^+ via the conversion of Asp to IA (imminoaspartate) followed by conversion to Qa (quinolinic acid) then to NaMN then NaAD and finally to NAD^+ (Rodionov *et al.*, 2008 a,b). The genes *nadA* and *nadB*

structure the quinolinate synthetase complex (Flachmann *et al.*, 1988) and *pncA* and *pncB* form part of the pyridine nucleotide cycles (Foster *et al.*, 1979).

Several members of the *Pasteurellaceae* are unable to synthesize or recycle NAD^+ and these organisms therefore rely on the extracellular environment for NAD^+. Pathway reductions, along with genome reductions are interconnected and it seems that evolution has preferred simplified pathways as is the case within the *Pasteurellaceae*. The NAD^+ requirement for members of the *Pasteurellaceae* is indicated as V-factor dependence or independence (Gerlach and Reidl, 2006). This is however not a definitive taxonomic criterion for species determination within the family *Pasteurellaceae* (Windsor *et al.*, 1993). Sources which satisfy V-factor requirements are NMN and NR. Figure 3.2 illustrates the genes that can be found within the NAD^+ pathways of the *Pasteurellaceae*. For the most part the V-factor dependent and independent pathways are identical except for the initial conversions. During the V-factor dependent route the enzymes from *pnuC* and $nadR^K$ sequentially convert NR to NMN and thereafter it is converted to NAD^+ through the chimeric enzyme of *nadR*. The V-factor independent route follows the conversions of Nam to NMN and then to NAD^+ through the enzymes encoded by *nadV* and *nadR* (Gerlach and Reidl, 2006).

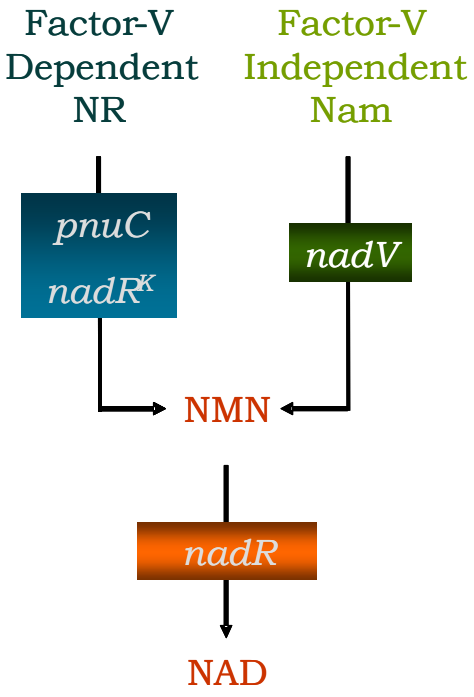


Figure 3.2. V-factor dependent and independent pathways found within *Pasteurellaceae*. The common routes include the conversion of NMN to NAD^+ while the the V-factor dependent route can be observed on the left-hand side and the V-factor independent route on the right-hand side.

Avibacterium paragallinarum has always been regarded as V-factor dependent until Mouahid and co-workers (1992) proved that V-factor independent strains did occur. The genes involved within the V-factor dependent or independent pathway have not yet been identified within *Av. paragallinarum*, however, Bragg and colleagues (1993) suggested that the V-factor independent trait of *Av. paragallinarum* might be carried on a plasmid. This phenomenon has also been observed within *Haemophilus ducreyi*, where a plasmid-encoded gene, namely *nadV*, was identified, and alongside this discovery it was illustrated that this gene was able to confer NAD^+ (or factor-V) independence when transformed within *Actinobacillus pleuropneumoniae* (Martin *et al.*, 2001). This chapter is devoted to the identification of genes involved in both the NAD^+-dependent and NAD^+-independent pathways of *Av. paragallinarum*.

3.3. Materials and Methods

3.3.1. Bacterial strains, growth conditions and plasmids

Avibacterium paragallinarum strains C-2 (Modesto) and 1750 were obtained from *Onderstepoort Biological Products*, Onderstepoort, South Africa. Strain Modesto is NAD^+-dependent and was cultivated in TM/SN medium containing 1% (m/v) biosate peptone, 1% (m/v) NaCl, 0.1% (m/v) starch and 0.05% (m/v) glucose with 5% (v/v) oleic acid-albumin complex, 1% (v/v) chicken serum, 1% (v/v) NAD^+ and 0.0005% (m/v) thiamine solution as supplements under oxygen limiting conditions. The NAD^+-independent strain 1750 was cultivated in the same manner as strain Modesto with the exclusion of NAD^+ and chicken serum supplements (Blackall and Yamamoto, 1990). Agar (1.7% m/v) was used in the case of solid media with cultivation occurring in a candle jar at 37˚C for 16 h.

Transformed *Escherichia coli* strain XL-10 Gold (Stratagene) was plated on Luria-Bertani (LB) plates containing agar (1.5% m/v) supplemented with 60 mg/l ampicillin, 10mg/l IPTG (isopropylthio-β-D-galactoside) and 40mg/l X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside). Cells for DNA mini-preparation were grown in LB broth under the appropriate antibiotic pressure (Sambrook *et al.*, 1989). Cloning vectors allowing blue-white selection used in this study are listed in Table 3.1.

Table 3.1: Plasmids used in this study.

Plasmid	Origin
pUC18	
pUC19	
pGem3Z	Promega
pGem® T-Easy	Promega

3.3.2. Enzymes, chemicals, kits and other consumables

Restriction enzymes from Fermentas were used and polymerase chain reaction (PCR) reagents were obtained from New England Biolabs® Inc (NEB).

Chemicals used during this study were of analytical or molecular biology grade and were used without further purification. All relevant chemicals were obtained from Sigma-Aldrich or Merck. NAD^+, bovine serum albumin fraction V along with proteinase K was obtained from Roche Molecular Biochemicals.

Oligonucleotides were obtained from Integrated DNA Technologies (IDT), Inqaba Biotechnical Industries or Bioneer.

3.3.3. Techniques applied

3.3.3.1. DNA extraction

Genomic DNA extraction was performed with the use of a QIAamp® DNA Mini kit from Qiagen. The manufacturer's recommendations were meticulously followed except for the elution procedure where genomic DNA was eluted in 2 X 80 µl 10mM Tris-HCl, pH 8. Plasmid DNA extraction was done with the use of the QIAamp® Maxi kit from Qiagen where the manufacturer's recommendations were followed.

3.3.3.2. Identification of *Av. paragallinarum*

Av. paragallinarum was identified through the use of an identification PCR developed specifically for this bacterium (Chen *et al.,* 1996). The reaction mixture contained 2 µM of the primer set AP-1F and AP-1R (Table 3.2) together with 0.5 µg of genomic DNA, 5 µl of the 10 x ThermoPol buffer, 200 µM dNTP's, 2 units *Taq* polymerase, filled to a final reaction volume of 50 µl using sterile

redistilled water. The reaction cycle included an initial heat denaturation of 10 min at 94˚C, this was done in order to break the cells for whole cell PCR enabling genomic DNA release after which 2 units *Taq* polymerase was added. The 10 min heat denaturation was not required where isolated genomic DNA was used as template. This was followed by 30 cycles of denaturation (94˚C for 25 sec); annealing (55˚C for 50 sec); elongation (72˚C for 45 sec) and a final elongation cycle of 7 min at 72˚C with an expected amplification product of approximately 500 bp.

Additionally, the 16S ribosomal DNA (rDNA) region was amplified in order to identify *Av. paragallinarum* (Mendoza-Espinoza *et al.*, 2008). The reaction mixture contained 2 µM of the primer set 27F and 1492R (Table 3.2) together with 5 µl of the 10 x ThermoPol buffer, 200 µM dNTP's, 2 units *Taq* polymerase, in a final reaction volume of 50 µl using sterile redistilled water. The reaction cycle included an initial denaturation of 2 min at 95˚C followed by 35 cycles of denaturation (95˚C for 30 sec); annealing (49˚C for 45 sec); elongation (72˚C for 1 min 30 sec) and a final elongation cycle of 10 min at 72˚C with an expected amplification result of approximately 1 500 bp. The PCR product obtained was excised from the agarose gel, purified and directly sequenced without any sub-cloning. This allowed detection of possible foreign genomic DNA contamination.

3.3.3.3. Polymerase chain reactions

All other PCR reactions were performed with the use of NEB *Taq* DNA polymerase according to the manufacturer's recommendations. The reaction mixture contained 5 µl of the 10 x ThermoPol buffer, 2 µM upstream and downstream primers, 200 µM dNTP's, 0.5 µg template DNA and 2 units *Taq* polymerase, filled to a final reaction volume of 50 µl using sterile redistilled water. Thermal cycling was performed using an Eppendorf Mastercycler

Personal with the following cycle profile: initial denaturation of 2 min at 94˚C, 35 cycles of denaturation (94˚C for 30 sec); annealing (temperature is primer dependent for 30 sec); elongation (at 72˚C for 1 min/kb) and a final elongation cycle of 7 min at 72˚C.

3.3.3.4. Inverse polymerase chain reactions

Inverse PCR reactions were performed as depicted by Ochman *et al.* 1988. All inverse PCR reactions were performed using KapaHiFi™ DNA polymerase from Kapa Biosystems (Pty) Ltd, South Africa. The reaction mixture contained 10 µl of the 5 x KapaHiFi Fidelity Buffer with $MgCl_2$, 0,3 µM upstream and downstream primers, 0,3 mM dNTP's, 0.5 µg template DNA and 0.5 units of KapaHiFi™ DNA polymerase in a final volume of 50 µl. Thermal cycling entailed initial denaturation (95˚C for 2 min), 35 cycles of denaturation (98˚C for 20 sec); annealing (temperature is primer dependent for 15 sec); elongation (at 72˚C for 30 sec/kb) and a final elongation cycle of 5 min at 72˚C.

Table 3.2: Oligonucleotide primers used in this study.

Primer	5'→3' Orientation	Application
AP-1F	TgA ggg TAg TCT TgC ACg CgA AT	ID PCR
AP-1R	CAA ggT ATC gAT CgT CTC TCT ACT	ID PCR
24F	AgA gTT TgA TCM Tgg CTC Ag*	16S rDNA
1492R	TAC ggY TAC CTT gTT ACg ACT T*	16S rDNA
NadR-F	CTT ggC gAC TCA ATT TCg ATA TAA gAC	PCR/Sequencing
NadR-R	gCT TAT TgA TgT CTA gTT TCg CCT ATT TAC	PCR/Sequencing
pnuC-F	ggA AAC CCT TTg AAg Tgg TAT gg	PCR/Sequencing
pnuC-R	gCA CCg TAC ATT AAA TAC ATC gC	PCR/Sequencing
ppnKcomp-F	CTC CCg TTT AAC TgT ATg CTT ATC Ag	PCR/Sequencing
ppnKcomp-R	gCA TCg TTC gTT TCA gAC TAT Tgg	PCR/Sequencing
RepB-2F	CCC CTg AAA TAA TgC CTT ATC TTT Cg	PCR/Sequencing
RepB-1R	ggA TCg ATA CgA TCT AgC TTC AC	PCR/Sequencing
RepBIN-F	gAA gAC gTA CCg AAT gAA gAT TTg C	Inverse PCR
RepBIN-R	gCA TAC AgA ACT ATT CTC TgC TCC	Inverse PCR
Int-1F	gCC TTg CTA TCg CCC AAg ATA AAC AgC	PCR/Sequencing
Int-2R	gTg AgC Tgg gAC ATA ATC gAg AAg	PCR/Sequencing
IntIN-F	CAC AAg ACg gAA ACC gAA TAg AC	Inverse PCR
IntIN-R	ggA CAT AAC CgA gAA gAA ATT ACA gC	Inverse PCR
NADVDEG-F	gTK AAT TTg CCR TTA YTR AAR SAW g*	PCR
NADVDEG-R	RCC ATY YTC AAA WAM CAg CTC K*	PCR
NadV1-F	CAC CTC gTA gCA ATg AAC AAg	PCR
NadV1-R	gCT gAA gTT AAA CCA TCA ACg TAA g	PCR
NadC-F	CAC AAg ACg ACC ATC TCA AAC C	PCR/Sequencing
NadC-R	ggA AgT AgC gAT gTA CCA TAT Tgg	PCR/Sequencing
NadCIN-F	gCC ACA TTC CCT TAT TCA ACg gTg	Inverse PCR
NadCIN-R	gTA ACC ACA gAA gTT CgA CgC AC	Inverse PCR
NadE-F	gCC TAT TTA ATT CAA Tgg gTT gAg C	PCR
nadE-R	CgA ATC CAT AAA ACA TTC ATA CAg Tgg	PCR

* Degenerate nucleotides: M = A/C; Y = C/T; R = A/g; S = g/C; W = A/T; K = g/T.

In order to avoid confusion between guanine and cytosine within the oligonucleotides guanine is indicated as lowercase.

3.3.3.5. General techniques

PCR, inverse PCR and DNA products were initially analyzed with the use of agarose gel electrophoresis. Agarose gels consisted of 1% (w/v) agarose in TAE buffer (0.1 M Tris; 0.05 M EDTA, pH 8; 0.1 mM glacial acetic acid) stained with Goldview™ (SBS Genetech) according to the manufacturer's recommendations. DNA was separated within the agarose gels at 5.6 V/cm for 45 min and documented using BioRad Gel Doc XR system. DNA to be isolated from agarose gels for further studies was visualized using a Dark Reader®.

PCR products were either purified from agarose gels or from solution with the use of the Bioflux BioSpin Gel Extraction Kit (Bioer Technology).

PCR products selected for sequencing were cloned into a suitable vector and modified plasmid vectors were transformed as follows: rubidium chloride competent *E. coli* cells (50 µl) (Hanahan, 1983) were transformed with the ligation mixture containing the DNA of interest ligated into a cloning vector of choice. The transformation was done according to Sambrook *et al.* (1989) followed by plating onto LB plates supplemented with the appropriate antibiotic. Plates were incubated at 37˚C for 16 h, after which positive transformants were selected for DNA mini-preparations and grown for 16 h at 37˚C with shaking. Plasmid DNA was extracted from the possible positive transformants using the lysis by boiling method according to Sambrook *et al.* (1989).

Restriction analysis, PCR screening and/or sequence analysis were used as screening methods for the correct recombinant plasmids. Recombinant plasmids used for sequencing were purified with the Bioflux BioSpin Plasmid Extraction Kit (Bioer Technology).

In order to determine the nucleotide composition of the relevant insert, purified positive templates were used in sequencing reactions. Sequence reactions were performed using the ABI Prism® Big Dye™ Terminator Cycle Sequencing Ready Reaction Kit v. 3.1 (Applied Biosystems, USA) according to the manufacturer's recommendations. Nucleotide composition was analyzed with an Applied Biosystems 3130*xl* Genetic Analyzer. Sequence analysis was performed using either Geneious Pro 4.6.1 (Biomatters) or CLC Combined Workbench 3.5.1 (CLC bio A/S).

Homology trees and protein alignments were constructed with the use of DNAMAN 5.2.9 from Lynnon Biosoft and plasmid maps with pDRAW32 1.1.104 from ACACLONE software.

3.4. Results and Discussion

3.4.1. NAD^+-independent pathway

In 2003, van Zyl reported on plasmids extracted from *Av. paragallinarum* NAD^+-independent strains 1742, 1345 and F11-3. Restriction analysis indicated that the plasmids obtained from these three isolates were identical. This plasmid was subjected to sequencing and it resembled the p250 plasmid of *Av. paragallinarum* discovered by Terry and co-workers (2003). The maps of these two plasmids are depicted in Figure 3.3.

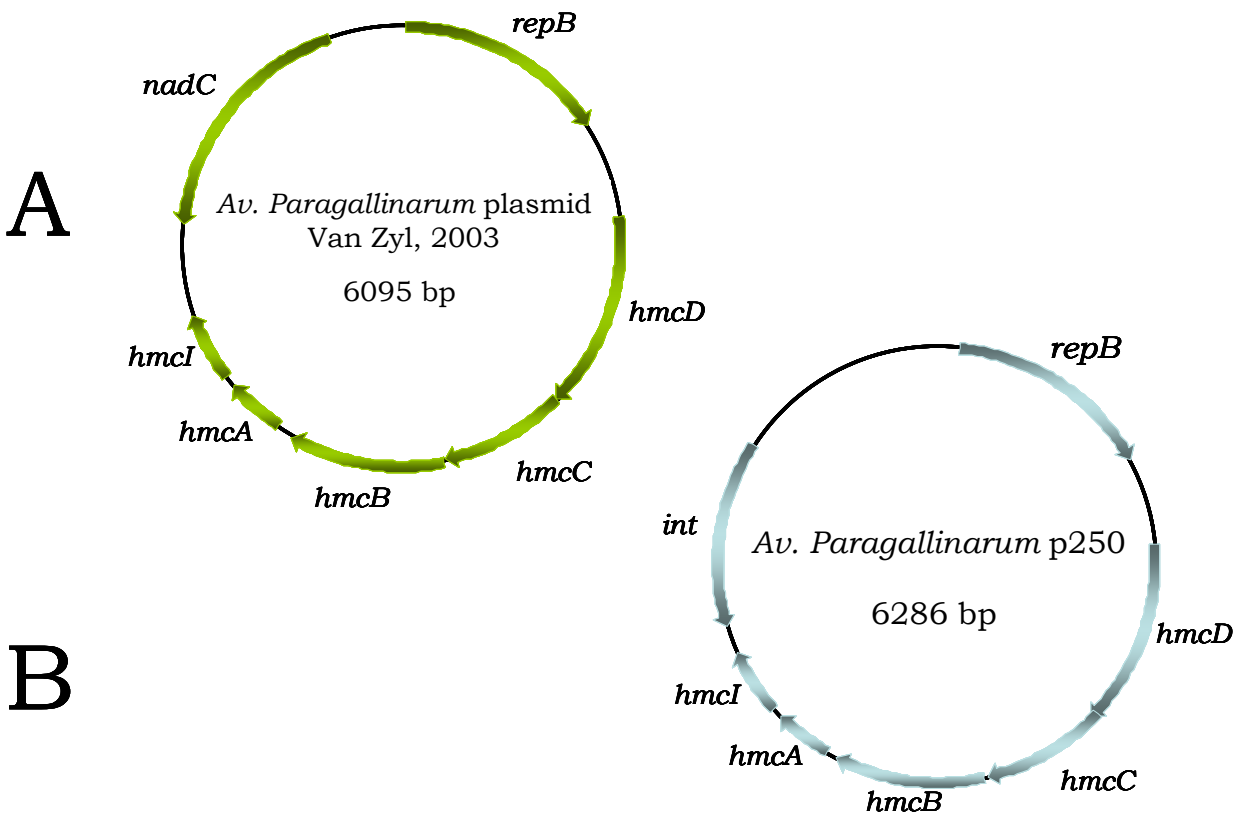


Figure 3.3. *Av. paragallinarum* plasmid illustrations. A illustrates the plasmid of van Zyl (2003) and B the plasmid p250 isolated by Terry and co-workers (2003). These plasmids are very similar except for the *int* gene which is present on p250 but absent from the van Zyl plasmid, and the *nadC* gene which is present on the van Zyl plasmid and absent from p250.

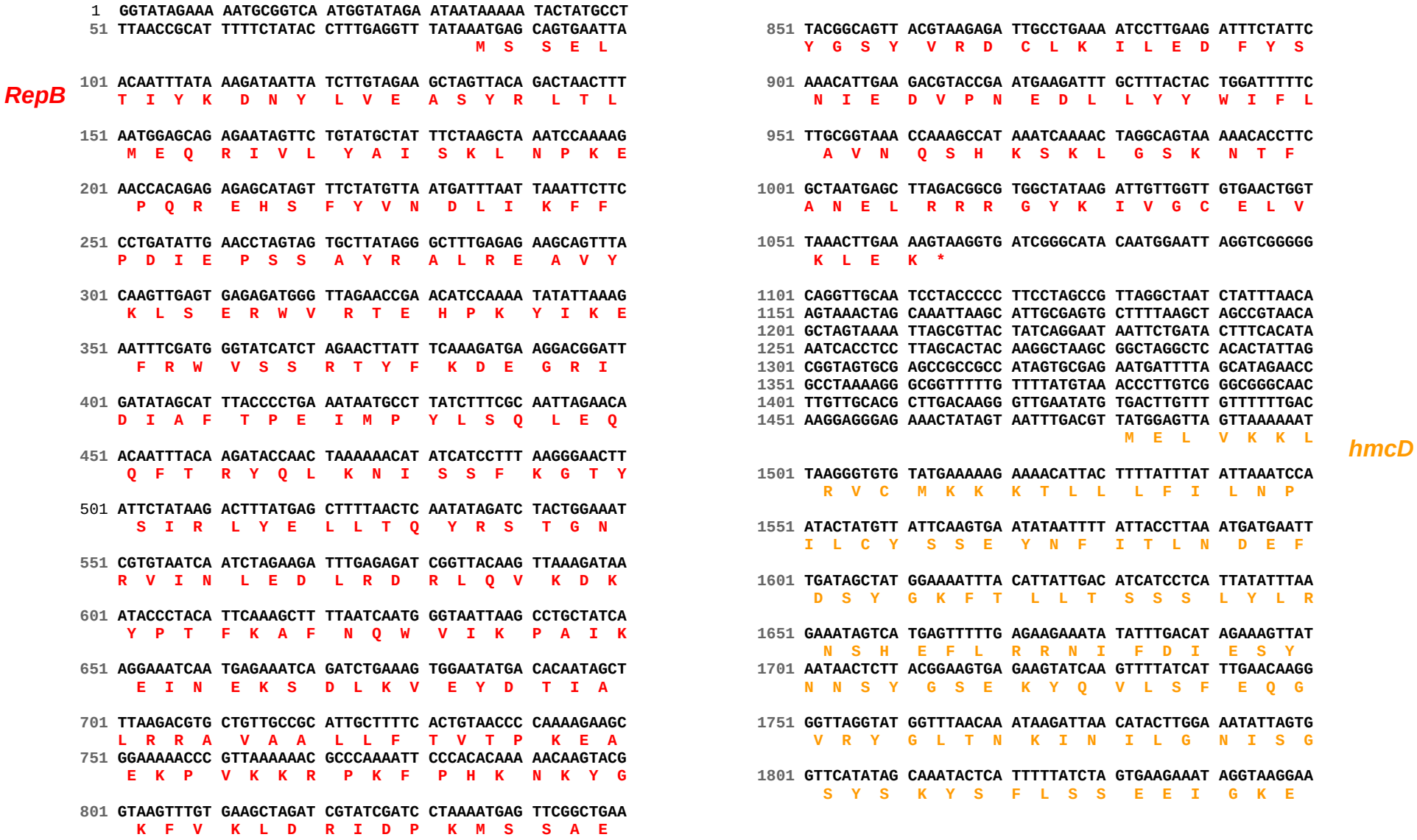
The sequence reported by van Zyl (2003) was re-analyzed and although it showed resemblance to p250, there were significant differences. These differences included the presence of an *int* gene on p250 that was substituted with a *nadC* gene on the van Zyl plasmid and between the *nadC* and *hmcI* genes on the van Zyl plasmid a partial *nadE* gene exists which is not present on p250. Both the haemocin operon and the *repB* from p250 were present on the van Zyl plasmid.

Cultivation of the original strains used by van Zyl (2003) was attempted but all the relevant strains were found to be non-viable, therefore plasmid extractions

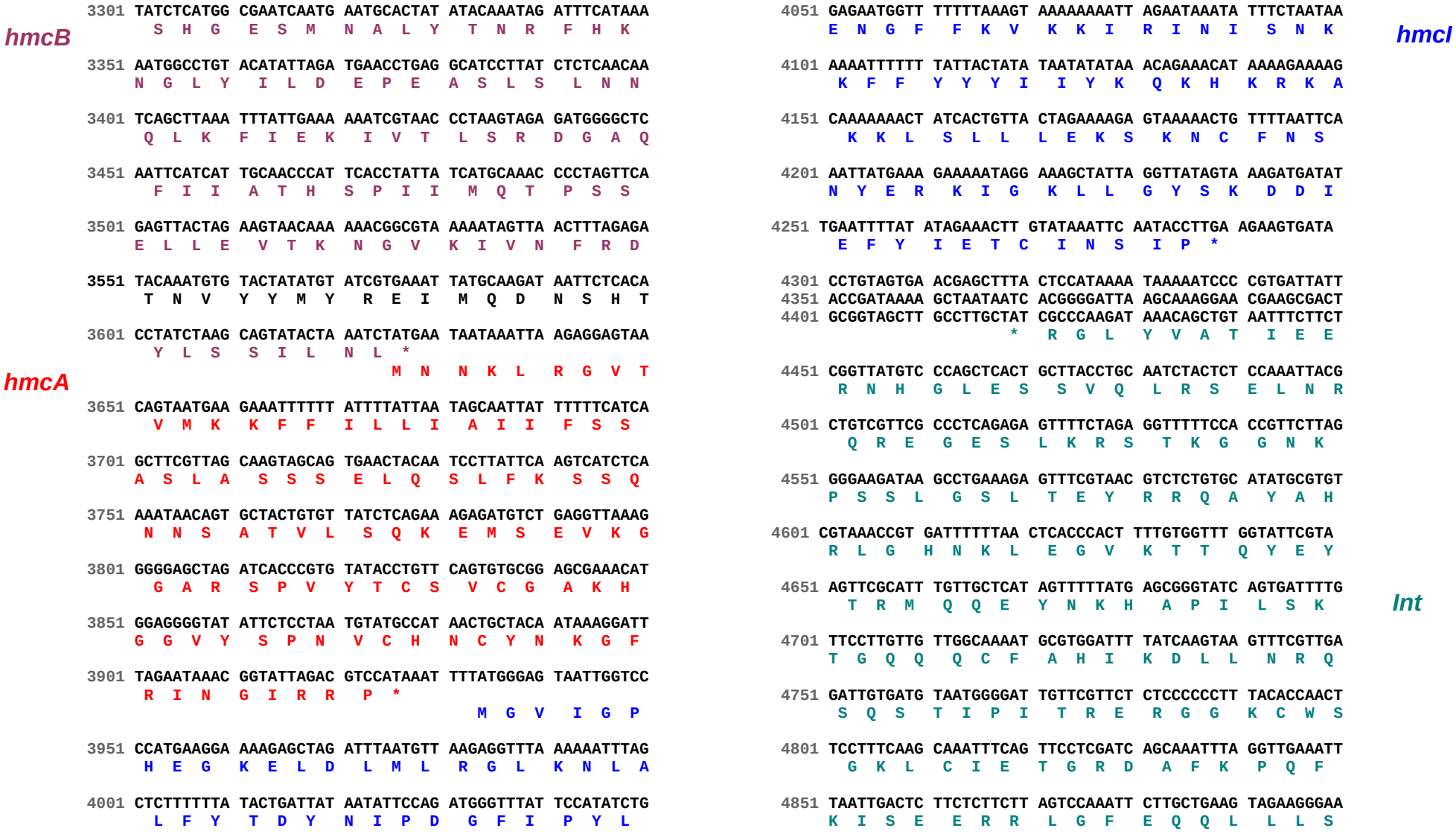
were performed on four other *Av. paragallinarum* NAD^+-independent strains 1750; 1233; 1932 and 1471. Plasmid DNA could only be obtained from two of the four strains and restriction analysis yielded an identical pattern. Plasmid DNA obtained from strain 1750 was used for further analysis.

Based on unpublished results obtained in this laboratory as well as published data, primers were designed to amplify fragments of the replication gene, *repB*, and the integrase gene, *int*, from the *Av. paragallinarum* p250 plasmid (Van Zyl, 2003; Terry *et al.*, 2003). A *repB* PCR, using primers RepB-2F and RepB-1R (annealing 53°C; elongation 1 min) alongside an *int* PCR using primers Int-1F and Int-2R (annealing 55°C; elongation 1 min). Inverse PCR reactions (Ochman *et al.*, 1988) using primer pairs RepBIN-F, RepBIN-R (annealing 52°C; elongation 6 min) or IntIN-F, IntIN-R (annealing 53°C elongation 6 min), were designed to obtain flanking regions adjacent to the *repB* and *int* genes.

These PCR reactions were performed in order to investigate the possibility of the extracted plasmid being plasmid p250 from *Av. paragallinarum* as described by Terry *et al.* (2003). PCR fragments obtained were sub-cloned followed by sequencing analysis and compared to the related genes present on the p250 plasmid. Sequence results indicated that the extracted plasmid is identical to plasmid p250. The full sequence obtained is depicted in Figure 3.4.


RepB
hmcD
1 GGTATAGAAA AATGCGGTCA ATGGTATAGA ATAATAAAAA TACTATGCCT
51 TTAACCGCAT TTTTCTATAC CTTTGAGGTT TATAAATGAG CAGTGAATTA
M S S E L
101 ACAATTTATA AAGATAATTA TCTTGTAGAA GCTAGTTACA GACTAACTTT
T I Y K D N Y L V E A S Y R L T L
151 AATGGAGCAG AGAATAGTTC TGTATGCTAT TTCTAAGCTA AATCCAAAAG
M E Q R I V L Y A I S K L N P K E
201 AACCACAGAG AGAGCATAGT TTCTATGTTA ATGATTTAAT TAAATTCTTC
P Q R E H S F Y V N D L I K F F
251 CCTGATATTG AACCTAGTAG TGCTTATAGG GCTTTGAGAG AAGCAGTTTA
P D I E P S S A Y R A L R E A V Y
301 CAAGTTGAGT GAGAGATGGG TTAGAACCGA ACATCCAAAA TATATTAAAG
K L S E R W V R T E H P K Y I K E
351 AATTTCGATG GGTATCATCT AGAACTTATT TCAAAGATGA AGGACGGATT
F R W V S S R T Y F K D E G R I
401 GATATAGCAT TTACCCCTGA AATAATGCCT TATCTTTCGC AATTAGAACA
D I A F T P E I M P Y L S Q L E Q
451 ACAATTTACA AGATACCAAC TAAAAAACAT ATCATCCTTT AAGGGAACTT
Q F T R Y Q L K N I S S F K G T Y
501 ATTCTATAAG ACTTTATGAG CTTTTAACTC AATATAGATC TACTGGAAAT
S I R L Y E L L T Q Y R S T G N
551 CGTGTAATCA ATCTAGAAGA TTTGAGAGAT CGGTTACAAG TTAAAGATAA
R V I N L E D L R D R L Q V K D K
601 ATACCCTACA TTCAAAGCTT TTAATCAATG GGTAATTAAG CCTGCTATCA
Y P T F K A F N Q W V I K P A I K
651 AGGAAATCAA TGAGAAATCA GATCTGAAAG TGGAATATGA CACAATAGCT
E I N E K S D L K V E Y D T I A
701 TTAAGACGTG CTGTTGCCGC ATTGCTTTTC ACTGTAACCC CAAAAGAAGC
L R R A V A A L L F T V T P K E A
751 GGAAAAACCC GTTAAAAAAC GCCCAAAATT CCCACACAAA AACAAGTACG
E K P V K K R P K F P H K N K Y G
801 GTAAGTTTGT GAAGCTAGAT CGTATCGATC CTAAAATGAG TTCGGCTGAA
K F V K L D R I D P K M S S A E
851 TACGGCAGTT ACGTAAGAGA TTGCCTGAAA ATCCTTGAAG ATTTCTATTC
Y G S Y V R D C L K I L E D F Y S
901 AAACATTGAA GACGTACCGA ATGAAGATTT GCTTTACTAC TGGATTTTTC
N I E D V P N E D L L Y Y W I F L
951 TTGCGGTAAA CCAAAGCCAT AAATCAAAAC TAGGCAGTAA AAACACCTTC
A V N Q S H K S K L G S K N T F
1001 GCTAATGAGC TTAGACGGCG TGGCTATAAG ATTGTTGGTT GTGAACTGGT
A N E L R R R G Y K I V G C E L V
1051 TAAACTTGAA AAGTAAGGTG ATCGGGCATA CAATGGAATT AGGTCGGGGG
K L E K *
1101 CAGGTTGCAA TCCTACCCCC TTCCTAGCCG TTAGGCTAAT CTATTTAACA
1151 AGTAAACTAG CAAATTAAGC ATTGCGAGTG CTTTTAAGCT AGCCGTAACA
1201 GCTAGTAAAA TTAGCGTTAC TATCAGGAAT AATTCTGATA CTTTCACATA
1251 AATCACCTCC TTAGCACTAC AAGGCTAAGC GGCTAGGCTC ACACTATTAG
1301 CGGTAGTGCG AGCCGCCGCC ATAGTGCGAG AATGATTTTA GCATAGAACC
1351 GCCTAAAAGG GCGGTTTTTG TTTTATGTAA ACCCTTGTCG GGCGGGCAAC
1401 TTGTTGCACG CTTGACAAGG GTTGAATATG TGACTTGTTT GTTTTTTGAC
1451 AAGGAGGGAG AAACTATAGT AATTTGACGT TATGGAGTTA GTTAAAAAAT
M E L V K K L
1501 TAAGGGTGTG TATGAAAAAG AAAACATTAC TTTTATTTAT ATTAAATCCA
R V C M K K K T L L L F I L N P
1551 ATACTATGTT ATTCAAGTGA ATATAATTTT ATTACCTTAA ATGATGAATT
I L C Y S S E Y N F I T L N D E F
1601 TGATAGCTAT GGAAAATTTA CATTATTGAC ATCATCCTCA TTATATTTAA
D S Y G K F T L L T S S S L Y L R
1651 GAAATAGTCA TGAGTTTTTG AGAAGAAATA TATTTGACAT AGAAAGTTAT
N S H E F L R R N I F D I E S Y
1701 AATAACTCTT ACGGAAGTGA GAAGTATCAA GTTTTATCAT TTGAACAAGG
N N S Y G S E K Y Q V L S F E Q G
1751 GGTTAGGTAT GGTTTAACAA ATAAGATTAA CATACTTGGA AATATTAGTG
V R Y G L T N K I N I L G N I S G
1801 GTTCATATAG CAAATACTCA TTTTTATCTA GTGAAGAAAT AGGTAAGGAA
S Y S K Y S F L S S E E I G K E


hmcD
1851 AAATCATTTA ATAAATTAAG GTTTGATACC ATAAATATTG GATTGTCCAC
K S F N K L R F D T I N I G L S T
1901 TAAATTAAAT GATTCTGAAT CAGACTGGAA TAAATCTATA TATTTTTCAG
K L N D S E S D W N K S I Y F S V
1951 TAGACGCCAT AAGCAAGAGA GAAAGTTTAT TCAAAAACTT TTATTTTGAC
D A I S K R E S L F K N F Y F D
2001 TATATAATAG ACAAAACAAT AGACCCTATA GTTTTATCTT TAAAAACAGG
Y I I D K T I D P I V L S L K T G
2051 TGTAAAATAT TTCTCAAAAA TAAAAAGCAA CAATCTCAAT TTTAAACCAT
V K Y F S K I K S N N L N F K P S
2101 CAAATGAAAT AAATATAAAG CCTAGAGTGG ATTTTTTAGT TAACCAAAAA
N E I N I K P R V D F L V N Q K
2151 ATATCCTTAA GTTTATCTAC AGAGATTCAA TTAAGAAGCA GTGAAAAATA
I S L S L S T E I Q L R S S E K Y
2201 CAATGGAAAA GTAACAAATG TATCTGGAAT AGAAAATTTC TTATCAATGG
N G K V T N V S G I E N F L S M G
2251 GGATTTCGTA TAATTTAGAT TTGAAAAATC GTTTATATTT TGAAACTTCT
I S Y N L D L K N R L Y F E T S
L K L L
2301 TTTAATACGA CGGGAAACTC AGGTGCTACA ATTTACCTTA ACCTTGAGAA
F N T T G N S G A T I Y L N L E K
L I R R E T Q V L Q F T L T L R R
hmcC
2351 GGATTTGTGA TATAATACTC CTATTATGTA TATCAGGTTC TCTTCTAGCC
D L *
I C D I I L L L C I S G S L L A
2401 AAAGAAATAA GCTCATACAA TGAACTTAAA AATTTTAATG TTATTAGGCA
K E I S S Y N E L K N F N V I R Q
2451 AACAAAAACC AATTCATGTG GAGCAGCGGC TCTTGCTACG ATGCTTAAAT
T K T N S C G A A A L A T M L K Y
2501 ATAAATTCCA TCTTAGCGAA ATTAACGAAG ATGTCATTTT ATCCAAGATG
K F H L S E I N E D V I L S K M
2551 AAAAATAAAA ATGAAGAAGC CTCCTTTTTT GAACTTGCAA AAATTTCAAA
K N K N E E A S F F E L A K I S K
2601 AACATTCAAT ATCAACGCTA TTGGAATATC CTTAACATTA AAGGAGTTAT
T F N I N A I G I S L T L K E L L
2651 TAAATATAAA GCAACCAGTA ATTGCTTATG TAAATAACAA TTTGAACAAC
N I K Q P V I A Y V N N N L N N
2701 GATCATTTTG TCATCATTAA TGGTATTATT AACAAAGAAT TATTAATATC
D H F V I I N G I I N K E L L I S
2751 AGATCCAGCG GTTGGTAATT ATTCATTAAG AGCATCAGAT TTTGAAAAAA
D P A V G N Y S L R A S D F E K I
2801 TATGGATATT AAGGGAGGGT GAAAAAGGAG ATGTTTTATA CTTATATAGT
W I L R E G E K G D V L Y L Y S
2851 GAAAAAAAAG ATTACTTAGA GTTTATTGAT AATATTGAAA AAAAACATAG
E K K D Y L E F I D N I E K K H R
2901 AATTTCCCTA AGGAATTAAA TTTGTATATA TATGAAATAT ATAGAAAAAA
I S L R N *
L Y I Y E I Y R K N
hmcB
2951 TAAAAATAAA TTTTTTCTTG ATTATTTATT TCCAATAAAA TTTAATAACC
K N K F F L D Y L F P I K F N N P
3001 CAATATCAAT CTTTTTTGGT GAGAATGGAA TAGGTAAATC GACTATAATG
I S I F F G E N G I G K S T I M
3051 GAATCTCTAG CCGTACACCT TGGATGCCCA GCCGAAGGAG GTTCAAAAAA
E S L A V H L G C P A E G G S K N
3101 TTTTAATTTT TCAACTGAAA ATACACACAT CCAAGTACCT GATATCGTGG
F N F S T E N T H I Q V P D I V V
3151 TAAAAAAAGG AACTAGATTT CCTAAGGATG TATTTTTCTA TAGAGCAGAA
K K G T R F P K D V F F Y R A E
3201 ACTTTCTATA CCTTTATGAG CGAGATGAAA AGGTTAGACA CCCCAGAATC
T F Y T F M S E M K R L D T P E S
3251 AGGAGGGGAG AAGATAAAAT ATTACTATGG GGGAGTTGAA TTGCATAAGT
G G E K I K Y Y Y G G V E L H K L


hmcB
hmcA
hmcI
Int
3301 TATCTCATGG CGAATCAATG AATGCACTAT ATACAAATAG ATTTCATAAA
S H G E S M N A L Y T N R F H K
3351 AATGGCCTGT ACATATTAGA TGAACCTGAG GCATCCTTAT CTCTCAACAA
N G L Y I L D E P E A S L S L N N
3401 TCAGCTTAAA TTTATTGAAA AAATCGTAAC CCTAAGTAGA GATGGGGCTC
Q L K F I E K I V T L S R D G A Q
3451 AATTCATCAT TGCAACCCAT TCACCTATTA TCATGCAAAC CCCTAGTTCA
F I I A T H S P I I M Q T P S S
3501 GAGTTACTAG AAGTAACAAA AAACGGCGTA AAAATAGTTA ACTTTAGAGA
E L L E V T K N G V K I V N F R D
3551 TACAAATGTG TACTATATGT ATCGTGAAAT TATGCAAGAT AATTCTCACA
T N V Y Y M Y R E I M Q D N S H T
3601 CCTATCTAAG CAGTATACTA AATCTATGAA TAATAAATTA AGAGGAGTAA
Y L S S I L N L *
M N N K L R G V T
3651 CAGTAATGAA GAAATTTTTT ATTTTATTAA TAGCAATTAT TTTTTCATCA
V M K K F F I L L I A I I F S S
3701 GCTTCGTTAG CAAGTAGCAG TGAACTACAA TCCTTATTCA AGTCATCTCA
A S L A S S S E L Q S L F K S S Q
3751 AAATAACAGT GCTACTGTGT TATCTCAGAA AGAGATGTCT GAGGTTAAAG
N N S A T V L S Q K E M S E V K G
3801 GGGGAGCTAG ATCACCCGTG TATACCTGTT CAGTGTGCGG AGCGAAACAT
G A R S P V Y T C S V C G A K H
3851 GGAGGGGTAT ATTCTCCTAA TGTATGCCAT AACTGCTACA ATAAAGGATT
G G V Y S P N V C H N C Y N K G F
3901 TAGAATAAAC GGTATTAGAC GTCCATAAAT TTTATGGGAG TAATTGGTCC
R I N G I R R P *
M G V I G P
3951 CCATGAAGGA AAAGAGCTAG ATTTAATGTT AAGAGGTTTA AAAAATTTAG
H E G K E L D L M L R G L K N L A
4001 CTCTTTTTTA TACTGATTAT AATATTCCAG ATGGGTTTAT TCCATATCTG
L F Y T D Y N I P D G F I P Y L
4051 GAGAATGGTT TTTTTAAAGT AAAAAAAATT AGAATAAATA TTTCTAATAA
E N G F F K V K K I R I N I S N K
4101 AAAATTTTTT TATTACTATA TAATATATAA ACAGAAACAT AAAAGAAAAG
K F F Y Y Y I I Y K Q K H K R K A
4151 CAAAAAAACT ATCACTGTTA CTAGAAAAGA GTAAAAACTG TTTTAATTCA
K K L S L L L E K S K N C F N S
4201 AATTATGAAA GAAAAATAGG AAAGCTATTA GGTTATAGTA AAGATGATAT
N Y E R K I G K L L G Y S K D D I
4251 TGAATTTTAT ATAGAAACTT GTATAAATTC AATACCTTGA AGAAGTGATA
E F Y I E T C I N S I P *
4301 CCTGTAGTGA ACGAGCTTTA CTCCATAAAA TAAAAATCCC CGTGATTATT
4351 ACCGATAAAA GCTAATAATC ACGGGGATTA AGCAAAGGAA CGAAGCGACT
4401 GCGGTAGCTT GCCTTGCTAT CGCCCAAGAT AAACAGCTGT AATTTCTTCT
* R G L Y V A T I E E
4451 CGGTTATGTC CCAGCTCACT GCTTACCTGC AATCTACTCT CCAAATTACG
R N H G L E S S V Q L R S E L N R
4501 CTGTCGTTCG CCCTCAGAGA GTTTTCTAGA GGTTTTTCCA CCGTTCTTAG
Q R E G E S L K R S T K G G N K
4551 GGGAAGATAA GCCTGAAAGA GTTTCGTAAC GTCTCTGTGC ATATGCGTGT
P S S L G S L T E Y R R Q A Y A H
4601 CGTAAACCGT GATTTTTTAA CTCACCCACT TTTGTGGTTT GGTATTCGTA
R L G H N K L E G V K T T Q Y E Y
4651 AGTTCGCATT TGTTGCTCAT AGTTTTTATG AGCGGGTATC AGTGATTTTG
T R M Q Q E Y N K H A P I L S K
4701 TTCCTTGTTG TTGGCAAAAT GCGTGGATTT TATCAAGTAA GTTTCGTTGA
T G Q Q Q C F A H I K D L L N R Q
4751 GATTGTGATG TAATGGGGAT TGTTCGTTCT CTCCCCCCTT TACACCAACT
S Q S T I P I T R E R G G K C W S
4801 TCCTTTCAAG CAAATTTCAG TTCCTCGATC AGCAAATTTA GGTTGAAATT
G K L C I E T G R D A F K P Q F
4851 TAATTGACTC TTCTCTTCTT AGTCCAAATT CTTGCTGAAG TAGAAGGGAA
K I S E E R R L G F E Q Q L L L S

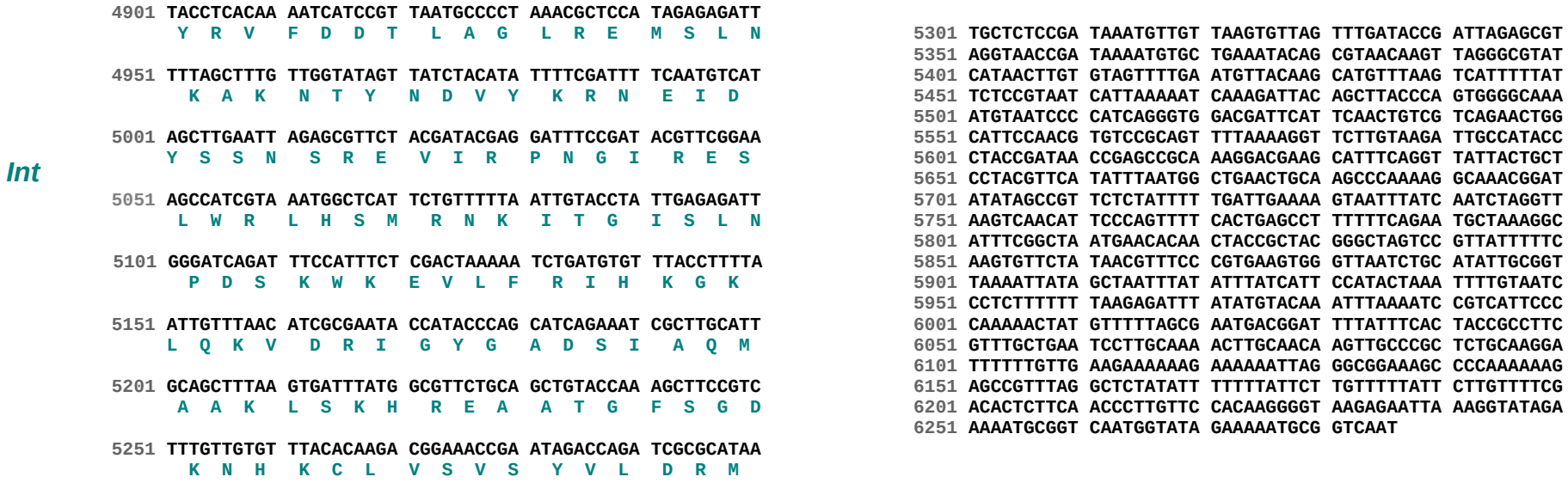


Figure 3.4. Nucleotide and coding sequences of the *Av. paragallinarum* plasmid from strain 1750. The 7 genes encoded by the plasmid are indicated by different colored font. All genes are transcribed in the same orientation, except the *Int* gene which is encoded in the 3' → 5' orientation.

Certain members of the *Pasteurellaceae* are NAD^+-independent due to the presence of a *nadV* gene either carried on a plasmid or located on their genomic DNA. Degenerate primers were designed based on the *nadV* gene of *Haemophilus ducreyi* 35000HP and *Actinobacillus pleuropneumoniae* serovar 7 str. AP76. Degenerate PCR reactions using primers NADVDEG-F and NADVDEG-R (annealing 49ºC; elongation 30 sec) were performed on both genomic DNA and plasmid DNA extracts from *Av. paragallinarum* 1750. Plasmid extracts were included to ensure that if a low copy plasmid encoding *nadV* is present, this gene could be detected even if plasmid DNA could not be detected; however no PCR product was obtained.

To circumvent this problem, specific primers (nadV1-F; nadV1-R) were designed for the nicotinamide phosphoribosyltransferase genes based on the *nadV1/nadV2* genes of *H. ducreyi* 35000HP. These PCR reactions (annealing 51ºC; elongation 30 sec) were performed on both the plasmid and genomic DNA of strain 1750 for the same reason as described above.

Based on unpublished data of Van Zyl (2003) primers were also designed to amplify the quinolinic acid phosphoribosyltransferase, *NadC*, gene. The primer pair, NadC-F and NadC-R (annealing 55ºC; elongation 1 min) was used to perform PCR reactions on both plasmid and genomic DNA extracts of *Av. paragallinarum* strain 1750 and genomic DNA of *Av. paragallinarum* strain Modesto (Figure 3.5). Again the plasmid fraction from strain 1750 was used to exclude the possibility of a plasmid other than p250, containing the relevant gene present at very low concentration.

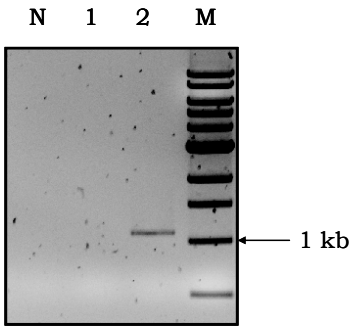


Figure 3.5. PCR amplification of the *nadC* gene of *Av. paragallinarum*. The negative control is indicated by lane N whereas lane M indicates the 1 kb molecular marker (NEB) the plasmid DNA fraction is represented by lane 1 and the genomic DNA by lane 2.

No product was found for amplifications performed on plasmid DNA or genomic DNA of *Av. paragallinarum* strain Modesto, however, a fragment with the correct size was amplified from the genomic DNA of *Av. paragallinarum* 1750 (Figure 3.5) and sequencing proved that this was indeed the *nadC* gene van Zyl (2003) reported on. Following sequence analysis the ORF of the *nadC* gene consist of 1 107 bp comprising a protein of 369 amino acids with a molecular weight of 91.9 kDa and a theoretical isoelectric point (pI) of 5.06 (Figure 3.6). This protein forms part of the TIM barrel superfamily proteins which share a structurally conserved phosphate binding motif and contain an eight beta/alpha closed barrel structure in general which was all identified within *Av. paragallinarum nadC*.

```
  1    MKEIEKKEKV NITRRPSQTF NFDEKIGEGY FSASYFLKAK HIAEQKRSKQ
 51    WVTMQFFQKK QAILCGVDEA IALLQRFARN PEGLKIWALN DGDVISPFES
101    VLIVEGFYED FAYLEGIIDG ILSRCTSIAT NVRETLNAAN GKPIIFMGDR
151    NDYFTLQPTD GYAAYIGGAT MQATDAMHQY QGGTGVGTMP HSLIQLFNGN
201    LIQACEAYMQ VYPNSPLTAL VDYNNDVITD SLKVAKHFGE KLYAVRVDTS
251    ANMVDRYFLT HQETFGQEDL RGVNVPLIKQ LRHQLDQAGF PWVKIVVSGG
301    FNQQKVQHFE QQNAPVDYYG IGASLLKISI DFTGDCVRLN GKAQAKAGRR
351    YRVNPRLSTV HFYHVFLM*
```

Figure 3.6. Protein sequence of the *nadC* gene from *Av. paragallinarum*.

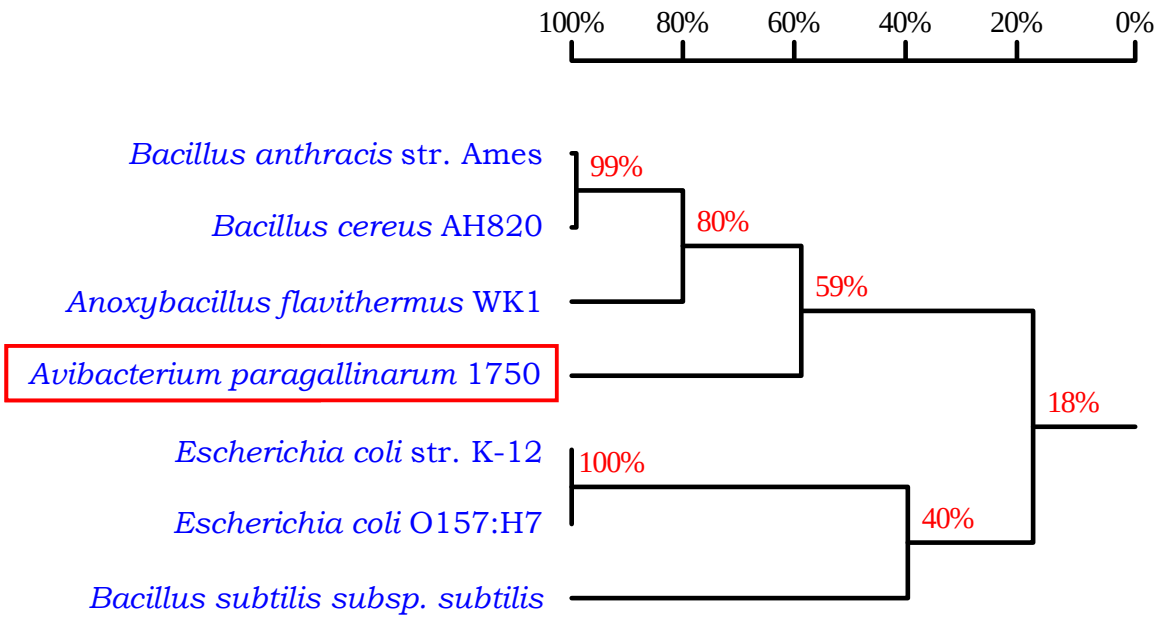


Figure 3.7. Homology tree of the *nadC* gene of *Av. paragallinarum*. *Av. paragallinarum nadC* gene is enclosed within a red box and show homology towards the *nadC* genes from various *Bacillus* sp.

The *nadC* gene of *Av. paragallinarum* display homology towards various *Bacillus* sp. *nadC* when compared to *nadC* genes from both *Bacillus* sp. and *E. coli* (Figure 3.7).

Inverse PCR primers were designed (NadCIN-F and NadCIN-R; annealing 55°C; elongation 6 min) in order to obtain flanking regions neighbouring the *nadC* gene of *Av. paragallinarum*. Various methods were employed to obtain the flanking regions adjacent to the *nadC* gene but none were successful.

As mentioned earlier a partial *nadE* sequence was also reported by van Zyl (2003) and was located between the *nadC* and *hmcI* genes. Primer set nadE-F and nadE-R was designed based on the plasmid sequence of van Zyl and a partial *nadE* was amplified (Figure 3.8) from extracted genomic DNA of NAD^+-independent strain 1750. The nucleotide sequence was subjected to the *ORF finder* search engine available on the National Centre for Bioinformatics (NCBI)

website (http://www.ncbi.nlm.nih.gov/), which in turn was subjected to a protein blast search on the NCBI database. The *ORF finder* (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size within a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes and the deduced amino acid sequence can be searched against the sequence database using the WWW BLAST server. The partial *nadE* sequence can be viewed in Figure 3.9.

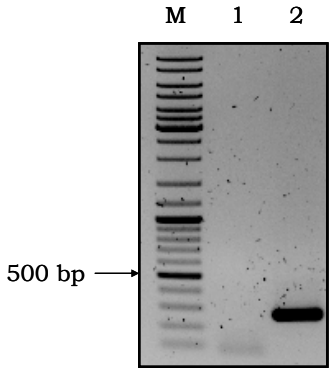


Figure 3.8. PCR amplification of a partial *nadE* from *Av. paragallinarum* 1750. Lane M indicates the molecular marker, GeneRuler™ (Fermentas); lane 1 is the negative control and lane 2 represent the *nadE* gene amplified from genomic DNA of *Av. paragallinarum* NAD^+-independent strain 1750.

```
1   MRTGAPVQGL LLPSATTSTQ DVEDAKLVAE SAKCPLLTLP ITPLYECFMD S   51
```

Figure 3.9. Protein sequence of a partial *nadE* gene from *Av. paragallinarum* 1750.

The *nadC* and *nadE* primers were used in different combinations in order to see if they are located close to each other thus forming an operon, in an attempt to isolate the full length *nadE* gene. However, no results were obtained.

As stated earlier, various members of the *Pasteurellaceae* are NAD^+-independent due to the presence of a *nadV* gene either carried on a plasmid or located on the genomic DNA. With the use of the NCBI database these members were identified together with their associated genes from the following organisms: *Aggregatibacter aphrophilus NJ8700: nadV; A. pleuropneumoniae* Serovar 3 *Str. JL03: nadV; H. ducreyi* plasmid *pNAD1: nadV* and *H. ducreyi 35000HP: nadV1* and *nadV2*. Genes implicated within the *de novo* pathway for NAD^+ synthesis, as set out for bacteria in general, present in members of the *Pasteurellaceae* were also identified and include: *Haemophilus influenzae R2846: modD* (possible *nadC*)*; Haemophilus parasuis 29755: nadA; Mannheimia succiniciproducens MBEL55E: nadE* and *Mannheimia haemolityca PHL213: nadB*. NAD^+ *de novo* synthesis genes were identified for *Escherichia coli* and *Bacillus* species (Table 3.3).

Identified genes within the *Pasteurellaceae*, *E. coli* and *Bacillus* sp. which play a role in the *de novo* synthesis of NAD^+ for bacteria in general, or in the V-factor independence pathway for *Pasteurellaceae* were used in a nucleotide-to-nucleotide comparative study with whole genome sequence data for the NAD^+-dependant *Av. paragallinarum* Modesto strain (Chapter 4). Alongside that Professor Henrik Christensen from the Department of Veterinary Disease Biology, Center for Applied Bioinformatics, Copenhagen University, Denmark supplied the contigs assembled of a titanium pyrosequencing run on the NAD^+-dependent *Av. paragallinarum* A-1 (0083) type strain. This contig set was analysed in parallel to that of the C-2 serovar. This study was performed in order to investigate the possibility of the presence of relevant genes within NAD^+-dependent *Av. paragallinarum* and to investigate the possibility of inactive genes leading to NAD^+-dependence in *Av. paragallinarum*. The nucleotide-to-nucleotide comparison was done using CLC Combined Workbench 3.5.1 from CLC bio A/S, parameters were set at a minimum base overlap of 35 nucleotides and the homology was set at a low confidence value.

Table 3.3: NAD^+ pathways related genes present in *Escherichia coli* and *Bacillus* species.

Name of Organism		NAD^+ *de novo* Pathway			NAD^+ Salvage Pathway				NAD^+ Uptake Pathway#		
	NCBI Accession Number	*nadB*	*nadA*	*nadC*	*nadD**	*nadE**	*pncA*	*pncB*	*pnuC*	*nadR*	*nadF* (*nadK*/ *ppnk*)
Escherichia coli O157H7 str. TW14359	NC_013008	YP_003079362	YP_003076737	YP_003076081	YP_003076642	YP_003078195	YP_003078224	YP_003076968	YP_003076738	YP_003081222	YP_003079397
Escherichia coli K-12 substr. MG 1655	NC_000913	NP_417069	NP_415271	NP_414651	NP_415172	NP_416254	NP_416282	NP_415451	NP_415272	NP_418807	NP_417105
Bacillus cereus AH187	NC_011658	YP_002340490	YP_002340488	YP_002340489	YP_002340392	YP_002338048	YP_002336727	YP_002340745			YP_002337329 YP_002340700
Bacillus subtilis subsp. Subtilis str. 168	NC_000964	NP_390665	NP_390663	NP_390664	NP_390442	NP_388195	NP_391054	NP_391053		NP_390667	NP_389043 NP_390832

*Genes present in the *de novo*, salvage and uptake pathways. #No *pncC* and *pnuE* genes were obtained for the above used organisms regarding the NAD^+ uptake pathway and these were omitted in the construction of the above table. Underlined genes were used in a comparative study to obtain genes present on the genome of *Avibacterium paragallinarum*. The above table was constructed according to Gerlach and Reidl, 2006.

Any contig that revealed a significant hit with any region of the compared gene were selected for further analysis. The selected contig was subjected to an ORF search with the use of the *ORF finder* search engine available on the NCBI website.

None of the implicated genes were identified based on a homology search with genome sequence data of the NAD^+-dependent *Av. paragallinarum* Modesto or *Av. paragallinarum* 0083. These results indicate that *Av. paragallinarum* NAD^+-dependency in these strains is probably due to the absence of *de novo* NAD^+ synthesis machinery.

3.4.2. NAD^+-dependent pathway

NAD^+-dependent pathways are referred to as NAD^+ salvage pathways, it is a basic pathway which consists of an uptake system with negligible resynthesis activity and for that reason these organisms rely on extracellular NAD^+ supply. The whole genome sequence of a NAD^+-dependent strain of *Av. paragallinarum* Modesto was obtained as described in chapter 4. An *in silico* study was performed to obtain all available sequence data pertaining to NAD^+-dependent pathway genes of the *Pasteurellaceae* available on the National Centre for Bioinformarics (NCBI) website (Figure 3.4a-c). Various genes were selected to use during the *in silico* study and were selected to include at least two representatives of any gene found within the NAD^+-salvage pathway within the *Pasteurellaceae*. The genes included: *H. influenzae* Rd KW20*: pnuC*, *nadR* and *ppnK (nadF)*; *Pasteurella multocida* subsp. *multocida* str. Pm70: *nadR*, *ppnK* (*nadF*) and *pncB* and *H. influenzae* 86-028NP*: nadN*. Sequence comparisons were performed with the selected genes against data obtained from the whole genome sequencing project of *Av. paragallinarum* Modesto (Chapter 4).

Table 3.4a: NAD^+ pathways related genes within the *Pasteurellaceae*.

Name of Organism		NAD^+-dependent Pathway			NAD^+-independent Pathway	Other genes found	NAD^+-dependent/ independent
	NCBI Accession Number	*pnuC*	*nadR*	*nadF* (*ppnK*)	*nadV*		
*Actinobacillus minor 202**	NZ_ACFT00000000						Dependent (Moller *et al.*, 1996)
*Actinobacillus minor NM305**	NZ_ACQL00000000		ZP_04752689	ZP_04752941		*pncB* ZP_04754006	Dependent (Moller *et al.*, 1996)
Aggregatibacter aphrophilus NJ8700	NC_012913	YP_003008262	YP_003008169	YP_003007717	YP_003006670		Independent (Nørskov-Lauritsen and Kilian, 2006)
Actinobacillus pleuropneumoniae L20	NC_009053	YP_001053866	YP_001052759	YP_001052925			Dependent (Lei *et al.*, 2009)
*Actinobacillus pleuropneumoniae Serovar 1 Str. 4074**	NZ_AACK00000000	ZP_00348309	ZP_00348320			*pncB* ZP_00133995 ZP_00133996	Dependent (Kucerova *et al.*, 2005)
Actinobacillus pleuropneumoniae Serovar 3 *Str. JL03*	NC_010278	YP_001652194	YP_001651097	YP_001651261	NC_010278		Dependent (Xu *et al.*, 2008)
Actinobacillus pleuropneumoniae Serovar 7 Str. AP76	NC_010939	YP_001969026	YP_001967840	YP_001968008			Dependent (Jacobsen *et al.*, 2005)
Actinobacillus succinogenes 130Z	NC_009655	YP_001345275	YP_001343798	YP_001344190		*pncB* YP_001343947 *nadE* YP_001343948	Dependent (Park and Zeikus, 1999)
Haemophilus ducreyi plasmid pNAD1	NC_005329				NP_957670		Independent (Martin *et al.*, 2001)
Haemophilus ducreyi 35000HP	NC_002940		NC_002940	NP_874165	*nadV1* NP_873863 *nadV2* NP_873871		Independent (Munson *et al.*, 2004)
*Haemophilus influenzae 6P18H1**	NZ_ABWW00000000		ZP_04464770	ZP_04465540			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae 7P49H1**	NZ_ABWV00000000		ZP_04465921	ZP_04467097			Dependent (Morton *et al.*, 2008)
Haemophilus influenzae 86-028NP	NC_007146	YP_248746	YP_248471	YP_247734		*nadN* YP_247923	Dependent (Morton *et al.*, 2008)

*Unfinished genome sequences. Underlined genes were used in a comparative study to obtain genes present on the genome of *Avibacterium paragallinarum*. The above table was constructed according to Gerlach and Reidl, 2006.

Table 3.4b: NAD^+ pathways related genes within the *Pasteurellaceae*.

Name of Organism		NAD^+-dependent Pathway			NAD^+-independent Pathway	Other genes found	NAD^+-dependent/ independent
	NCBI Accession Number	*pnuC*	*nadR*	*nadF* (*ppnK*)	*nadV*		
*Haemophilus influenzae PittAA**	NZ_AAZG00000000	ZP_01789946	ZP_01790534	ZP_01790960			Dependent (Morton *et al.*, 2008
Haemophilus influenzae PittEE	NC_009566	YP_001291061	YP_001291347	YP_001290402			Dependent (Morton *et al.*, 2008)
Haemophilus influenzae PittGG	NC_009567	YP_001292992	YP_001292706	YP_001291978			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae PittHH**	NZ_AAZH00000000		ZP_01792348	ZP_01792391			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae PittII**	NZ_AAZI00000000		ZP_01794048	ZP_01795245			Dependent (Morton *et al.*, 2008)
Haemophilus influenzae Rd KW20	NC_000907	NC_000907	NP_438922	NP_438244			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae R2846**	NZ_AADO00000000	ZP_00155654				*modD* (Possible *NadC*) ZP_00155095	Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae R2866**	NZ_AADP00000000	ZP_00156919					Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae R3021**	NZ_AAZE00000000	ZP_01785950	ZP_01786288 ZP_01786287	ZP_01786407 ZP_01786411			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae 22.1-21**	NZ_AAZD00000000		ZP_01784152 ZP_01784153	ZP_01784860			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae 22.4-21**	NZ_AAZJ00000000	ZP_01796343	ZP_01796431 ZP_01796432 ZP_01796433	ZP_01797728 ZP_01797725			Dependent (Morton *et al.*, 2008)
*Haemophilus influenzae 3655**	NZ_AAZF00000000	ZP_01788176	ZP_01787851	ZP_01789099			Dependent (Morton *et al.*, 2008)
Haemophilus parasuis SH0165	NC_011852	YP_002476518	YP_002476399	YP_002476023 YP_002475837 YP_002475651		*pncB* YP_002475991	Dependent (Blanco *et al.*, 2004)

*Unfinished genome sequences. Underlined genes were used in a comparative study to obtain genes present on the genome of *Avibacterium paragallinarum*. The above table was constructed according to Gerlach and Reidl, 2006.

Table 3.4c: NAD^+ pathways related genes within the *Pasteurellaceae*.

Name of Organism		NAD^+-dependent Pathway			NAD^+-independent Pathway	Other genes found	NAD^+-dependent/ independent
	NCBI Accession Number	*pnuC*	*nadR*	*nadF* (*ppnK*)	*nadV*		
*Haemophilus parasuis 29755**	NZ_ABKM00000000		ZP_02477968	ZP_02477641		*nadA* ZP_02479500	Dependent (Blanco *et al.*, 2004)
Haemophilus somnus 2336	NC_010519		YP_001785295	YP_001784320		*pncB* YP_001785207	Independent (Challacombe *et al.*, 2007)
Haemophilus somnus 129PT	NC_008309		YP_718294	YP_718862		*pncB* YP_718207	Independent (Challacombe *et al.*, 2007)
Mannheimia succiniciproducens MBEL55E	NC_006300	YP_087385	YP_087361	YP_087934		*pncB* YP_089081 *nadE* YP_089080	Dependent (Hong *et al.*,2004)
*Mannheimia haemolityca PHL213**	NZ_AASA00000000	ZP_04977896	ZP_04978727		ZP_04978905	Possible *nadB* ZP_04979036	
Pasteurella multocida subsp. Multocida str. Pm70	NC_002663		NP_246326	NP_245270		*pncB* NP_245936	Dependent (Kielstein *et al.*, 2001)

*Unfinished genome sequences. Underlined genes were used in a comparative study to obtain genes present on the genome of *Avibacterium paragallinarum*. The above table was constructed according to Gerlach and Reidl, 2006.

The nucleotide-to-nucleotide comparison was done using CLC Combined Workbench 3.5.1 from CLC bio A/S, parameters were set at a minimum base overlap of 35 nucleotides and the homology was set at a low confidence value. Any contig that revealed a significant hit with any region of the compared gene were selected for further analysis. The selected contig was subjected to an ORF search with the use of the *ORF finder* search engine available on the NCBI website. Within each contig all the possible ORF's identified were subjected to a protein blast search on the NCBI database. Through this method, possible role players in the *Av. paragallinarum* NAD^+ salvage pathway, were identified namely *nadR*, *pnuC* and *ppnK*.

Oligonucleotides were designed initially to confirm the presence of the relevant gene as indicated by the *in silico* study, and thereafter to amplify the complete ORF of the relevant gene (Table 3.5).

Table 3.5: Primers used to confirm or to obtain the complete ORF of the relevant genes implicated in the NAD^+-salvage pathway of *Av. paragallinarum*.

Primer	Application
NadR-F NadR-R	Confirm presence of *nadR*
NadRcomp-F NadRcomp-R	Obtain complete ORF of *NadR*
pnuC-F pnuC-R	Confirm presence of *pnuC*
pnuCcomp-F pnuCcomp-R	Obtain complete ORF of *pnuC*
ppnKcomp-F	Confirm presence of *ppnK*
ppnK-R ppnKcomp-2R	Obtain complete ORF of *ppnK*

The nucleotide-to-nucleotide comparison study implicated a possible *nadR* present on a contig 20 341 bp in size (Figure 3.10). Primer set NadR-F and NadR-R was used to amplify a fragment of the implicated *nadR* from both *Av. paragallinarum* NAD⁺-dependent strain Modesto and *Av. paragallinarum* NAD⁺-independent strain 1750. Even though strain 1750 is capable of synthesizing NAD⁺, NAD⁺ salvage mechanisms still need to exist. Following sequence confirmation, primer set NadRcomp-F and NadRcomp-R was used to amplify the complete ORF of *nadR* from *Av. Paragallinarum* (Figure 3.10 and Figure 3.11).

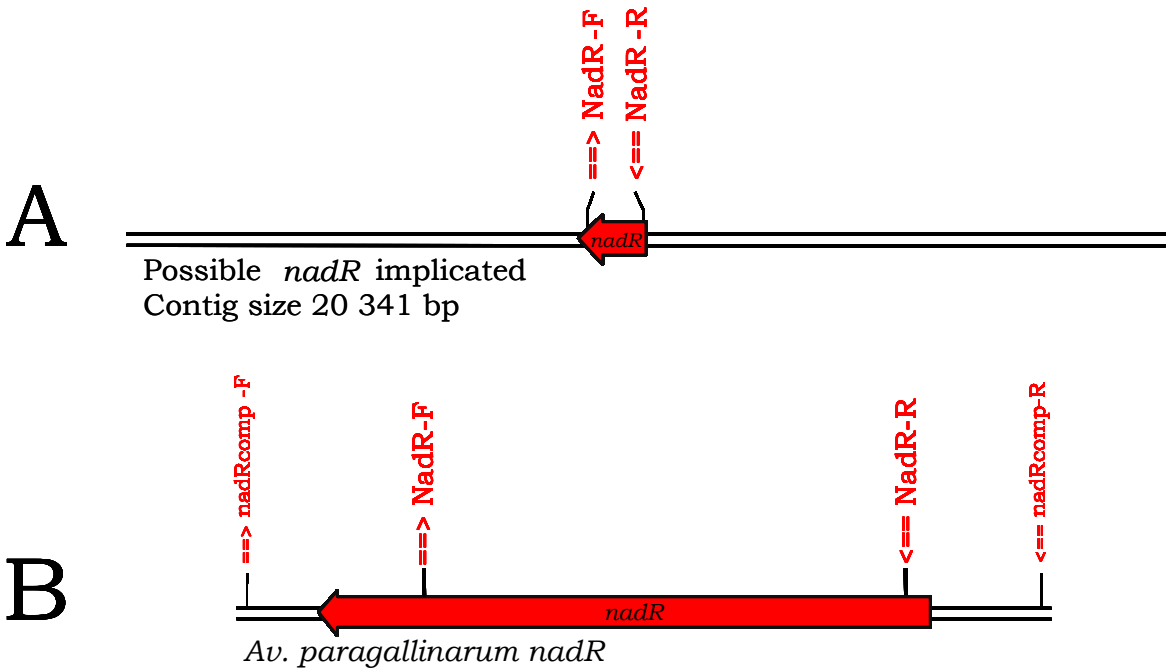


Figure 3.10. *nadR* gene of *Av. paragallinarum*. Primers nadR-F and nadR-R were used to confirm the implicated *nadR* gene (A). Thereafter primers nadRcomp-F and nadRcomp-R (B) were used to amplify the complete *nadR* ORF from *Av. paragallinarum*.

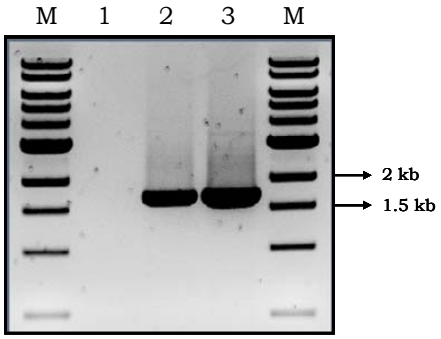


Figure 3.11. PCR amplification of the *nadR* gene of *Av. paragallinarum* together with flanking regions where primer set nadRcomp was used and an expected size of 1 720 bp. NEB 1 kb ladder (lane M) was used as molecular marker where lane 1 indicates the negative control; lane 2 the complete *nadR* gene amplified from *Av. paragallinarum* NAD^+-dependent strain Modesto and lane 3 the complete *nadR* gene amplified from *Av. paragallinarum* NAD^+-independent strain 1750.

The *nadR* genes from both the NAD^+-dependent and the NAD^+-independent strains from *Av. paragallinarum* consists of 1 293 bp which translates into an ORF of 431 amino acids with a molecular weight of 106.3 kDa and a theoretical pI of 5.08. A 99.7% identity exists between these two genes when they are compared with one another on protein level (Figure 3.12). Conserved domain analysis revealed that the substrate binding site, the active site motif and the nicotinamide recognition loop is present within *Av. paragallinarum nadR* genes and these regions form the NMNAT domain of the nadR protein. The NadR protein of *H. influenzae* is a bifunctional enzyme possessing both NMN adenylytransferase (NMNAT) and ribosylnicotinamide kinase activities. Its function is essential for the growth and survival of *H. influenzae* and thus may present a new highly specific anti-infectious drug target. The N-terminal domain that hosts the NMNAT activity is closely related to archaeal NMNAT.

```
Av. paragallinarum Modesto nadR MLLRSLLMSSFAYLQQKRKQLHLKVSEVCEQANITRAYFNQLVSGKIKNPSAAKLSALHK
Av. Paragallinarum   1750  nadR MLLRSLLMSSFAYLQQKRKQLHLKVSEVCEQANITRAYFNQLVSGKIKNPSAAKLSALHK

Av. paragallinarum Modesto nadR VLQITEQLDKKVGVIFGKFYPVHTGHINMIYEAFSKVDEVHVVVCSDTERDLKLFYDSKM
Av. paragallinarum   1750  nadR VLQITEQLDKKVGVIFGKFYPVHTGHINMIYEAFSKVDEVHVVVCSDTERDLKLFYDSKM

Av. paragallinarum Modesto nadR KRMPTVQDRLRWMQQIFKYQKNQIVIHHLVEDGLPSYPNGWEAWAERVKQLFNEKNVNPS
Av. paragallinarum   1750  nadR KRMPTVQDRLRWMQQIFKYQKNQIVIHHLVEDGLPSYPNGWEAWAERVKQLFNEKNVNPS

Av. paragallinarum Modesto nadR VVFSSEPQDKAPYEKYLNLQVELVDPQRQFFNVSATKIRNNPFHYWKFIPKEVRPFFAKT
Av. paragallinarum   1750  nadR VVFSSEPQDKAPYEKYLNLQVELVDPQRQFFNVSATKIRNNPFHYWKFIPKEVRPFFAKT

Av. paragallinarum Modesto nadR IAILGGESSGKTVLVNKLATVFNTTSAWEYGREFVFEQLGGDEQAMQYSDYPQMALGHQR
Av. paragallinarum   1750  nadR IAILGGESSGKTVLVNKLATVFNTTSAWEYGREFVFEQLGGDEQAMQYSDYPQMALGHQR

Av. paragallinarum Modesto nadR YIDYAVRHAHKVAIVDTDFITTQAFCIQYEGKAHPFLDSMIKEYPFDITILLNNNTKWVD
Av. Paragallinarum   1750  nadR YIDYAVRHAHKVAIVDTDFITTQAFCIQYEGKAHPFLDSMIKEYPFDITILLNNNTKWVD

Av. paragallinarum Modesto nadR DGLRSLGSHKQRQRFQQLLKKLLDKYNVPYIEIESPSYLERYNKVKEIVEKILNEEELPE
Av. paragallinarum   1750  nadR DGLRSLGSHKQRQRFQQLLKKLLDKYNVSYIEIESPSYLERYNKVKEIVEKILNEEELPE

Av. paragallinarum Modesto nadR LIHDKLLFEE
Av. paragallinarum   1750  nadR LIHDKLLFEE
```

Figure 3.12. nadR protein comparison. The nadR from *Av. paragallinarum* NAD⁺-dependent strain Modesto was compared to the nadR from *Av. paragallinarum* NAD⁺-independent strain 1750 based on identity.

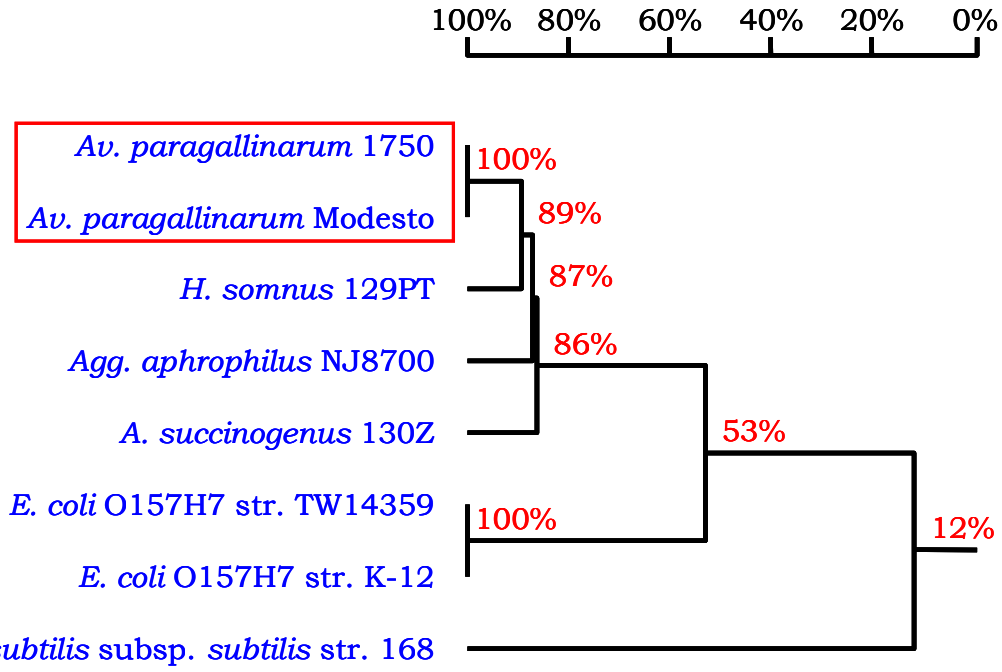


Figure 3.13. Homology tree of the *nadR* genes from various members of the *Pasteurellaceae*. *Av. paragallinarum nadR* genes are enclosed within a red box.

Av. paragallinarum nadR genes seem to closely resemble the *nadR* gene of *Haemophilus somnus* 129PT based on the homology tree in Figure 3.13.

A possible *pnuC* present on a contig 5 340 bp in size was implicated through the nucleotide-to-nucleotide comparison study (Figure 3.14). Primer set pnuC-F and pnuC-R was used to amplify a fragment of the implicated *pnuC* from both *Av. paragallinarum* NAD$^+$-dependent strain Modesto and *Av. paragallinarum* NAD$^+$-independent strain 1750. Following sequence confirmation, primer set pnuCcomp-F and pnuCcomp-R was used to amplify the complete ORF of *nadR* from *Av. paragallinarum* (Figure 3.14 and Figure 3.15).

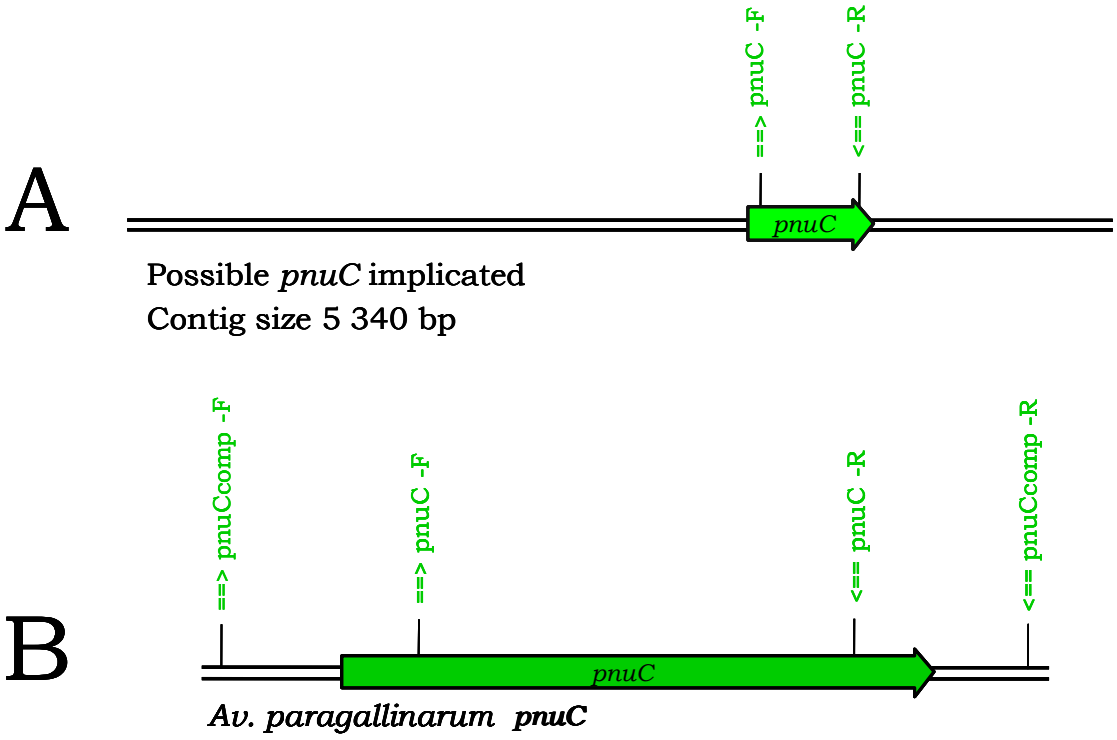


Figure 3.14. *pnuC* gene of *Av. paragallinarum*. Primers pnuC-F and pnuC-R (A) were used to confirm the presence of the implicated *pnuC* gene. Thereafter primers pnuCcomp-F and pnuCcomp-R (B) were used to amplify the complete *pnuC* ORF from *Av. paragallinarum*.

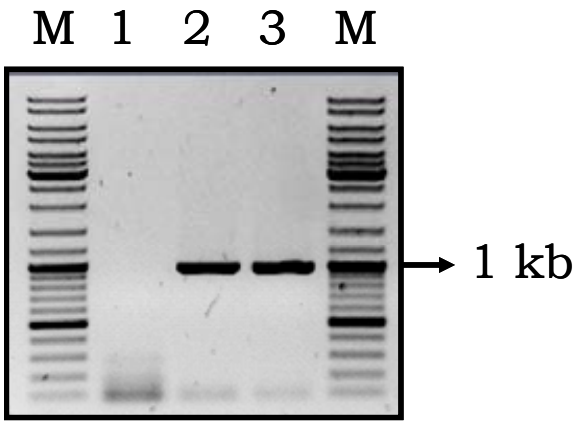


Figure 3.15. PCR amplification of the complete *pnuC* gene of *Av. paragallinarum* together with flanking regions where primer set pnuCcomp was used with an expected size of 975 bp. GeneRuler™ (Fermentas) was used as molecular weight marker (lane M) where lane 1 represents the negative control, lane 2 indicates the complete *pnuC* gene amplified from *Av. paragallinarum* NAD^+-dependent strain Modesto and lane 3 the complete *pnuC* gene amplified from *Av. paragallinarum* NAD^+-independent strain.

The *pnuC* genes from the NAD^+-dependent and NAD^+-independent strains of *Av. paragallinarum* both consists of 681 bp which translates into an ORF of 227 amino acids with a molecular weight of 57.9 kDa and a theoretical pI of 5.1. Conserved domain analysis identified the *pnuC* of *Av. paragallinarum* as part of the superfamily of integral membrane proteins that are involved in the transport of nicotinamide mononucleotide. The outer most N- and C-terminal regions of this protein are poorly conserved but a conserved region corresponding to most of the length of *pnuC* was identified and this region is mainly found in pathogens. A 99.6 % identity exists between the two genes from *Av. paragallinarum* if compared with one another on protein level (Figure 3.16).

```
Av. paragallinarum Modesto pnuC  MNLSQTLKNEFFGGWKPFEVVWLFLFIAAQVFAYVQSPVSWLEALA
Av. Paragallinarum   1750  pnuC  MNLSQTLKNEFFGGWKPFEVMWLFLFIAAQVFAYVQSPVSWLEALA

Av. paragallinarum Modesto pnuC  GISGILCVVFVGKG KISNYLFGLIFAYSYFYVSLDNKFYGEMTTT
Av. paragallinarum   1750  pnuC  GISGILCVVFVGKG KISNYLFGLIFAYSYFYVSLDNKFYGEMTTT

Av. paragallinarum Modesto pnuC  LYVYIPAQFIGYFLWKENMRNNVAQSETVIAKFLTAKGWTTLIALV
Av. paragallinarum   1750  pnuC  LYVYIPAQFIGYFLWKENMRNNVAQSETVIAKFLTAKGWTTLIALV

Av. paragallinarum Modesto pnuC  SLGSIAFISILKTTDGNSIGLDGVTTVLVIAAQLLMILRYSEQWIL
Av. paragallinarum   1750  pnuC  SLGSIAFISILKTTDGNSIGLDGVTTVLVIAAQLLMILRYSEQWIL

Av. paragallinarum Modesto pnuC  WIIINILSVILWAETPAMYLMYGAYLLNSLYGYYNWSKLAKNA
Av. paragallinarum   1750  pnuC  WIIINILSVILWAETPAMYLMYGAYLLNSLYGYYNWSKLAKNA
```

Figure 3.16. pnuC protein comparison. The pnuC from *Av. paragallinarum* NAD⁺-dependent strain Modesto was compared to the pnuC from *Av. paragallinarum* NAD⁺-independent strain 1750 based on identity.

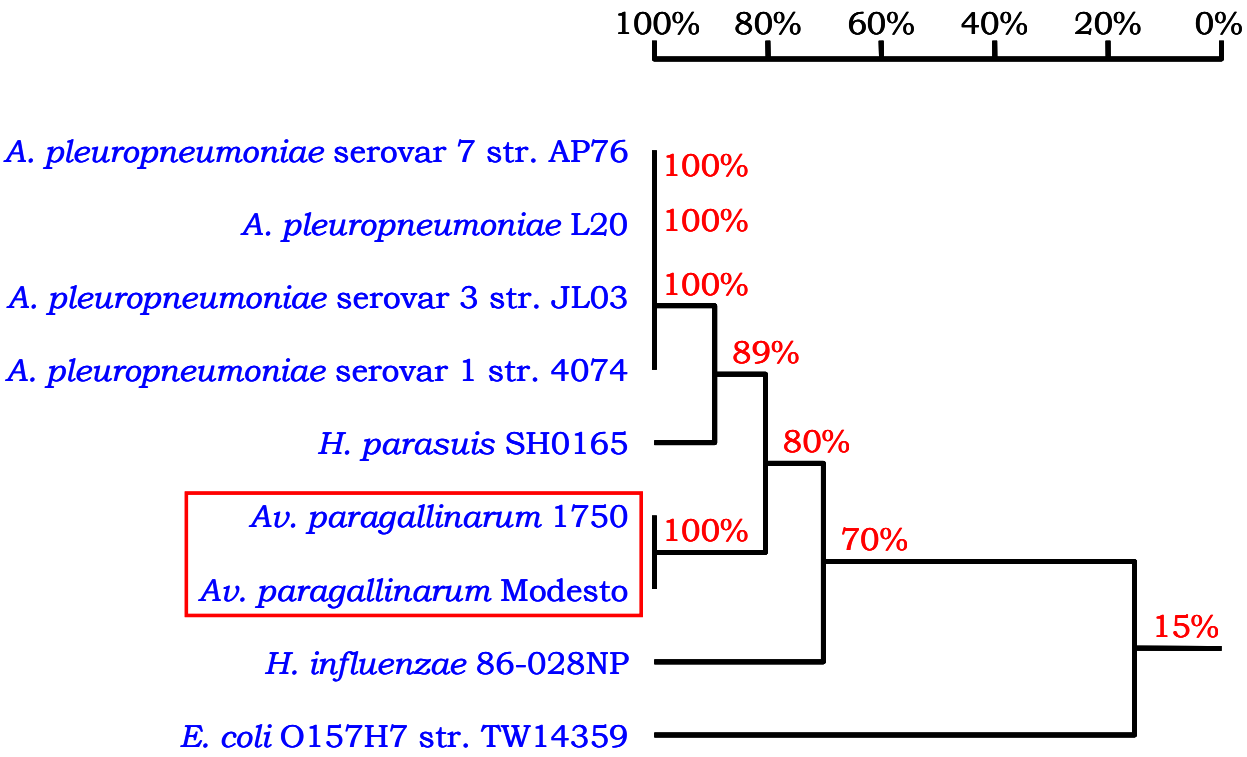


Figure 3.17. Homology tree of the *pnuC* genes from various members of the *Pasteurellaceae*. *Av. paragallinarum pnuC* genes are enclosed within a red box.

Based on a homology tree constructed of the *pnuC* genes from *Av. paragallinarum* and the *pnuC* genes from other members of the *Pasteurellaceae*, these genes show homology towards the *pnuC* gene of *H. parasuis* SH0165 (Figure 3.17).

The nucleotide-to-nucleotide comparison study implicated a possible *ppnK* (*nadF*) present on a contig 17 981 bp in size (Figure 3.18). Primer set ppnKcomp-F and ppnK-R was used to amplify a fragment of the implicated *ppnK* from both *Av. paragallinarum* NAD^+-dependent strain Modesto and *Av. paragallinarum* NAD^+-independent strain 1750. Following sequence confirmation primer set ppnKcomp-F and ppnKcomp-2R was used to amplify the complete ORF of *nadR* from *Av. paragallinarum* (Figure 3.18 and Figure 3.19).

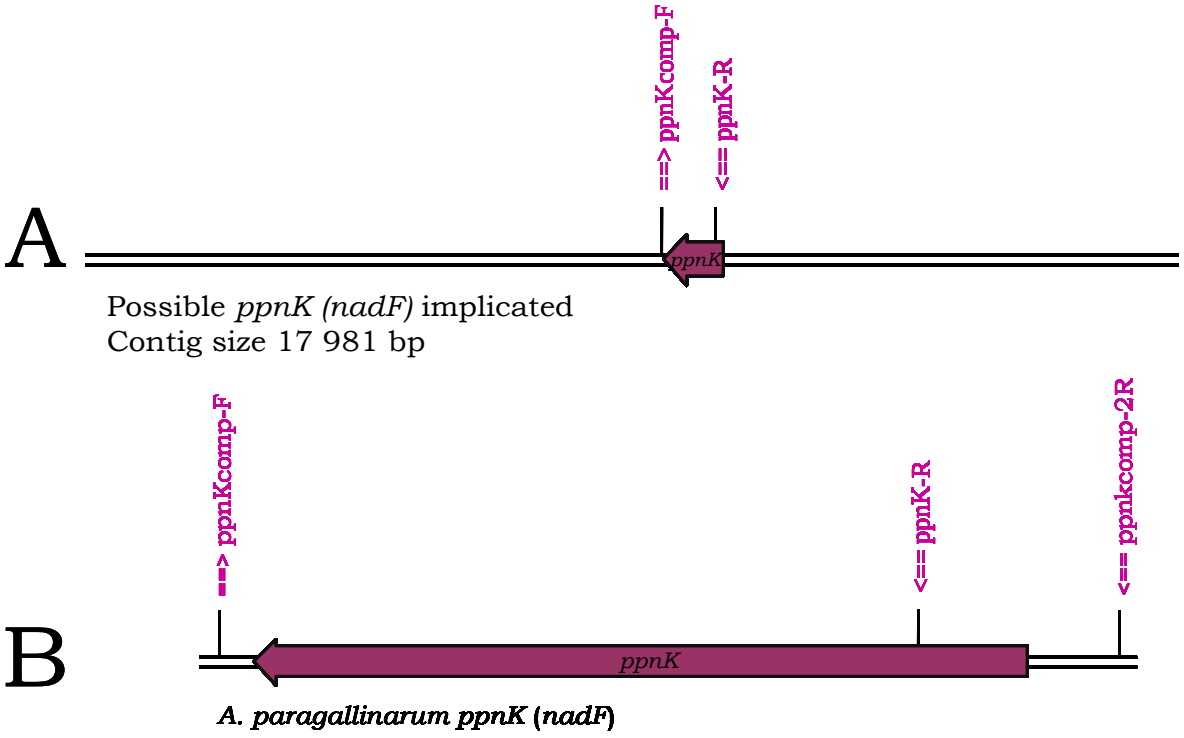


Figure 3.18. *ppnK* gene of *Av. paragallinarum*. Primers ppnKcomp-F and ppnK-R (A) was used to confirm the presence of the *ppnK* gene. Thereafter primers ppnKcomp-F and ppnKcomp-R (B) was used to amplify the complete *pnuC* ORF from *Av. paragallinarum*.

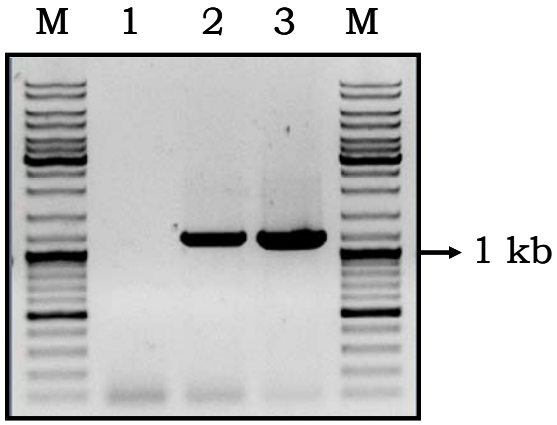


Figure 3.19. PCR amplification of the complete *ppnK* gene of *Av. paragallinarum* together with flanking regions. Primer set ppnKcomp was used with an expected size of 1 186 bp. GeneRuler™ (Fermentas) was used as molecular weight marker (lane M) where lane 1 indicates the negative control; lane 2 the complete *ppnK* gene amplified from *Av. paragallinarum* NAD^+-dependent strain Modesto and lane 3 the complete *ppnK* gene amplified from *Av. paragallinarum* NAD^+-independent strain 1750.

The *ppnK* genes of both the NAD^+-dependent and NAD^+-independent strains from *Av. paragallinarum* consists of 978 bp which translates into an ORF of 326 amino acids with a molecular weight of 82 kDa and a theoretical pI of 5.08. Conserved domain analysis revealed that the *ppnK* genes of *Av. paragallinarum* are capable of catalysing the conversion of NAD^+ to NADP. A diacylglycerol kinase catalytic domain was also identified. The diacylglycerol is a second messenger that acts as a protein kinase C activator. The catalytic domain is assumed from the finding of bacterial homologues and reveals an active site in the inter-domain cleft formed by four conserved sequence motifs, revealing a novel metal-binding site. The residues of this site are conserved across this family of proteins. When the two *Av. paragallinarum pnuC* genes are compared with one another a 98.7 % identity exists on protein level (Figure 3.20).

```
Av. paragallinarum Modesto ppnK   MCNIAANFLNSSAVHFSRIMTMSPNIIDNNTLHRSFQTIGLVGRPRNDVN
Av. Paragallinarum   1750  ppnK   MCNIAANFLNSSAVHFSRIMTMSPNIIDNNTLHRSFQTIGLVGRPRNDVN

Av. paragallinarum Modesto ppnK   LQMHKNLFQWLSEKGYQVLVEKPIGERLGLPAAHLAELEEIGRNAQLAIV
Av. paragallinarum   1750  ppnK   LQMHKNLFQWLSEKGYQVLVEKPIGERLGLPAEHLAELEEIGRNAQLAIV

Av. paragallinarum Modesto ppnK   IGGDGNMLGRARILAKYDIALIGINRGNLGFLTDIDPKNAYAQLAACLDN
Av. paragallinarum   1750  ppnK   IGGDGNMLGRARILAKYDIALIGINRGNLGFLTDIDPKNAYAQLAACLDN

Av. paragallinarum Modesto ppnK   GEFFVEERFLLETAIEREGQIIATSNAVNEVVIHPAKIAHMIDFHVYIND
Av. paragallinarum   1750  ppnK   GEFFVEERFLLETAIEREGQIIATSNAVNEAVIHPAKIVHMIDFHVYIND

Av. paragallinarum Modesto ppnK   KLAFSQRSDGLVVSTPTGSTAYSLSAGGPILTPQLNAIALVPMFPHTLSS
Av. paragallinarum   1750  ppnK   KLAFSQRSDGLIVSTPTGSTAYSLSAGGPILTPQLNAIALVPMFPHTLSS

Av. paragallinarum Modesto ppnK   RPLVIDGESKISIRFADYNTDQLEVGCDSQVALPFTPDDVVHIYKSQHKL
Av. Paragallinarum   1750  ppnK   RPLVIDGESKISIRFADYNTDQLEVGCDSQVALPFTPDDVVHIYKSQHKL

Av. paragallinarum Modesto ppnK   RLLHLKNYNYYNVLSTKLGWLRNSA
Av. paragallinarum   1750  ppnK   RLLHLKNYNYYNVLSTKLGWLRNSA
```

Figure 3.20. ppnK protein comparison. The ppnK from *Av. paragallinarum* NAD⁺-dependent strain Modesto was compared to the ppnK from *Av. paragallinarum* NAD⁺-independent strain 1750 based on protein identity.

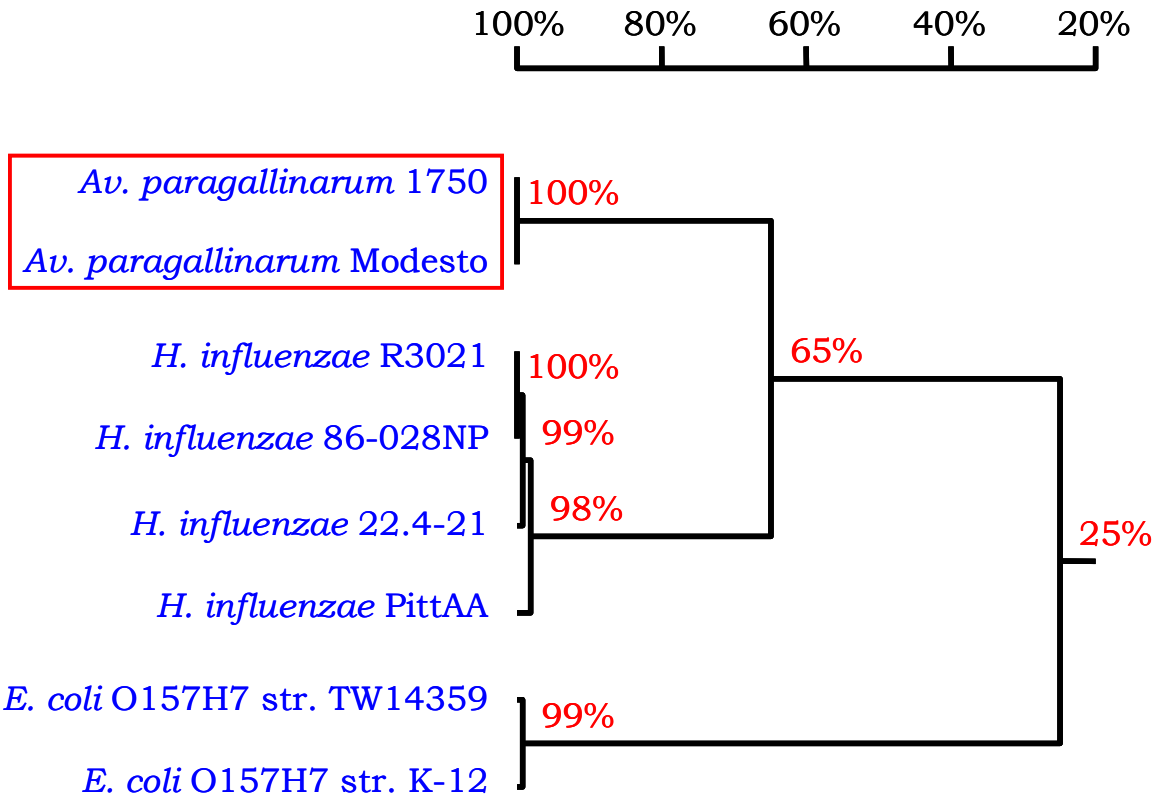


Figure 3.21. Homology tree of the *ppnK* genes from various members of the *Pasteurellaceae*. *Av. paragallinarum ppnK* genes are enclosed within a red box.

From Figure 3.21 it is clear that the *ppnK* genes of *Av. paragallinarum* show homology towards the *ppnK* genes of various *Haemophilus sp.*

The raw contig set from the titanium pyrosequencing run on the NAD$^+$-dependent *Av. paragallinarum* 0083 strain was compared to the genes obtained for the NAD$^+$-dependent pathway for *Av. paragallinarum* Modesto and the results proved to be identical as in the case of *Av. paragallinarum* Modesto.

3.5. Concluding remarks

With the use of data obtained from the whole genome sequencing project of *Av. paragallinarum* Modesto and *Av. paragallinarum* 0083, the genes forming the NAD$^+$-dependent or the NAD$^+$ salvage pathway of *Av. paragallinarum* were identified (Figure 3.22).

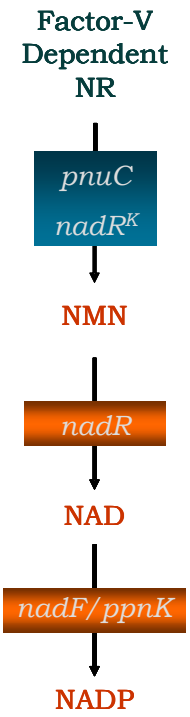


Figure 3.22. NAD$^+$-dependent pathway for *Av. paragallinarum*. This diagram illustrates all the genes identified in *Av. paragallinarum* responsible for the NAD$^+$-dependent pathway.

The NAD^+-dependent as well as the NAD^+-independent strains of *Av. paragallinarum* contains the pathway illustrated in Figure 3.22. The pnuC protein, encoded by the *pnuC* gene, is an integral membrane protein that is responsible for the transport of nicotinamide mononucleotide. The *nadR* gene expresses a bi-functional protein which serves as both a regulatory protein and is responsible for conversion of NMN to NAD^+ followed by the conversion of NAD^+ to NADH catalysed by the ppnK protein. The presence of these genes implies that *Av. paragallinarum* follows the classical route as proposed for the *Pasteurellaceae* by Gerlach and Reidl (2006) for either NAD^+-dependence or NAD^+-salvage.

Through genetic evolution the *Pasteurellaceae* seem to favour a shortened pathway for NAD^+-independence. This however is not the case for *Av. paragallinarum*. It was expected that a plasmid would be identified from NAD^+-independent strains carrying a gene or genes enabling NAD^+-independence within these strains. It seems that *Av. paragallinarum* may be unique within the *Pasteurellaceae* following the *de novo* synthesis pathway for the production of NAD^+ as set out for bacteria in general. It may also be speculated that we are currently witnessing an evolutionary shift within *Av. paragallinarum*; this organism may have already completed the shift towards a shortened pathway for NAD^+ recovery and may be shifting towards the shortened pathway for NAD^+-independence as well. An *nadC* and a partial *nadE* gene were identified for *Av. paragallinarum* NAD^+-independent strain 1750 from genomic DNA but not from plasmid DNA extracts in contrast to what was initially expected. The Quinolinic acid phosphoribosyltransferase (*nadC*) is responsible for the conversion of Qa to NaMN and the NAD^+ synthetase is responsible for the final conversion within the NAD^+ *de novo* pathway, converting nicotinate adenine dinucleotide to nicotinamide adenine nucleotide. However the presence of these proteins alone does not confer

NAD^+-independence, it only suggests less dependence on NAD^+ from the extracellular environment as illustrated in Figure 3.23.

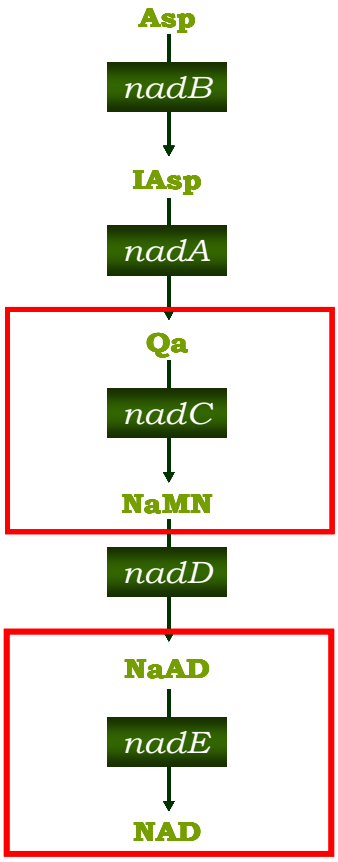


Figure 3.23. Possible NAD^+-independent pathway for *Av. paragallinarum*. The red box indicates the *nadC* gene that was identified for bacteria in the *de novo* NAD^+ synthesis pathway as set out in general for bacteria (Rodionov *et al.*, 2008 a,b).

The next step within the NAD^+ pathways for *Av. paragallinarum* would be to perform whole genome sequencing on a NAD^+-independent strain. That would be the only way to shed light on the enigmatic occurrence that is currently observed within *Av. paragallinarum*.

3.6. Literature cited

Blackall, P. J., and R. Yamamoto. 1990. Infectious coryza. In: A laboratory manual for the isolation and identification of avian pathogens, 3rd ed. H. G. Parchase, L. H. Arp, C. H. Domermuth and J. E. Pearson, eds. *American association of avian pathologists, Ames, Iowa.* Pp. 27-31.

Blanco, I., L. Galina-Pantoja, S. Oliveira, C. Pijoan, C. Sa´nchezd, A. Canals. 2004. Comparison between *Haemophilus parasuis* infection in colostrums-deprived and sow-reared piglets. *Vet. Microbiol.* **103**:21-27.

Bragg, R. R., L. Coetzee and J. A. Verschoor. 1993. Plasmid-encoded NAD^+ independence in some South African isolates of *Haemophilus paragallinarum*. *Onderstepoort J. Vet. Res.* **60**:147-152.

Challacombe, J. F., A. J. Duncan, T. S. Brettin, D. Bruce, O. Chertkov, J. C. Detter, C. S. Han, M. Misra, P. Richardson and R. Tapia. 2007. Complete Genome Sequence of *Haemophilus somnus* (*Histophilus somni*) Strain 129Pr and Comparison to *Haemophilus ducreyi* 35000HP and *Haemophilus influenzae* Rd. *J. Bacteriol.* **189**:1890-1898.

Chen, X., J. K. Miflin, P. Zhang and P. J. Blackall. 1996. Development and amplification of DNA probes and PCR tests for *Haemophilus paragallinarum*. *Avian Dis.* **40**:398-407.

Flachmann, R., N. Kunz, J. Seifert, M. Gütlich, F. Wientjes, A. Läufer and H. G. Gassen. 1988. Molecular biology of pyridine nucleotide biosynthesis in *Escherichia coli* – Cloning and characterization of quinolinate synthesis genes *nadA* and *nadB*. *Eur. J. Biochem.* **175**:221-228.

Foster J. W., D. M. Kinney and A. G. Moat. 1979. Pyridine Nucleotide Cycle of *Salmonella typhimurium* : Regulation of Nicotinic Acid Phosphoribosyltransferase and Nicotinamide Deamidase. *J. Bacteriol.* **138**:957-961.

Gerlach, G., J. Reidl. 2006. NAD^+ utilization in *Pasteurellaceae*: Simplification of a Complex Pathway. *J. Bacteriol.* **188**:6719-6727.

Hanahan, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166**:557-580.

Hong, S. H., J. S. Kim, S. Y. Lee, Y. H. In, S. S. Choi, J. K. Rih, C. H. Kim, H. Jeong, C. G. Hur and J. J. Kim. 2004. The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat. Biotechnol.* **22**:1275-1281.

Jacobsen, I., I. Hennig-Pauka, N. Baltes, M. Trost and G. F. Gerlach. 2005. Enzymes Involved in Anaerobic Respiration Appear to Play a Role in *Actinobacillus pleuropneumoniae* Virulence. *Infect. Immun.* **73**:226-234.

Kielstein, P., H. H. Wuthe, Ø. Angen, R. Mutters and P. Ahrens. Phenotypic and genetic characterization of NAD^+-dependent *Pasteurellaceae* from the respiratory tract of pigs and their possible pathogenetic importance. *Vet. Microbiol.* **81**:243-255.

Lei, L., C. Du., P. Yang, F. Xie, P. Ou, W. Han and J. Wang. 2009. Screening of strain-specific *Actinobacillus pleuropneumoniae* genes using a combination method. *J. Microbiol. Methods.* **77**:145-151.

Lima, W. C., A. M. Varani, C. F. M. Menck. 2009. NAD^+ Biosynthesis Evolution in Bacteria: Lateral Gene Transfer of Kynurenine Pathway in Xanthomonadales and Flavobacteriales. *Mol. Biol. Evol.* **26**:399-406.

Lin, H. 2007. Nicotinamide adenine dinucleotide: beyond a redox coenzyme. *Org. Biomol. Chem.* **5**:2541-2554.

Martin, P. R., R. J. Shea and M. H. Mulks. 2001. Identification of a Plasmid-Encoded Gene from *Haemophilus ducreyi* Which Confers NAD^+ Independence. *J. Bacteriol.* **183**:1168-1174.

Mendoza-Espinoza, A., Y. Koga and A. I. Zavaleta. 2008. Amplified 16S Ribosomal DNA Restriction Analysis for Ientification of *Avibacterium paragallinarum*. *Avian Dis.* **52**:54-58.

Møller, K., V. Fussing, P. A. D. Grimont, B. J. Paster, F. E. Dewhirst and M. Kilian. 1996. *Actinobacillus minor* sp. nov., *Actinobacillus porcinus* sp. nov., and *Actinobaillus indolicus* sp. nov., Three New V-Factor Independent Species from the Respiratory Tract of Pigs. *Int. J. Syst. Bacteriol.* **46**:951-956.

Morton, D. J., T. M. Van Wagoner, T. W. Seale, P. W. Whitby and T. L. Stull. 2008. Catalase as a source of both X- and V-factor for *Haemophilus influenzae*. *FEMS Microbiol. Lett.* **279**:157-161.

Mouahid, M., M. Bisgaard, A. J. Morley, R. Mutters and W. Mannheim. 1992. Occurrence of V-factor (NAD^+) independent strains of *Haemophilus paragallinarum*. *Vet. Microbiol.* **31**:363-368.

Munson, R. S., H. Zhong, R. Mungur, W. C. Ray, R. J. Shea, G. G. Mahairas and M. H. Mulks. 2004. *Haemophilus ducreyi* strain ATCC 27722 contains a genetic element with homology to the Vibrio RS1 element that can replicate as a plasmid and confer NAD^+ independence on *Haemophilus influenzae*. *Infect. Immun.* **72**:1143-1146.

Nørskov-Lauritsen, N. and M Kilian. 2006. Reclassification of *Actinobacillus actinomycetemcomitans*, *Haemophilus aphrophilus*, *Haemophilus paraphrophilus* and *Haemophilus segnis* as *Aggregatibacter actinomycetemcomitans* gen. nov., comb. nov., *Aggregatibacter segnis* comb. nov., and emended description of *Aggregatibacter aphrophilus* to include V factor-dependent and V factor-independent isolates. *Int. J. Syst. Evol. Microbiol.* **56**:2135-2146.

Ochman, H., A. S. Gerber and D. L. Hartl. (1988). Genetic Application of an Inverse Polymerase Chain Reaction. *Genetics.* **120**:621-623.

Rodionov, D. A., L. Xiaoqing, I. A. Rodionova, C. Yang, L. Sorci, E. Dervyn, D. Martynowski, H. Zhang, M. S. Gelfand and A. L. Osterman. 2008a. Transcriptional regulation of NAD^+ metabolism in bacteria : genomic reconstruction of *NiaR* (*YrxA*) regulon. *Nucleic Acids Res.* **36**:2032-2046.

Rodionov, D. A., J. De Ingeniis, C. Mancini, F. Cimadamore, H. Zhang, A. L. Osterman and N. Raffaelli. 2008b. Transcriptional regulation of NAD^+ metabolism in bacteria : *NrtR* family of Nudix-related regulators. *Nucleic Acids Res.* **36**:2047-2059.

Sambrook, J., E. T. Fritsch and T. Maniatis. 1989. Molecular cloning: A laboratory manual. 2nd edition. *Cold spring harbor laboratory press. Cold Spring Harbor N. Y.*

Terry, D. T., Y. M. Zalucki, S. L. Walsh, P. J. Blackall and M. P. Jennings. 2003. Genetic analysis of a plasmid encoding haemocin production in *Haemophilus paragallinarum*. *Microbiology.* **149**:3177-3184.

Van Zyl, A. E. 2003. Characterization of a plasmid conferring NAD^+ independence in *Haemophilus paragallinarum*. M.Sc. thesis. University of the Free State, Bloemfontein, South Africa.

Windsor, H. M., R. C. Gromkova and H. J. Koornhof. 1993. Growth Characteristics of V Factor-Independent Transformants of *Haemophilus influenzae*. *Int. J. Syst. Bacteriol.* **43**:799-804.

Xu, Z., Y. Zhou, L. Li, R. Zhoul, S. Xiao, Y. Wan, S. Zhang, K. Wang, W. Li, L. Li, H. Jin, M. Kang, B. Dalai, T. Li, L. Liu, Y. Cheng, L. Zhang, T. Xu, H. Zheng, S. Pu, B. Wang, W. Gu, X. L. Zhang, G. F. Zhu, S. Wang, G. P. Zhao and H. Chen. 2008. Genome Biology of *Actinobacillus pleuropneumoniae JL03*, and Isolate of Serovar 3 Prevalent in China. PLoS ONE **3**(1): e1450. doi:10.1371/journal.pone.0001450.

Chapter 4
Whole genome sequencing of *Avibacterium paragallinarum*

4.1. Abstract

The whole genome of a NAD^+-dependent *Avibacterium paragallinarum* strain was sequenced with the use of pyrosequencing technology. A pyrosequencing run was performed on a GS-20 instrument initially, with an average read length of 100 bp and resulted in a 11 X sequence coverage. Analysis of this data set produced 787 contigs of which the smallest was 85 bp long and the largest 43 928 bp. A second pyrosequencing run was performed on a GS-FLX Titanium instrument with an average read length of 400 bp resulting in a 27 X sequence coverage. The Titanium run delivered 300 contigs and with the use of contra-mapping the total contig size could be decreased from 300 to 268. Contigs were submitted for annotation which in turn suggests that *Av. paragallinarum* contains multiple proteins involved in the cell envelope together with mobile and extrachromosomal element functions.

4.2. Introduction

Pyrosequencing technology has recently emerged as the preferred method to perform high-throughput DNA sequencing. The technology, developed by 454 Life Sciences (currently owned by Roche Diagnostics), initiated the next-generation movement by exposing three bottlenecks faced by the research community which included library preparation; template preparation and sequencing (Rothberg and Leamon, 2008).

The basis for miniaturizing was provided by the concept of sequencing through synthesis. The two opposing technologies available were cyclic reversible termination based on sequencing by sequential addition of labelled bases followed by fluorescent detection and cleavage of that fluorescent base on the one hand and on the other was detection of pyrophosphate release with an enzymatic cascade ending in luciferase and the detection of emitted light. The second approach was chosen for the 454 platform because direct incorporation of natural nucleotides was more appealing than repeated cycles of incorporation, detection and cleavage. 454 sequencing concluded with a pyrosequencing reaction where both the template preparation step and the pyrosequencing chemistry moved to the solid phase (Rothberg and Leamon, 2008).

Pyrosequencing is based on the detection of light whenever a nucleotide is incorporated therefore no physical separation process like electrophoresis are required to resolve the next base in the DNA strand. This meant that, unlike electrophoresis, pyrosequencing could be reduced to any reaction volume that generates detectable levels of light. This in turn enabled sequencing to be performed in parallel unlike the usual sequential route. The sequencing template and required enzymes are immobilized on a solid support instead of occurring free in solution and the sequencing templates are physically

separated by discrete wells of a picotiter plate as can be seen in Figure 4.1 (Rothberg and Leamon, 2008).

Figure 4.1. Optimized bead deposition within the picotitre plate. The size of the wells along with the size of the beads are optimized in such a way that only one library bead fits into one well along with enzyme beads for the sequencing reaction and packing beads for holding each bead within a well (Mardis, 2008).

Another crucial obstacle was supplying reaction reagents to the large quantities of simultaneous reactions. Each reaction requires a constant source of fresh reagents alongside each base addition. An elegant solution was provided in a laminar flow concept in which diffusion from a laminar flow stream itself is sufficient to bring fresh reagents to each reaction and to wash both by-products and unincorporated reagents away. Higher densities of reactions per cm^2 were accomplished by placing the reactions in wells in order to sequence tens or hundreds of thousands of reactions. In addition, well depth further isolate individual reactions from one another (Rothberg and Leamon, 2008; Mardis, 2008; Shendure and Ji, 2008).

The next problem to overcome was imaging and this was solved by adapting commercially available astrological grade cameras that employed fibre-optic bundles physically attached to the surface of a charge-couple device (CCD). These bundles are then narrowed to allow a large image area to be projected onto a smaller CCD surface (Rothberg and Leamon, 2008).

454 was able to manufacture a plate with millions of wells which would serve as individual reactors allowing separate enzymatic reactions to take place in each of the individual wells by etching the thin slides from a fibre-optic block. This can be achieved in two steps; first the fibre-optic bundles are sliced into thin disposable slides that resemble microscope slides, thereafter, wells are etched on one side of the slides by removing the glass between the fibre-shell with acid (Rothberg and Leamon, 2008).

Once the plate is placed in a flow cell allowing delivery of the reagents this fibre-optic slide would serve as a monolithic substrate for many reactions in parallel where the individual wells limit diffusion and thereby allowing individual sequencing reactions to take place without interference from their neighbours and further allowing the light from the reactions to be captured through placing the slide on top of a fibre-optic bundle permanently attached to a high-end imaging CCD system (Rothberg and Leamon, 2008; Mardis, 2008; Shendure and Ji, 2008).

The development of the 454 pyrosequencing technology resulted in the establishment of commercial pyrosequencing practices. In this chapter the application of the 454 pyrosequencing technology in order to reveal the genome sequence of *Avibacterium paragallinarum* will be discussed.

4.3. Materials and Methods

4.3.1. Bacterial strains and growth conditions

The NAD^+-dependent *Avibacterium paragallinarum* strain C-2 (Modesto) was obtained from *Onderstepoort Biological Products*, Onderstepoort, South Africa. This bacterium was cultivated in TM/SN medium which contains 1%

(m/v) biosate peptone, 1% (m/v) NaCl, 0.1% (m/v) starch and 0.05% (m/v) glucose with oleic acid-albumin complex, 1% (v/v) chicken serum, 1% (v/v) NAD^+ and 0.0005% (m/v) thiamine solution as supplements under limiting oxygen conditions (Blackall and Yamamoto, 1990). Agar (1.7%) was used to solidify the above mentioned media with cultivation in a candle jar at 37˚C for 16 h.

4.3.2. Enzymes, chemicals, kits and other consumables

Primers were obtained from Inqaba Biotechnical Industries. Bioflux DNA purification kits were obtained from Bioer Technology. Polymerase chain reaction (PCR) reagents were obtained from New England Biolabs® Inc (NEB).

Chemicals used during this study were analytical or molecular biology grade, used without further purification. All relevant chemicals were obtained from Sigma-Aldrich or Merck except NAD^+, bovine serum albumin fraction V along with proteinase K which was obtained from Roche Molecular Biochemicals.

Qiagen® DNA extraction kit was used for genomic DNA isolation from *Av. paragallinarum* and used according to the manufacturer's recommendations.

Consumables and kits used during 454 pyrosequencing were obtained from Roche Applied Sciences.

4.3.3. Techniques applied

4.3.3.1. DNA extraction

Genomic DNA extraction was performed with the use of the Qiagen® DNA extraction kit. *Av. paragallinarum* cells were harvested from 20 h old TM/SN plates. From hereon the manufacturer's recommendations were followed.

4.3.3.2. Identification of *Av. paragallinarum*

With the use of a PCR assay based on species-specific DNA probes *Av. paragallinarum* was identified using primers AP-F and AP-R (Chen *et al.*, 1996). The reaction cycle included an initial denaturation of 10 min at 94˚C applied to whole cell PCR in order to break to cells to enable DNA release after which the 2 units *Taq* polymerase was added or 2 min at 94˚C otherwise. This was followed by 30 cycles of denaturation (94˚C for 25 sec); annealing (55˚C for 50 sec); elongation (72˚C for 45 sec) and a final elongation cycle of 7 min at 72˚C with an expected amplification result of approximately 500 bp.

16S rDNA amplification according to Mendoza-Espinoza and co-workers (2007) was done on genomic DNA in order to confirm that there were no foreign, other than *Av. paragallinarum*, genomic DNA present in extracts. Primer pair 27F and 1492R was used to amplify the 16S rDNA region. The reaction cycle included an initial denaturation of 2 min at 94˚C followed by 35 cycles of denaturation (94˚C for 30 sec); annealing (49˚C for 45 sec); elongation (72˚C for 1 min 30 sec) and a final elongation cycle of 10 min at 72˚C with an expected amplification result of approximately 1 500 bp.

4.3.3.3. Polymerase chain reactions

All PCR reactions were performed with the use of NEB *Taq* DNA polymerase according to the manufacturer's recommendations. The reaction mixture contained 5 µl of the 10 x ThermoPol buffer, 2 µM upstream and downstream primers, 200 µM dNTP's, 0.5 µg template DNA and 2 units *Taq* polymerase filled to a final reaction volume of 50 µl using sterile redistilled water.

Table 4.1: Oligonucleotide primers used in this study.

Primer	5'→3' Orientation	Application	Melting temp.
AP-1F	TgA ggg TAg TCT TgC ACg CgA AT	ID PCR	64.55˚C
AP-1R	CAA ggT ATC gAT CgT CTC TCT ACT	ID PCR	62.86˚C
27F	AgA GTT TgA TCM Tgg CTC Ag*	16S PCR	52.2 - 54.3 °C
1492R	TAC ggY TAC CTT gTT ACg ACT T*	16S PCR	53.5 - 55.6 °C

* Degenerate nucleotides: M = A/C; Y = C/T.

In order to avoid confusion between guanine and cytosine within the oligonucleotides guanine is indicated as lowercase.

4.3.3.4. Sequencing of the 16S rDNA region

In order to determine the nucleotide composition of the 16S rDNA region, purified positive templates were used in sequencing reactions. Sequencing reactions were performed using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, USA) along with the 3130*xl* Genetic Analyzer (Applied Biosystems, USA) according to the manufacturer's recommendations. Sequences were assembled using Geneious 4.7.4 from Biomatters.

4.3.3.5. 454 Pyrosequencing

The Genome Sequencer 20 System and the Genome Sequencer Titanium System from Roche Applied Sciences was employed for whole genome sequencing of *Av. paragallinarum*. The GS-20 System services were provided by Inqaba Biotechnical Industries (Pty) Ltd, South Africa and the Titanium System by Agowa Genomics, Germany.

The first step in the pyrosequencing process is DNA library preparation. The double stranded genomic DNA is broken into fragments with the use of high pressure nitrogen. This process is referred to as nebulization. The genomic DNA is fragmented to obtain a size range between 300 – 800 bp. These fragments are then purified, blunt-ended and phosphorylated. These steps are illustrated in Figure 4.2.

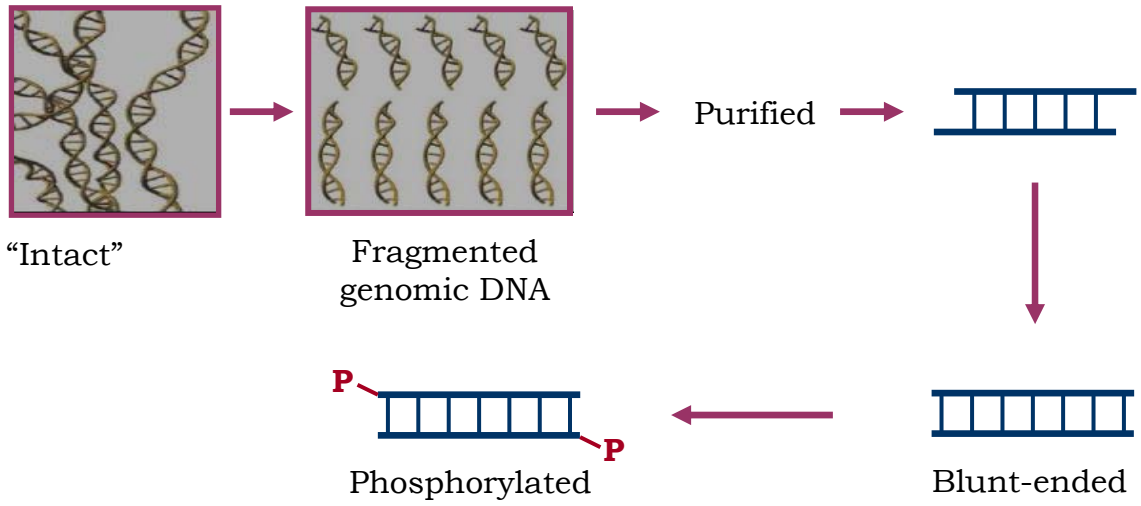


Figure 4.2. DNA library preparation. Genomic DNA is fragmented, purified, blunt-ended and phosphorylated (Adapted from Roche Applied Sciences).

Thereafter, two short adaptors, A and B, are ligated to the fragments one on each end. These adaptors provide priming sites during amplification and

sequencing where they provide the key sequence during sequencing as well. The key sequence is a short stretch of 4 nucleotides in a specific order that is used by the system to calibrate base calling and to recognize legitimate library fragments. The B adaptor contains a 5' biotin tag for selective mobilization. As a result 4 types of products are produced. Two types with one A and one B adaptor, one type with two A adaptors and the last type with 2 B adaptors. The formation of these types can be observed in Figure 4.3.

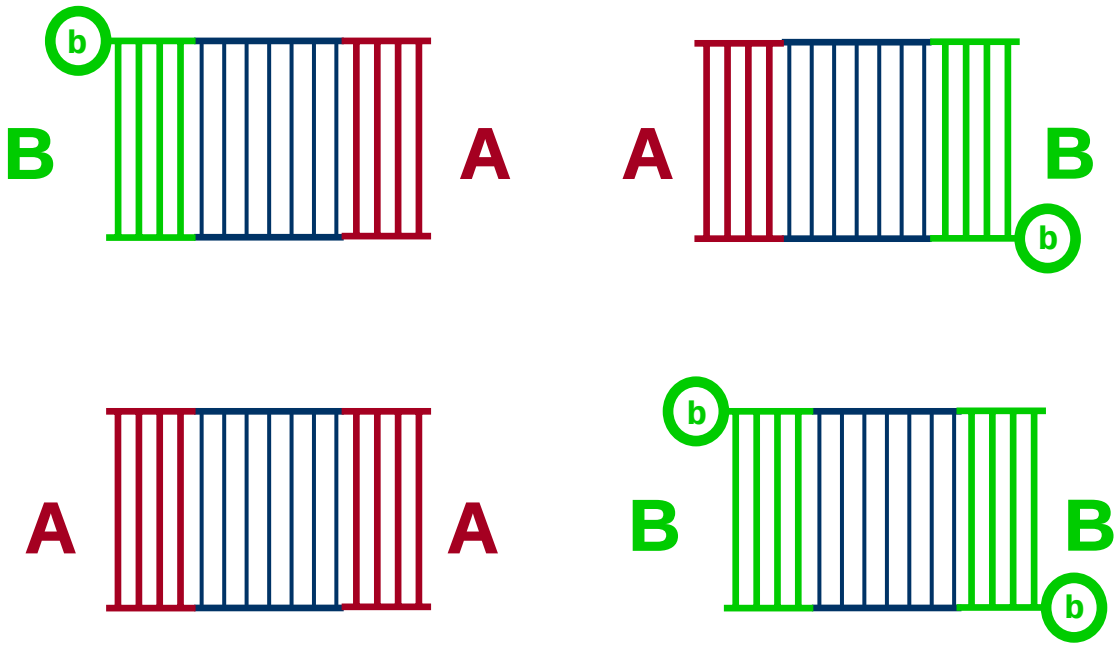


Figure 4.3. The formation of the four types of DNA fragments. Two fragment types containing A and B adaptors, one fragment type only A adaptors and the last type of fragment only B adaptors (Adapted from Roche Applied Sciences).

The reason for the attachment of these adaptors is to collect fragments containing only one A and one B adaptor. In order to obtain only fragments containing both a A and B adaptor a magnetic bead coated with streptavidin is used. The biotin tag on the B adaptor has the ability to bind streptavidin very tightly. This property in turn is then used during selective mobilization. The biotin on the B adaptor binds to the streptavidin and fragments containing

only A adaptors are washed away. Fragments containing two B adaptors are bound to the streptavidin on the magnetic bead. With the addition of NaOH double stranded DNA fragments attached to the beads are denatured and as a result single stranded DNA fragments containing only **one** adaptor A and **one** adaptor B are released from the beads. These will form the single stranded template DNA for sequencing.

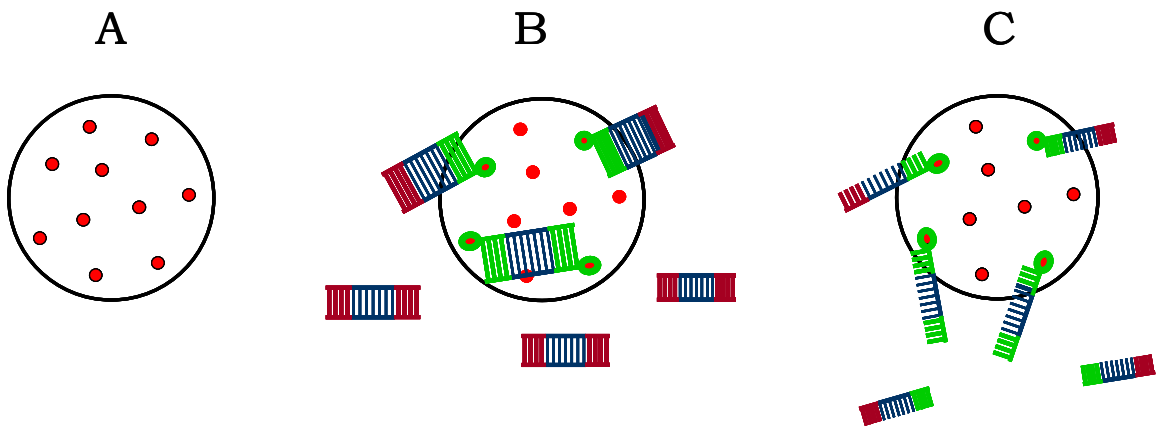


Figure 4.4. The production of single stranded DNA containing A-B adaptors. A magnetic bead coated with streptavidin (diagram A) is used to bind fragments containing B, green, adaptors (diagram B). Adaptors containing only A adaptors are washed away. NaOH are used to denature double stranded DNA fragments, fragments labelled with only B adaptors stay bound to the magnetic bead and fragments containing A-B adaptors can now be separated from the beads (diagram C) and are employed in downstream applications (Adapted from Roche Applied Sciences).

DNA is amplified through emulsion based polymerase chain reaction, emPCR. This step eliminates expensive and time-consuming processes of cloning and colony picking. The single stranded template DNA are purified and quantified in order to determine the concentration of single stranded template DNA required for amplification. Thereafter the single stranded template DNA fragments are immobilized onto primer coated capture beads through hybridization. The forward primer is biotin labelled which is used during the

enrichment step in which one fragment is amplified to produce many copies of the captured fragment. This process is optimized to produce beads where a single library fragment is bound onto each bead and the end result is one library fragment bound to one bead which in turn would lead to **ONE READ**. Figure 4.5 illustrates the primer coated capture beads.

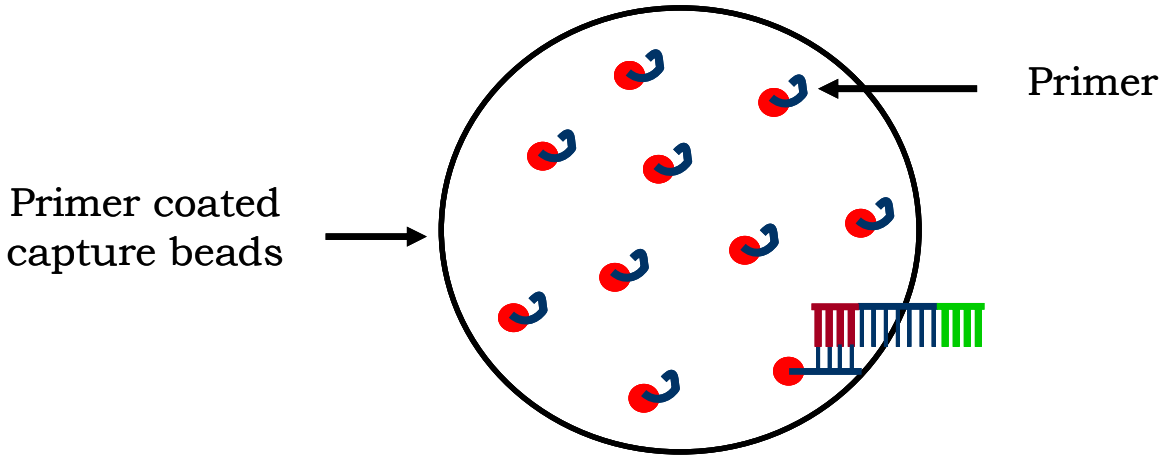


Figure 4.5. Primer coated capture beads. Single stranded template DNA fragments are immobilized onto primer coated capture beads (Adapted from Roche Applied Sciences).

A water-in-oil-emulsion is created containing the bead library along with the amplification reagents, Figure 4.6. A single fragment is within its own emulsion micro-reactor along with amplification reagents. Within each micro-reactor independent clonal amplification takes place. All micro-reactors are processed simultaneously resulting in millions of beads each with a unique clonally amplified fragment.

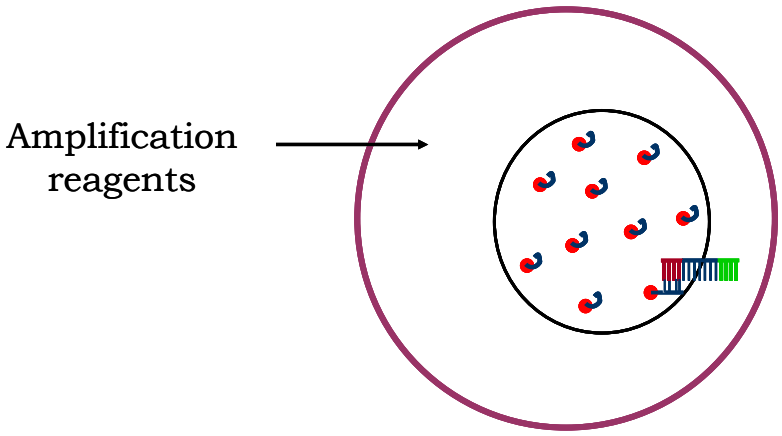


Figure 4.6. Water in oil emulsion micro-reactors. Within each micro-reactor there are sufficient reagents for amplification alongside a single fragment attached to one bead for amplification (Adapted from Roche Applied Sciences).

Following amplification, the micro-reactors are broken thereby releasing the DNA positive beads. All of the double stranded DNA fragments contained on a bead are identical. Streptavidin coated magnetic enrichment beads are added to capture, and thereby purify, successfully amplified fragments, Figure 4.7. The streptavidin enrichment beads only bind to the amplified fragments that incorporated the biotinylated primer. Single stranded DNA fragments are released through the addition of NaOH. As a result single stranded DNA template are now enriched in carrying beads and the amplified DNA library is ready for sequencing.

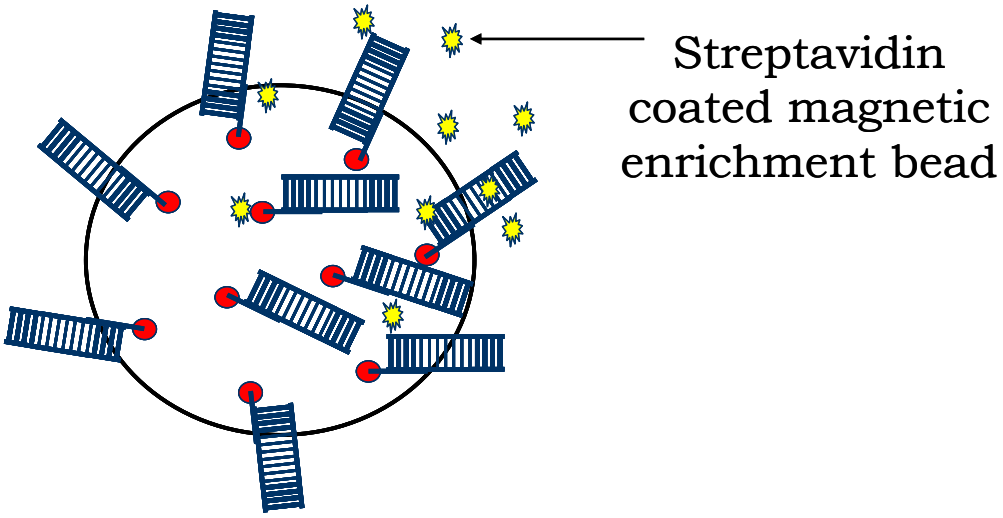


Figure 4.7. The enrichment process. Library beads are captured with the use of streptavidin coated magnetic enrichment beads (Adapted from Roche Applied Sciences).

Preparation for sequencing entails a sequencing primer annealed to the single stranded DNA fragments. The single stranded template DNA fragments, attached to beads, are mixed with polymerase and co-factor after which they are layered onto a pico-titer plate, followed by enzyme beads and packing beads. The deposition process is optimized in such a way that only one bead will be deposited within a well. Sequencing is performed simultaneously in all the wells. Figure 4.8 illustrates a summary of the sequencing process. There is a continuous overflow of reagents containing buffers and nucleotides over the pico-titer plate. The first thing the system does is look for the key sequence TACG employed by the A and B adaptors, if this sequence is not present in this order that specific well will not be read. The enzyme beads contain sulfurylase and luciferase and when a nucleotide is complimentary and as a result incorporated a pyrophosphate molecule is released. The pyrophosphate molecule is converted to ATP by sulfurylase using adenosine phosphosulfate. Luciferase hydrolyzed the ATP using luciferon producing oxy–luciferon and light. The light in turn is detected by the CCD camera attached at the base of the pico-titer plate. Signal intensity indicates incorporation and

the intensity is in turn directly proportional to the number of nucleotides incorporated.

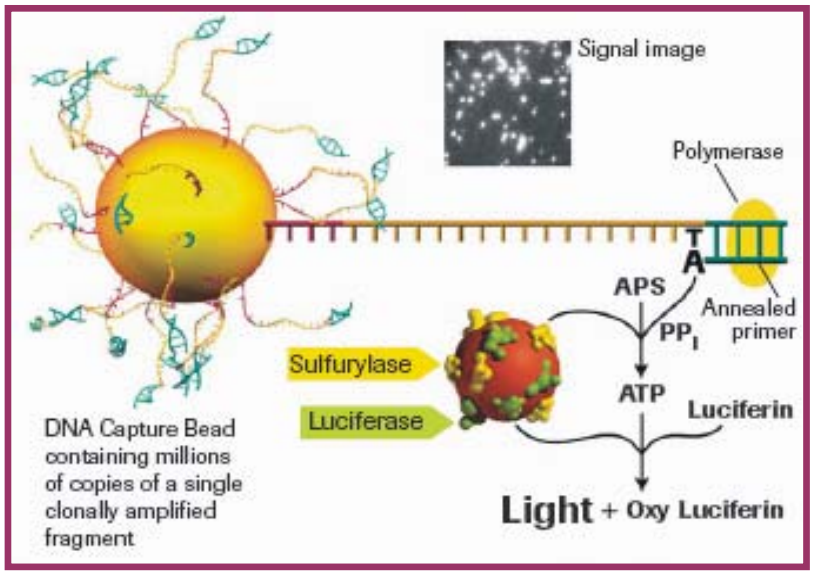


Figure 4.8. Sequencing of library beads. This diagram illustrates the bead-library being sequenced with the sequential enzymatic reactions producing a light signal which in turn is detected by the CCD camera (Roche Applied Sciences).

There are a few differences between the Genome Sequencer 20 System and the Genome Sequencer Titanium System with the most significant the read length each system can deliver. The Genome Sequencer 20 System delivers up to 100 bp read lengths whilst the Genome Sequencer Titanium System delivers up to 400 bp read lengths. The other differences include series reagents to support the extended read length (Roche Applied Sciences).

4.4. Results and Discussion

4.4.1. DNA extraction

In order to perform pyrosequencing at least 3 – 10 µg of very pure, high-grade unsheared genomic DNA is required. For this purpose genomic DNA was extracted from multiple samples of *Av. paragallinarum* NAD$^+$-dependent strain Modesto (Figure 4.9).

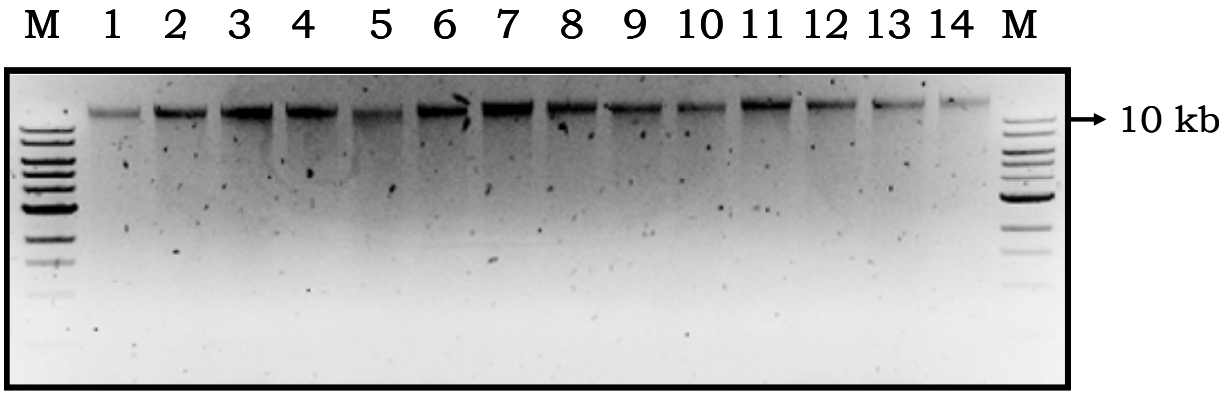


Figure 4.9. Genomic DNA extraction of *Av. paragallinarum* NAD$^+$-dependent strain Modesto. NEB 1 kb ladder (lane M) was used as molecular marker.

4.4.2. Identification of *Av. paragallinarum*

Av. paragallinarum identification PCR's were performed on cultures prior to genomic DNA extraction using whole cells as well as on genomic DNA extracted using primer pair AP-F and AP-R. The expected 500 bp fragment can be seen in Figure 4.10 for all 14 samples.

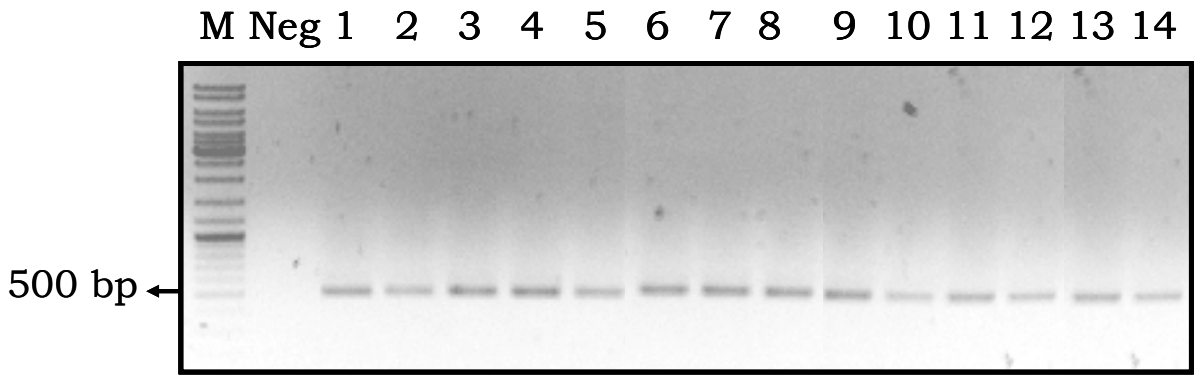


Figure 4.10. *Av. paragallinarum* identification through PCR. In the above diagram M indicates the molecular marker; lane Neg indicates the negative control and lane 1 – 14 indicates the different genomic DNA samples. GeneRuler™ (Fermentas) was used as molecular marker.

4.4.3. Sequencing of the 16S rDNA region

The 16S rDNA region of *Av. paragallinarum* was amplified using primer pair 27F and 1492R from each of the 14 genomic DNA extracts. These PCR products were directly sequenced without any sub-cloning. This method was applied in order to confirm the purity of the extracted genomic DNA. Sequence analysis indicated that none of the samples showed the presence of any foreign prokaryotic genomic DNA.

4.4.4. 454 Pyrosequencing

Genomic DNA samples were sent to Inqaba Biotechnical Industries (Pty) Ltd, South Africa for a ¼ plate GS-20 pyrosequencing run. The GS-20 pyrosequencing read length is up to 100 bp, it should be noted that this however does not mean that all reads will be 100 bp but it only indicates that the instrument's maximum read length is 100 bp. Based on close relatives of *Av. paragallinarum* whose genome sequencing projects have been completed namely: *Haemophilus influenzae* 86-028NP genome size of 1.9 Mb; *Pasteurella*

multocida subsp. *multocida* str. Pm70 genome size 2.2 Mb and *Aggregatibacter aphrophilus* NJ8700 genome size 2.2 Mb it can be speculated that the genome size of *Av. paragallinarum* should be between 1.9 and 2.2 Mb. The data obtained from Inqaba Biotechnical Industries (Pty) Ltd., showed that roughly 23.4 million nucleotides were read by the instrument. Working with the upper limit of a probable genome size of 2.2 Mb for *Av. paragallinarum* the amount of nucleotides sequenced could cover the genome of *Av. paragallinarum* 11 times therefore a 11 X coverage. This however does not mean that every base within the genome was read 11 times. The 234 000 reads were assembled by Inqaba Biotechnical Industries (Pty) Ltd, South Africa using the Newbler 1.0.53.17 software from the 454 Life Sciences Corporation. An overlap minimum match identity of 90% was used and this resulted in 787 contigs of which the smallest was 85 bp long and the largest 43 928 b. The excessive amount of contigs obtained was not expected. In order to obtain a good assembly for a genome with the GS-20 sequencer a 20 X coverage is required which was not achieved in this instance.

Since the initial GS-20 pyrosequencing was performed the technology was improved with firstly the release of the GS-FLX in January 2007 followed by the GS-FLX Titanium kit in October 2008. The latter allowed an increase read length of up to 400 bp. The only difference in technology between the GS-FLX and the GS-FLX Titanium are the reagents used during the procedure.

Genomic DNA samples were sent to Agowa Genomics, Germany for a ¼ plate GS-FLX Titanium pyrosequencing run. The total number of nucleotides read was 6.03 million which resulted in a higher coverage of 27 X when working with the expected upper limit genome size. The total number of reads were 160 743 and it resulted in an average read length of 375 bp. The reads were assembled using Newbler 2.0.01.14 from 454 Life Sciences Corporation with an overlap minimum match identity of 90% and an overlap minimum match

length of 40 bases. The total number of contigs obtained were 300 with the largest contig 78 465 bp and the average contig size longer than 500 bp were 10 727 bp.

Initially the GS-20 contigs were mapped to the Titanium contigs with a minimum overlap of 25 bp and a high alignment identity with the CLC Main Workbench 5.2. Several discrepancies were observed during the mapping and an example of this is illustrated in Figure 4.11.

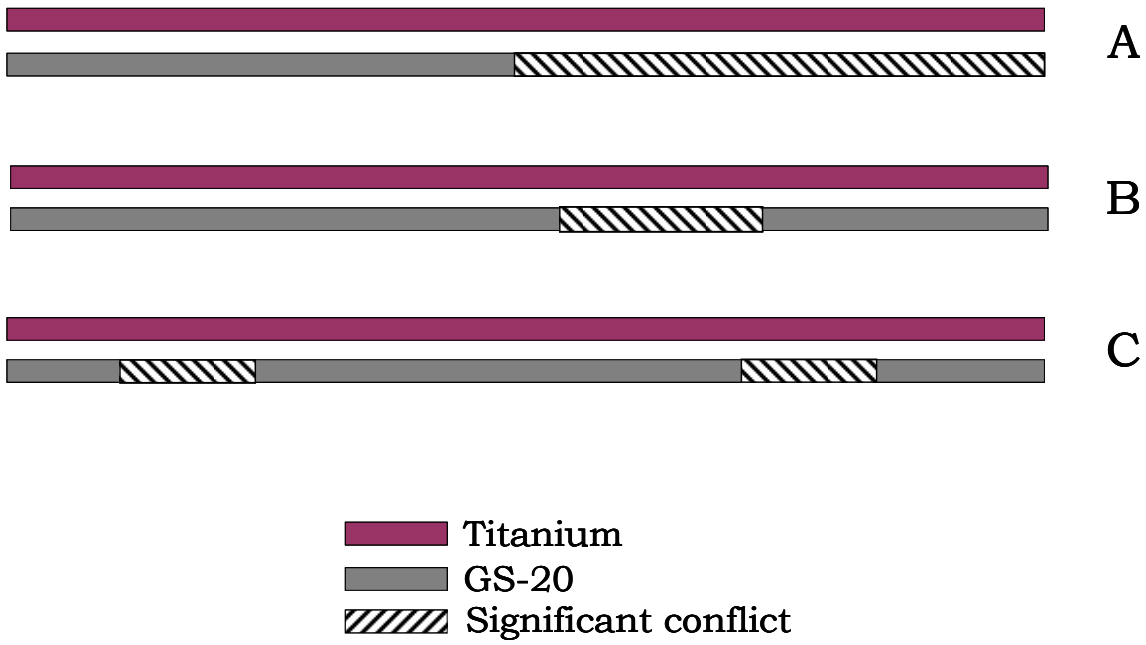


Figure 4.11. GS-20 and GS-FLX Titanium mapping conflicts. Various types of conflicts are illustrated in A where half of a contig align 100% and half show significant conflict. In B most of the contig align 100% with a region showing significant conflict and in C there is more than one region showing significant conflict.

Some of the GS-20 contigs mapped 100% to the Titanium contigs but in other cases half of a GS-20 contig fit perfectly to a Titanium contig whereas the other half did not fit and in other cases one or more regions of the GS-20 contig did not fit whereas other regions of the GS-20 contig fit perfectly to the Titanium

contig (Figure 4.11). From this it was concluded that the GS-20 data could not be used individually as a result of the low coverage.

Figure 4.12 illustrates the relationship between read length and coverage. When obtaining GS-20 short read lengths more reads are required in order to assemble a single long contig due to the short fragments. On the other hand with the GS-FLX Titanium longer reads, less reads are required to assemble in a similar size contig.

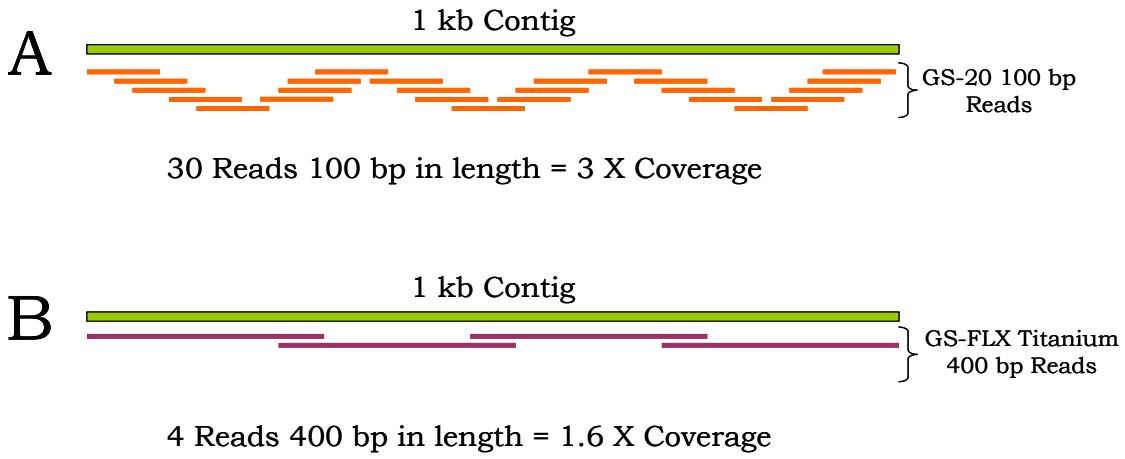


Figure 4.12. The relationship between read length and coverage. In this diagram it can be seen that the shorter the reads, as in the case of A the higher the required coverage. On the other hand when working with longer reads as in B less reads are required to cover a contig therefore a lower coverage.

With the GS-20 run, 12 085 reads were implicated to form part of repeated regions, this figure was however dramatically decreased with the longer GS-FLX Titanium reads where only 162 reads were deemed to form part of repeated regions. From the above mentioned results it was clear that *Av. paragallinarum* contain numerous repeat regions, with some repeat regions as small as 48 bp. Analysing a genome sequence containing this amount of repetitive regions, shorter reads (as in the case with the GS-20 data) will not suffice.

Another technique that is advised to be used in combination with pyrosequencing is paired-end sequencing. This technique are used to determine the orientation and relative positions of contigs produced by *de novo* shotgun sequencing and assembly. This technique allow for gap closure between contigs as well. However, this technique will not be able to solve problems encountered within *Av. paragallinarum* seeing that a longer read-lenth is needed for effective paired-end analysis within *Av. paragallinarum* (Personal communication, Roche Diagnostics).

The next step was therefore to investigate the probability of using a joint assembly between the GS-20 and the GS-FLX Titanium reads. The reads from the GS-20 run as well as the reads from the GS-FLX Titanium run were used in an assembly using Newbler 2.0.01.14 from 454 Life Sciences Corporation with an overlap minimum match identity of 90% and a overlap minimum match length of 30 bases. The total number of contigs obtained was 859 with the largest contig 60 491 bp and the average contig size larger than 500 bp were 5 696 bp. The joint assembly were mapped to the GS-FLX Titanium assembly followed by contra-mapping (Figure 4.13) with the use of CLC Main Workbench 5.2. During these assemblies a minimum aligned read length of 20 bp was used with medium alignment stringency. These requirements were used in order not to overlook contigs that may align at the beginning or the end of the GS-FLX Titanium assemblies.

With the use of contra-mapping the GS-FLX Titanium contigs were reduced to 268 contigs of which the smallest contig was 115 bp and the largest 78 465 bp with an overall average contig size of 9 558 bp. The total genome size of *Av. paragallinarum* Modesto, when computing all the contigs thereof, is 2 561 700 bases. Based on the genome sizes of members from the *Pasteurellaceae* and working with the upper limit of 2.2 Mb this indicates that

the *Av. paragallinarum* genome is at least 16% larger than the largest completed genome for the *Pasteurellaceae*.

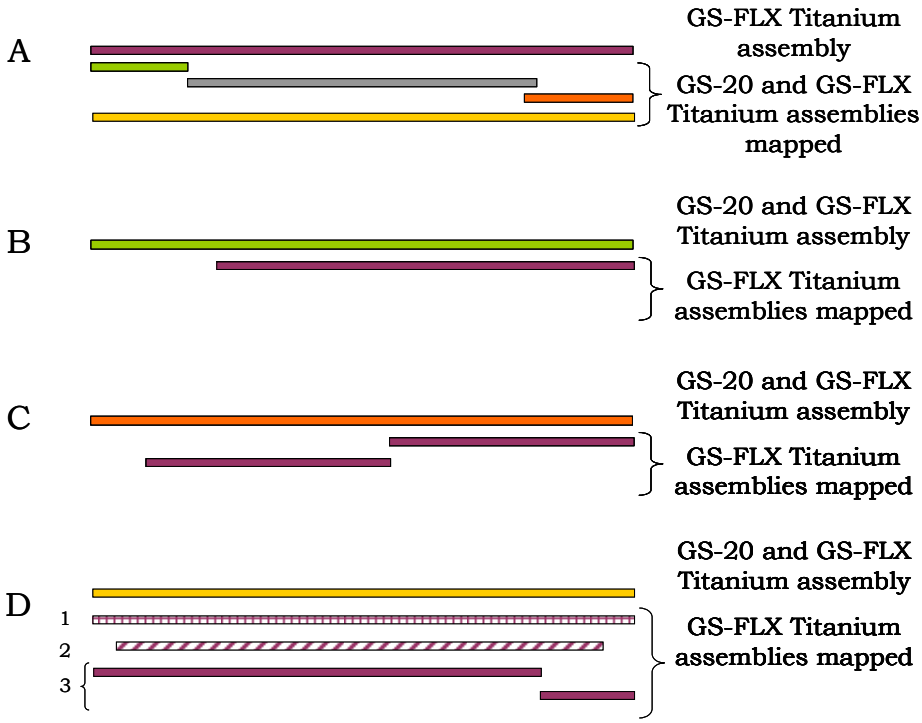


Figure 4.13. Contra-mapping. The mapping of the combined GS-20 and GS-FLX Titanium assemblies to the GS-FLX Titanium assemblies are illustrated in A. Contigs that align at the ends of the GS-FLX Titanium assemblies were subjected to contra-mapping where those contigs serve as the reference against which the GS-FLX Titanium assemblies were mapped as illustrated in B, C and D. In some cases nothing more can be done as illustrated in B and D_1. In other cases it is possible then to close a gap between two GS-FLX Titanium contigs as illustrated in C and D_3. In cases observed as illustrated by D_2, some contigs could be extended at least.

Although the minimum number of contigs obtained was 268, it does seem that the contig set represent the whole genome of *Av. paragallinarum* Modesto based on the predicted genome size. It is clear that the complete genome could not be assembled due to various repeated regions which have been observed. In some cases certain consecutive bases repeated within a contig (Figure 4.14) and in other a range of bases repeated between different contigs (Figure 4.15).

GCCTTATTTTTAAGTCAAACACCAAGGCTGTTTGGATAAATAATCTTGTAAAAAATCAATCAATAATGCCGTGCGTGGGGCAAGGGTATGGT
GGCGGTAATACACGGCATAAATGGGTTGTGTGAGTGATTGGTGTAAATCAGGCAGGATTTCGACCAATTCGCCATTTTGGCGTGCGTTGTGT
GTCATCATATCCGAAAGGCATACCACACCAAGATGATTGAGTGCCAATGCCAATAGGCTTTCGCCGCTAGAGGCAGAGATATGCGGCTGTAT
GTCAAGCCAATGCCCTGCGTTGTCTAAAATTGGCCATTGGTTGAGGTGAGCTGGTTGACTAAAACCTAATAAACGTTGTTGGCATAATTGCT
CTAGATTGTGAATGTTGCCGTGTTCAGCTAAATATTGCGGGCTAGCTAGAATACGAACGGCGGTTTTACCAAGCAAGCGGGCGTGCATGGTG
GAATCTTGTAGCGTGCCAATGCGAATGGCAACATCGGTGCGTTCCTCTAAGAGATCGATGTTTAACTTATTGCTGTTTAGCTCCAACTGGAG
CTGGGGATAGCGTGGCAAAAATTCGGCGATAGCTGGAATAAGAACGTGTCGCATCACGGGGGTGGAAGTATTCACACGCAATAATCCGCTGA
CATCACTACGCCGAGCGAGCAGTTTTTCTtCCGCACTTTGCACACTGTCTAAAATTTGCCGTGCTTGTTGCAACAACAATTCCCCTTCATCA
GTGAGGTGCATTTTTCGCGTACTGCGTTGCAATAAGGTCATTCCCAGCTCTCGTTCTAGTTTTGTCAAAGCACGGCTCACCATTGAAACAGG
CTGTTCCAAACGCATTGCCGCCCCGCTAAGCGACCCGCTTTCCACGATATGAACAAAGATATTTAGGGCTTCCAGATTCAGTTTCACGACAC
GTTCCCCAACACAAAaGAATGGGCTTATTATTGCAAAAaTAAAAaTTAAATAACAATTATTTTTGCGCAAATCGCAAAAGTGATTTGCTGTT
TCTCGGTTTTTTTCTTCCATTTTTTTGCATATAATCCCTCACATATCGCCACATCGCGTGGCTTCTATTTCAGCAAAaTAAAaGGAAAAaCA
ATGAACATTTTATTACTCAACGGTGGCAAAGCATTTGGCCATTCTCACGGTGAACTCAATGCCATCTTACACAACGAAGCCAAAACCCTGTT
AAGTGAATTGGGGCATCAGGTTCAAGAAACCCATATTGACGCGGGCTACGACATTGAAACCGAAATTGAAAAATACCTCTGGGCTGACGCAG
TCATCTACCAAATGCCAGGTTGGTGGATGGGCGAGCCTTGGATCGTGAAAAAATATATGGACGAAGTGTTCACCATCGGACACGGCAAACTG
TATCAAAGCGACGGCCGCCACCGCGTCAATCCCACCGAAGGCTACGGCACAGGTGGCTTGCTGCAAGGGCGGAAACATATGCTTTCTCTCAC
TTGGAACGCACCAATTGAAGCCTTTACCCGCGAGGGTGATTTTTtCAACGGCGTGGGCGTTGATGGCGTGTATCTGCACTTTCACAAAGCCA
ACGAATTTCTCGGCTTTACCCCATTGCCAACCTTTATTtGTAATGATGTGATTAAATCACCGAATGTGCCTGAATATTTAGCGAATTACCGT
CAGCATTTGACAGAGGTTTTTGGCAAAGCTGAATAAATGGGCATTAATCTTTTTTATCAGCCTAAATTAAAGGCTTTCGCATAGATTATTTC
TCCTTGTGGAGGGGTTTCTGTAAAGGGAAGAACACGCAATACCCCTTCTGGCAAGGGGAAGGGAGCGGTAATTCCGTAATGAGAAGCCGTTT
GATAGCCAAAGCGAGAATAATACTCAGGGTGTCCTAATACGACGATAGCAGCATATTCCTTGTTTTTAAGCTGTTGATGTGCTGCTAAAATC
AGTGCTTTTCCCACACCTTGTTTTTGCCACTGAGGTAATACCGAGAGGGGAGCTAAGGCTAACACGGGTTGTTGTCCAATGGTGGCGGGGCT
AAATAAAATATGTCCGATTATTTTCCCTTCATTTTCAGCAACTAAACTCAGTACAACCTCGTTGTTTTCATATAAACGGCGGACAAGTAAGG
CTTCTTTTCCATCACTAAATGGCATATTAGCGAAGGCTTCTTGAATCACATTTTCGATTTCTTGATTATCCTCGATACGAGATTGACGCAAA
ATAAAATGGGATTGTGTCATTTTATTTCTCCTAAATTATTGATAAAAAACGGATTGTATGGCGGAAAAGATTTCAAGGCAAGAAACAAAAGT
GCGGTGATTTTTATCATTGGAATTTTACAACGTAAAAAATTGACAAAATGATGATTTTATTTCTTAATTTATGCTCTAGCTTGTAGAATAAT
AAAAATTTTTTGATAAAAAAaGTTGACATAAAAAAaGGACACATCATAATGCCTGACAGCACAGCACAGCACAGCACAGCACAGCACAGCAC
AGCACAGCACAGCACAGCACAGCACAGCACAGCACAGCACAGCACAGCACA

Figure 4.14. Consecutive base repeats. This is an illustration of consecutive base repeats observed within a contig.

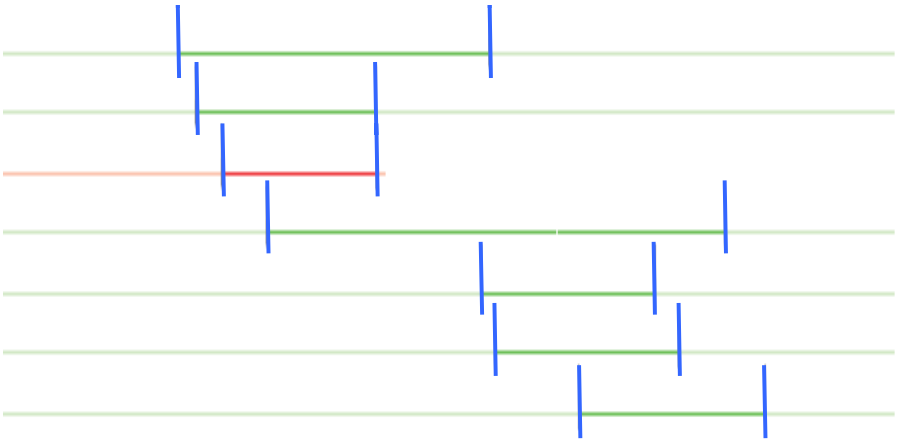


Figure 4.15. Repeats between different contigs. This diagram illustrates repeated regions that is shared between various contigs. Each horizontal line represents a different contig and the repeated region that is shared between the contigs can be found between the vertical lines.

A group from Denmark under the leadership of Professor Henrik Christensen, Department of Veterinary Disease Biology, Center for Applied Bioinformatics

performed a titanium pyrosequencing run on *Av. paragallinarum* A-1 (0083). This data set revealed a total number of 485 contigs with the largest contig 51 124 bp and the smallest contig 106 bp. Once more the difficulty experienced with an attempt to close the genome of *Av. paragallinarum* A-1 could be ascribed to repeats observed. Moreover this group sequenced the genome of another *Pasteurellaceae* family member (type strain of *P. multocida* subsp. *septica*) for which they received only 30 contigs. This contig set proves that the difficulties experienced here are species related rather than strain specific.

Both the GS-20 and GS-FLX Titanium reads were mapped against *Haemophilus influenzae* 86-028NP; *Pasteurella multocida* subsp. *multocida* str. Pm70 and *Aggregatibacter aphrophilus* NJ8700 using the GS-Reference Mapper 2.0.00 from 454 Life Sciences Corporation in order to reduce the number of contigs. The GS-20 and GS-FLX Titanium reads were mapped to these genomes together with the GS-FLX Titanium reads alone (Table 4.2).

Table 4.2. Reference genome mapping.

	***Haemophilus influenzae* 86-028NP**		***Pasteurella multocida* subsp. *multocida* str. Pm70**		***Aggregatibacter aphrophilus* NJ8700**	
Results	**GS-20 GS-FLX T***	**GS-FLX Titanium**	**GS-20 GS-FLX T***	**GS-FLX Titanium**	**GS-20 GS-FLX T***	**GS-FLX Titanium**
Fully mapped reads	0.43%	0.12%	0.50%	0.12%	0.50%	0.12%
Partially mapped reads	6.47%	12.03%	6.21%	11.73%	6.21%	11.73%
Largest contig size (bp)	2 301	2 301	2 347	2 337	2 343	2 337
Total number of contigs (n)	361	336	370	349	370	349

*GS-20 and GS-FLX Titanium reads combined.

Unfortunately using a reference genome did not aid in lowering the total amount of contigs for *Av. paragallinarum* (Table 4.2). There is also not a

significant difference when using only GS-FLX Titanium reads or in combination with the GS-20 reads. In all of the cases the percentage of fully mapped reads was less than 1%. The largest contig size was between 2 347 bp and 2 301 bp meaning that (n-1) contigs were all smaller and as a result it could not assist the assembly of the genome of *Av. paragallinarum*. A reference assembly using contigs from *Av. paragallinarum* Modesto and *Av. paragallinarum* 0083 was attempted in order to investigate the possibility that either one can aid a complete assembly of the other but this method delivered no results.

The initial 300 contigs from the GS-FLX Titanium assembly were sent to the J. Craig Venter Institute (JCVI) for annotation. This institute uses the Glimmer system to identify genes in bacterial, archaeal or viral genomes. Following the identification of each putative open reading frame (ORF), these candidate genes, are translated and protein sequences is searched against an internal non-identical amino acid database consisting of all proteins available from GenBank (NCBI); Protein Information Resource (PIR); SWISS-PROT (Expasy) and Omniome (JCVI). Initially a BLAST search is performed where after a modified Smith-Waterman alignment is executed against the BLAST outcomes. The Smith-Waterman alignment identifies potential frameshifts or point mutations by extending the gene 300 nucleotides up- and downstream. Frameshifts, point mutations or probable sequence errors are then indicated as pairwise and multiple alignments. Several additional sequence based searches are performed including SignalP.

Data received from the JCVI indicated a total of 2 842 ORF. ORF's obtained could either be assigned a function, were conserved hypothetical proteins, proteins of unknown function or hypothetical proteins and finally a group of proteins that could not be assigned a function (Figure 4.16).

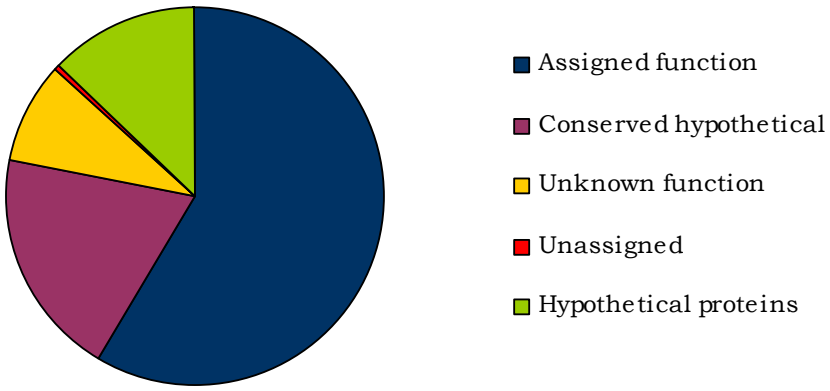


Figure 4.16. Preliminary ORF analysis. This diagram illustrates the quantity of ORF that could either be assigned a function or not. Hypothetical and conserved hypothetical proteins are not assigned to a function.

ORF's that could be assigned to a protein and a corresponding function were then further divided based on their function into: amino acid biosynthesis; purines, pyrimidines, nucleosides and nucleotides; fatty acid and phospholipids metabolism; biosynthesis of co-factors, prosthetic groups and carriers; central intermediary metabolism; energy metabolism; transport and binding proteins; DNA metabolism; transcription; protein synthesis; protein fate (modification, folding, stabilization etc.); regulatory functions; signal transduction; cell envelope; cellular processes; mobile and extrachromosomal element functions (Figure 4.17).

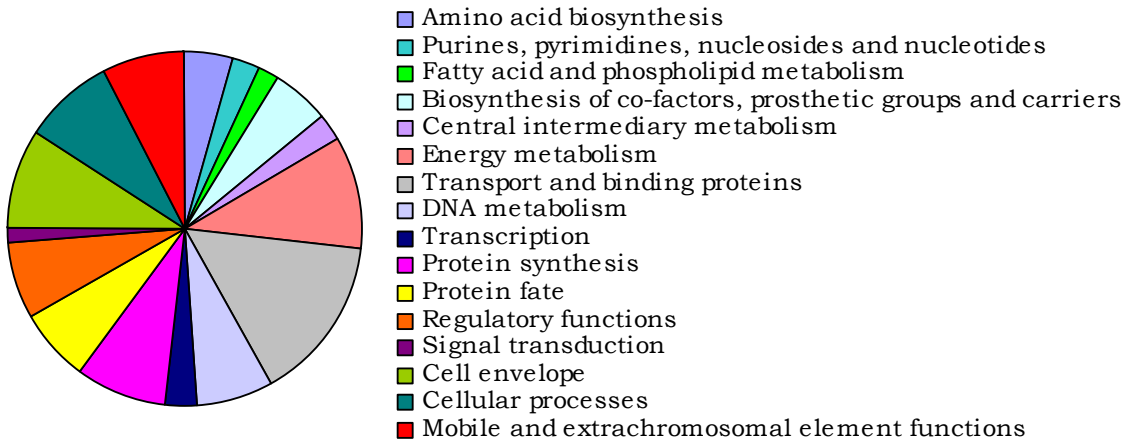


Figure 4.17. Protein functionality assignment. The proteins that could be assigned to a function could then be divided into a category based on that function as illustrated here.

Looking at the functionality assignment it is interesting to note that *Av. paragallinarum* currently has more proteins involved in the cell envelope as in cellular processes. This could contribute to the variety of serovars and serotypes observed within *Av. paragallinarum* and towards difficulties observed regarding immunogenicity. Mobile and extrachromosomal element functions also seem to play a significant part within *Av. paragallinarum*.

4.5. Concluding remarks

Through the use of pyrosequencing the genome of *Av. paragallinarum* was sequenced. Two pyrosequencing runs was performed one with a GS-20 instrument and the other with a GS-FLX Titanium instrument. Working with a genome as complex and repetitive as the genome of *Av. paragallinarum* short reads (GS-20) increase the difficulty during assembly and therefore it can be concluded that the shorter the read and the more complex the genome the higher the coverage required. Read length and required coverage are indirectly proportional to each other. Even though the GS-20 coverage was not adequate

the data was still useful through contra-mapping in closing gaps between the GS-FLX Titanium contigs. With the GS-FLX Titanium assemblies the increased read length had a dramatic decrease in total amount of contigs but read length still leaves much to desire when repetitive moieties are involved. When contra-mapping was performed the total number of contigs decreased from 300 to 268.

With current techniques available it is not possible to close the genome of *Av. paragallinarum* due to various types of repeats observed within the genome of *Av. paragallinarum*. With the analysis of whole genome data obtained from Professor Henrik Christensen (Denmark) for *Av. paragallinarum* 0083 it seems that repeats observed within the genome are species related rather than strain specific.

Mapping of *Av. paragallinarum* reads to reference genomes indicated that this option can only be attempted when the organism of interest is very closely related to the reference genome used.

The contigs obtained from the GS-FLX Titanium assembly was send to the JCVI for annotation. Well over half of the ORF's found could be assigned to a protein, though there is still a significant amount of ORF's that are hypothetical or conserved hypothetical proteins and could not be assigned to a protein. The ORF assigned to proteins could be further divided based on the associated function of the protein. Most of the proteins assigned are involved in the transport and binding of proteins. More proteins were assigned to the cell envelope than cellular metabolism which could explain the variety of serovars and serotypes observed within *Av. paragallinarum*. A great deal of proteins was assigned to mobile and extrachromosomal element functions which may explain the difficulty experienced during the assembly of the genome.

4.6. Literature cited

Blackall, P. J., and R. Yamamoto. 1990. Infectious coryza. In: A laboratory manual for the isolation and identification of avian pathogens, 3rd ed. H. G. Parchase, L. H. Arp, C. H. Domermuth and J. E. Pearson, eds. *American association of avian pathologists, Ames, Iowa.* Pp. 27-31.

Chen, X., J. K. Miflin, P. Zhang and P. J. Blackall. 1996. Development and amplification of DNA probes and PCR tests for *Haemophilus paragallinarum*. *Avian Dis.* **40**:398-407.

Mardis, E. R. 2008. Next-Generation Sequencing Methods. *Annu. Rev. Genom. Human Genet.* **9**:387-402.

Mendoza-Espinoza, A., Y. Koga and A. I. Zavaleta. 2007. Amplified 16S Ribosomal DNA Restriction Analysis for Identification of *Avibacterium paragallinarum*. *Avain Dis.* **52**:54-58.

Roche Applied Sciences. https://www.roche-applied-science.com

Rothberg, J. M., J. H. Leamon. 2008. The development and impact of 454 sequencing. *Nat. Biotechnol.* **26**:1117-1124.

Shendure, J. and H. Ji. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**:1135-1145.

Chapter 5

Genetic diversity through integration motifs within *Avibacterium paragallinarum*

5.1. Abstract

The whole genome sequencing annotation implicated various genes involved in mobile and extrachromosomal elements. An integrated p250 from *Av. paragallinarum* was identified. Nine site-specific integrases were identified and they share little homology with one another. Ten transposases were identified of which three of the transposases are insertions sequence elements (IS) and the remainder phage-transposases. These transposases share little homology with each other including the two transposases associated with the same IS-family. Numerous prophage functions were assigned to open reading frames identified during the annotation process which led to a putative Mu-like prophage identified for *Av. paragallinarum*, ΦAvpmuC-2M, as well as a HP2-like prophage ΦAvpC-2M-HP2.

5.2. Introduction

In 2003 Terry *et al.* reported the first sequence of a plasmid isolated from *Avibacterium paragallinarum* (Figure 5.1). This plasmid is 6 286 bp in size and it contains seven open reading frames, ORF's. These ORF's include a haemocin operon, *hmc*, a putative replication (*repB*) protein and a putative integrase (*int*) (Chapter 3).

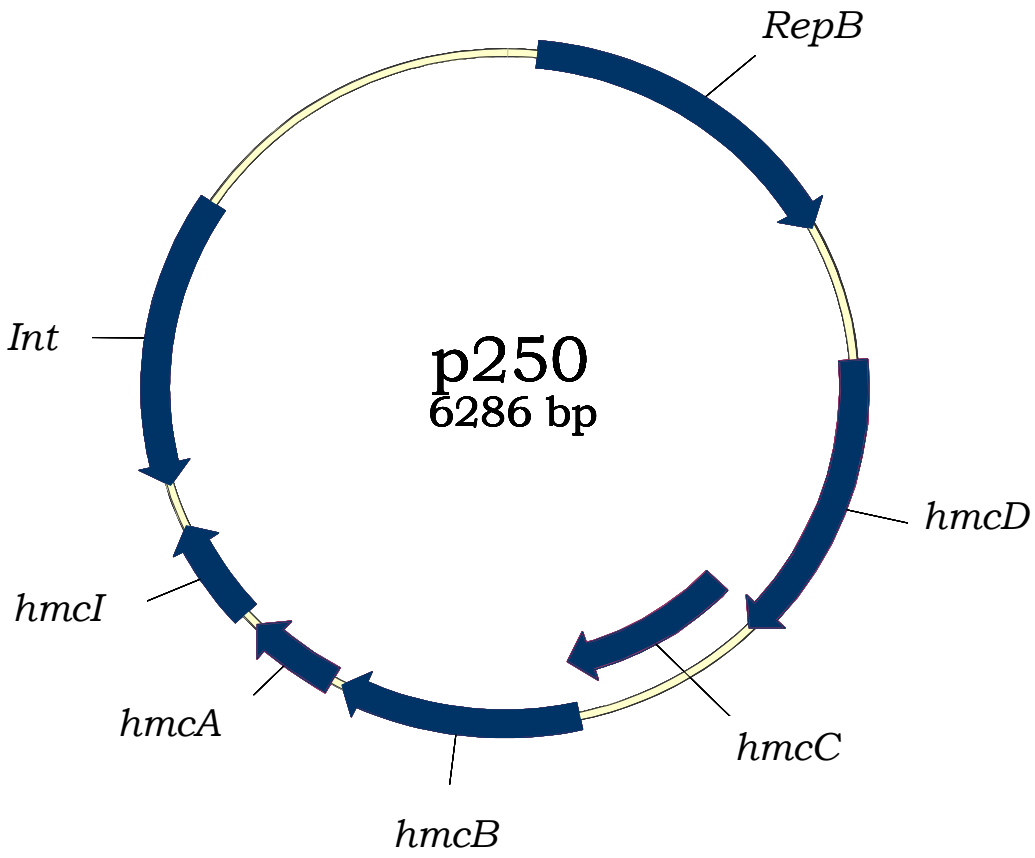


Figure 5.1. *Avibacterium paragallinarum* plasmid p250 illustrating the seven ORF's found on this plasmid namely *repB,* the putative replication protein; *hmcD*, *hmcC*, *hmcB*, *hmcA* and *hmcI* comprising the haemocin operon and *int,* the putative integrase (Terry *et al.*, 2003).

The haemocin operon is similar to the haemocin operon of *Haemophilus influenzae* and these genes are responsible for the production, processing,

export and immunity of haemocin. The mechanism of action of this bacteriocin is unknown, it is however suggested that this property may enhance the colonization of *Av. paragallinarum* in the sinuses of chickens. This property was also observed in strains that do not contain a plasmid and it was proposed to be chromosomally located (Terry *et al.*, 2003). Haemocin-like activity which could not be attributed to plasmid p250 has also been reported for *Av. paragallinarum* by Hsu *et al.* (2007). These authors also reported on another putative pathogenesis plasmid, plasmid pA14, from this bacterium. Although the full plasmid sequence was not obtained it was reported that this plasmid contains a truncated MglA protein and an RNase II protein which could be associated with virulence (Hsu *et al.*, 2007).

Temperate bacteriophages play an intricate role in the generation of microbial diversity and in the evolution of bacterial genomes through the mediation of rearrangement within the bacterial chromosome. Well documented horizontal gene transfer has been implicated to have a vast impact on bacterial pathogenesis (Balding *et al.*, 2005). Within the family *Pasteurellaceae* various bacteriophages have been reported in *Haemophilus influenzae*, *Actinobacillus actinomycetemcomitans*, *Pasteurella multocida* and *Mannheimia haemolytica*, and bacteriophage-like sequences have been reported in *Histophilus somni* (Highlander *et al.*, 2006). Phage integration requires an attachment site on the phage namely *attP* and a homologues attachment site on the bacterial genome namely *attB* (Figure 5.2.). The phage integration is then achieved through the phage integrases that perform recombination between the attachment site on the phage and the attachment site on the bacterial genome (Groth *et al.*, 2000).

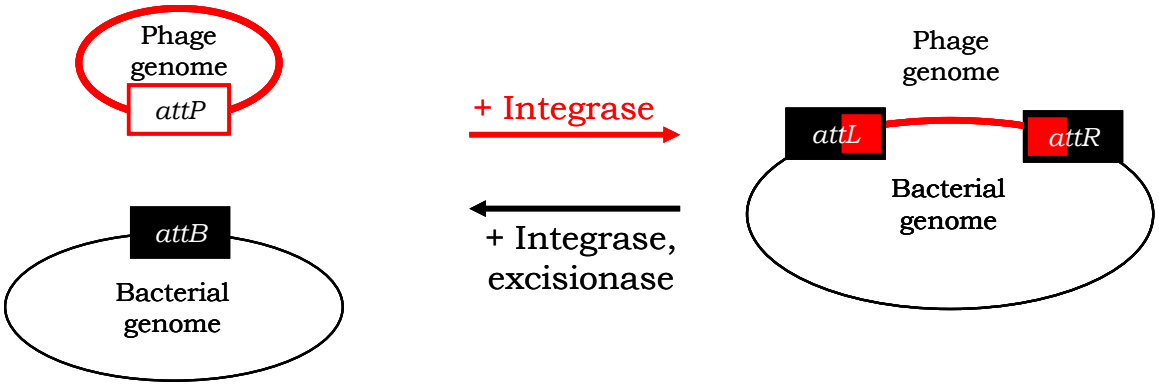


Figure 5.2. Bacteriophage integration and excision. The bold (red) circle represents the bacteriophage genome along with the phage attachment site (*attP*) and the narrow circle (black) represents the bacterial genome along with the bacterial attachment site (*attB*). When an integrase is present, in some cases, a host factor is required as well, the integration reaction can occur resulting in the phage genome being inserted within the bacterial genome flanked by two hybrid attachment sites namely *attL* and *attR*. In the presence of an excisionase and an integrase the phage genome is excised from the bacterial chromosome resulting in a phage genome with an *attP* site and a bacterial genome with an *attB* site. In some cases the excision reaction requires host factors and/or cofactors. (Groth and Calos, 2003)

Transposons and insertions sequence elements (IS) add to genetic diversity in their own unique approach. Transposons may be referred to as jumping gene systems and are present in a variety of forms recognized through their structure, genetic relatedness and mechanism of transposition. Transposons have the ability to move a variety of genes both inter- and intra-molecularly. These elements can relocate from one site in a DNA molecule to another or from one DNA molecule to another DNA molecule (Bennet, 2008). IS differ from transposons mainly due to the fact that they are the smallest autonomously mobile elements, they are phenotypically cryptic in nature and they may or may not encode functions that alter the phenotypic appearance of the cell. On the other hand, transposons encode at least one function that alter the phenotypic appearance of the cell (Bennet, 2008; Rousseau *et al.*, 2004; Mahillon and Chandler, 1998).

This chapter is dedicated to the genetic diversity observed within *Av. paragallinarum* through integrative plasmids, bacteriophages, transposons and IS elements.

5.3. Materials and Methods

5.3.1. Bacterial strains

Avibacterium paragallinarum strains C-2 (Modesto) and 1750 were obtained from *Onderstepoort Biological Products*, Onderstepoort, South Africa. Strain Modesto is NAD^+-dependent and was cultivated in TM/SN medium containing 1% (m/v) biosate peptone, 1% (m/v) NaCl, 0.1% (m/v) starch and 0.05% (m/v) glucose with 5% (v/v) oleic acid-albumin complex, 1% (v/v) chicken serum, 1% (v/v) NAD^+ and 0.0005% (m/v) thiamine solution as supplements grown under oxygen limiting conditions. The NAD^+-independent strain 1750 was cultivated in the same manner as strain Modesto with the exclusion of NAD^+ and chicken serum supplements (Blackall and Yamamoto, 1990). Agar (1.7% m/v) was used in the case of solid media with cultivation occurring in a candle jar at 37˚C for 16 h.

5.3.2. Enzymes, chemicals, kits and other consumables

Restriction enzymes were obtained from Fermentas and polymerase chain reaction (PCR) reagents were obtained from New England Biolabs® Inc (NEB).

Chemicals used during this study were of analytical or PCR grade and were used without further purification. All relevant chemicals were obtained from Sigma-Aldrich or Merck, NAD^+ (PCR grade), bovine serum albumin fraction V and proteinase K were obtained from Roche Molecular Biochemicals.

Oligonucleotides were either obtained from Integrated DNA Technologies (IDT) or Bioneer.

5.3.3. Techniques applied

5.3.3.1. DNA extraction

Genomic DNA extractions were performed with the use of a QIAamp® DNA Mini kit from Qiagen. The manufacturer's recommendations were meticulously followed except for the elution procedure where genomic DNA was eluted in 2 X 80 µl 10mM Tris-HCl, pH 8.

Plasmid DNA extraction was done with the use of the QIAamp® Maxi kit from Qiagen where the manufacturers' recommendations were followed.

5.3.3.2. Identification of *Av. paragallinarum*

Av. paragallinarum was identified through the use of an identification PCR developed specifically for *Av. paragallinarum* (Chen *et al.,* 1996). The reaction mixture contained 2 µM of the primer set AP-1F and AP-1R (Table 4.2) together with 0.5 µg of genomic DNA, 5 µl of the 10 x ThermoPol buffer, 200 µM dNTP's, 2 units *Taq* polymerase mix, filled to a final reaction volume of 50 µl using sterile redistilled water. When identification PCR's were performed on cultures taken from plates, 2-3 colonies were picked and resuspended within the PCR mixture and when PCR's were performed on liquid culture, 5 µl of the liquid culture was resuspended in the PCR mixture. The reaction cycle included an initial heat denaturation of 10 min at 94˚C; this was done in order to break the cells enabling genomic DNA release after which 2 units *Taq* polymerase was added. The heat denaturation was not required where

genomic DNA was directly used as template. This was followed by 30 cycles of denaturation (94˚C for 25 sec); annealing (55˚C for 50 sec); elongation (72˚C for 45 sec) and a final elongation cycle of 7 min at 72˚C with an expected amplification result of approximately 500 bp.

Additionally, the 16S ribosomal DNA, rDNA, region was amplified in order to identify *Av. paragallinarum* (Mendoza-Espinoza *et al.*, 2008). The reaction mixture contained 2 µM of the primer set 27F and 1492R (Table 4.2) together with 5 µl of the 10 x ThermoPol buffer, 200 µM dNTP's, 2 units *Taq* polymerase, filled to a final reaction volume of 50 µl using sterile redistilled water. The reaction cycle included an initial denaturation of 2 min at 95˚C followed by 35 cycles of denaturation (95˚C for 30 sec); annealing (49˚C for 45 sec); elongation (72˚C for 1 min 30 sec) and a final elongation cycle of 10 min at 72˚C with an expected amplification result of approximately 1500 bp. The PCR product obtained was excised from the agarose gel, purified and directly sequenced without any sub-cloning. This allowed discrimination against possible genomic DNA contamination.

5.3.3.3. Polymerase chain reactions

All other PCR reactions were performed with the use of NEB *Taq* DNA polymerase according to the manufacturer's recommendations. The reaction mixture contained 5 µl of the 10 x ThermoPol buffer, 2 µM upstream and downstream primers, 200 µM dNTP's, 0.5 µg template DNA and 2 units *Taq* polymerase, filled to a final reaction volume of 50 µl using sterile redistilled water. Thermal cycling was performed using an Eppendorf Mastercycler Personal with the following cycle profile: initial denaturation of 2 min at 94˚C, 35 cycles of denaturation (94˚C for 30 sec); annealing (temperature is primer dependent for 30 sec); elongation (at 72˚C time is dependent on the size of the anticipated product) and a final elongation cycle of 7 min at 72˚C.

Table 5.1: Oligonucleotide primers used in this study.

Primer	5'→3' Orientation	Application
AP-1F	TgA ggg TAg TCT TgC ACg CgA AT	ID PCR
AP-1R	CAA ggT ATC gAT CgT CTC TCT ACT	ID PCR
24F	AgA gTT TgA TCM Tgg CTC Ag*	16S rDNA
1492R	TAC ggY TAC CTT gTT ACg ACT T*	16S rDNA
Int-1F	gCC TTg CTA TCg CCC AAg ATA AAC AgC	PCR
Int-2R	gTg AgC Tgg gAC ATA ATC gAg AAg	PCR
Rep3F	Cgg CAA TTA CgC AAg AgA TTg C	PCR

* Degenerate nucleotides: M = A/C; Y = C/T.

In order to avoid confusion between guanine and cytosine within the oligonucleotides guanine is indicated as lowercase.

5.3.3.4. General techniques

PCR and DNA product were analyzed with the use of agarose gel electrophoresis. Agarose gels consisted of 1% (w/v) agarose in TAE buffer (0.1 M Tris; 0.05 M EDTA, pH 8; 0.1 mM glacial acetic acid) stained with Goldview™ (SBS Genetech) according to the manufacturer's recommendations. DNA was separated within the agarose gels at 5.6 V/cm for 45 min and documented using BioRad Gel Doc XR system. DNA to be isolated from agarose gels for further studies was visualized using a Dark Reader®.

PCR products were either purified from agarose gels or from solution with the use of the Bioflux BioSpin Gel Extraction Kit from Bioer Technology Co., Ltd.

PCR products were sequenced directly with the use of the ABI Prism® Big Dye™ Terminator Cycle Sequencing Ready Reaction Kit v. 3.1 (Applied Biosystems) according to the manufacturer's recommendations. Nucleotide composition was analyzed with an Applied Biosystems 3130*xl* Genetic Analyzer. Sequence

analysis was performed using either Geneious Pro 4.6.1 from Biomatters or CLC Combined Workbench 3.5.1 from CLC bio A/S.

Genetic diagrams were constructed with the use of pDRAW32 from ACACLONE software and homology trees were constructed with the use of DNAMAN 5.2.9 from Lynnon Biosoft.

5.4. Results and Discussion

Terry *et al.* (2003) described plasmid p250 from *Av. paragallinarum* and reported that various strains showed haemocin activity even though no plasmid DNA was obtained. Those strains therefore seemed to contain the haemocin operon located in their genome. Chapter 3 reports on the screening of plasmid extracted from four NAD^+-independent *Av. paragallinarum* strains. The strains that showed the presence of plasmid was screened for the p250 plasmid of *Av. paragallinarum* and the screening process indicated that *Av. paragallinarum* NAD^+-dependent strain Modesto possibly carry the haemocin operon within the genome.

Genome sequence data of *Av. paragallinarum* strain Modesto was analysed to determine if the p250 plasmid containing the haemocin operon is in fact integrated as proposed by various authors (Terry *et al.*, 2003, Hsu *et al.*, 2007). The integrated p250 plasmid was identified in only one of the contigs obtained from the genome sequence data (Figure 5.3).

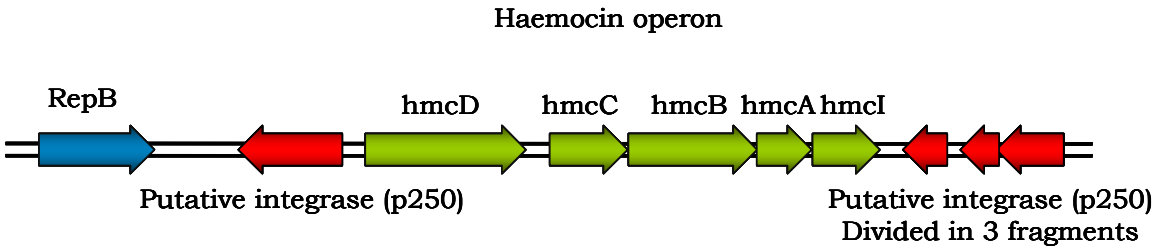


Figure 5.3. Schematic representation of the integration of p250 into the genome of *Av. paragallinarum* strain Modesto.

Sequence data revealed that there are two integrase genes present on opposite sides of the haemocin operon and both of these are identical to the integrase gene found on the p250 plasmid. One of the integrase genes is divided into 3 fragments however; the reason for this remains unclear. Furthermore, it seems that the inverted repeats present between the *repB* and *int* genes from p250 are not present in the integrated copy.

To confirm the integration of the p250 plasmid a PCR was performed on genomic DNA from *Av. paragallinarum* Modesto and plasmid p250 from *Av. paragallinarum* strain 1750 as template together with primer pair RepB-3F and Int-2R. If p250 was integrated within the genome a fragment size of 671 bp was expected and when p250 was not integrated a fragment size of 3 617 bp was expected. Genomic DNA from *Av. paragallinarum* Modesto and plasmid p250 from *Av. paragallinarum* strain 1750 was used as template during the p250 integration PCR (Figure 5.4).

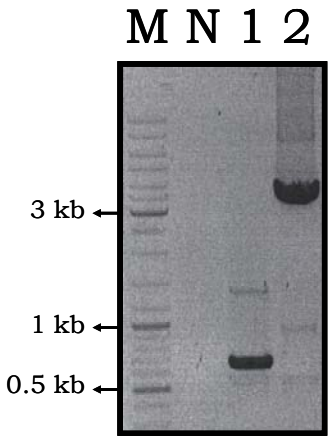


Figure 5.4. p250 Integration PCR. Lane M indicates the molecular marker (GeneRuler™, Fermentas); lane N shows the negative control; lane 1 is the integration PCR performed on *Av. paragallinarum* strain Modesto and lane 2 the integration PCR performed on *Av. paragallinarum* strain 1750.

Figure 5.4 clearly indicates the integration of p250 within the genome of *Av. paragallinarum* Modesto, thereby reconfirming the genome sequence data. Furthermore, the contig which harbours the p250 integration from *Av. paragallinarum* Modesto was compared to the *Av. paragallinarum* 0083 raw contig set and it was found that the p250 integration in both Modesto and 0083 are identical.

A partial plasmid pA14 was described by Hsu *et al.* in 2007 and this was termed a pathogenesis plasmid. The two genes identified on this partial plasmid (a putative truncated MglA protein and an RNase II protein) were used in a nucleotide-to-nucleotide comparative study against the whole genome sequencing data of *Av. paragallinarum* Modesto and *Av. paragallinarum* 0083. Both these genes were found to be chromosomally located and were identified on two separate contigs. If these two contigs are artificially linked to one another in every possible orientation the smallest amount of bases that separated the two genes are 11 724 bases. Furthermore, analysis of the surrounding regions does not indicate the presence of either an integrase or

replication like gene. This indicates that these genes are unlikely to form a plasmid in *Av. paragallinarum* Modesto and *Av. paragallinarum* 0083.

Annotated whole genome sequencing data of *Av. paragallinarum* Modesto obtained from the J. Craig Venter Institute (JCVI) indicated the presence of nine phage integrase genes. These integrase genes could either be repeats of each other or not related to one another. This relationship was determined through sequence homology analysis (Figure 5.5).

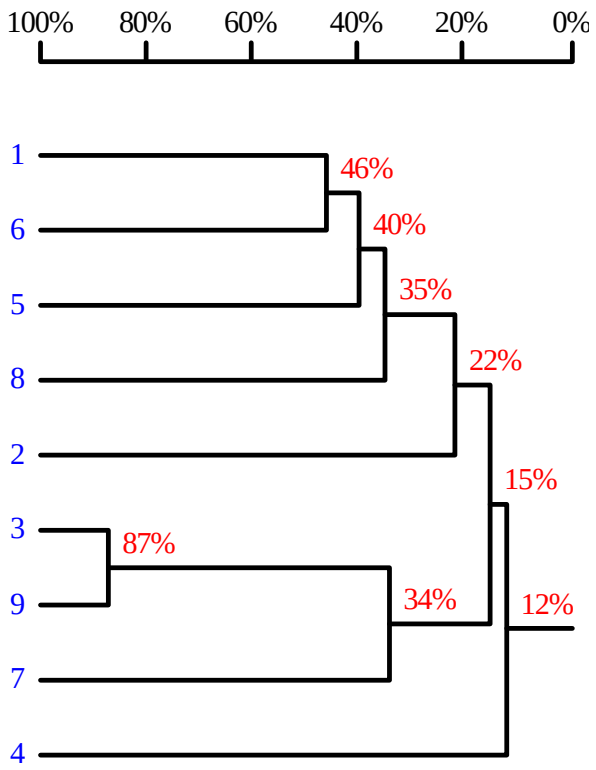


Figure 5.5. A homology tree of integrase genes identified within *Av. paragallinarum* Modesto. This diagram indicates that all the integrase genes identified shows a high level of variability.

Figure 5.5 illustrates that all nine integrase genes identified within the genome of *Av. paragallinarum* Modesto are unique. Only two of the genes (number three and number nine) show a high degree of similarity. However, the level of similarity indicates that these two are not repeats of each other. Furthermore,

all the other integrase genes show a low percentage of similarity indicating that all of them are unique.

Transposons and IS are both integrative elements that are directed through transposases and the main difference between these two elements are that transposons carry at least one trait that alters the phenotypic appearance of the cell. From the annotation of the whole genome sequence of *Av. paragallinarum* Modesto, 10 transposases were identified. As in the case of the identified integrase genes, the transposases can either be repeats from one another or they could be unique and these elements can either form part of transposons or IS elements. A homology tree was constructed to investigate the homology of these transposases to one another (Figure 5.6).

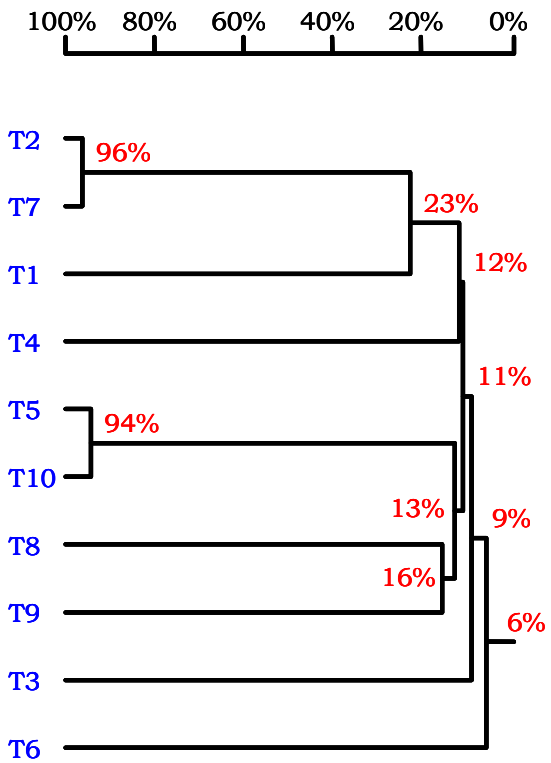


Figure 5.6. A homology tree of the transposase genes identified within *Av. paragallinarum* Modesto. This diagram indicates that a high degree of homology exists only between transposase T2 and T7 and between T5 and T10 respectively.

In Figure 5.6 it can be observed that most of the transposase genes identified within *Av. paragallinarum* Modesto are unique with a very low similarity observed between them. Only transposase T2 and T7 and T5 and T10 show a high similarity between each other respectively. Three of the transposases identified belong to two IS families namely: T1 and T9 to the IS150 family and T6 to IS605 family. The remaining seven transposases were identified as Mu-like transposases.

The presence of prophages has been reported within members of the *Pasteurellaceae* (Resch *et al.*, 2004; Highlander *et al.*, 2006). The annotation results from JCVI implicated 7% of the assigned protein functions (141 ORF's) to that of prophage functions. The various prophage genes were present on 60 contigs. Genes associated with prophage functions were mapped to the various contigs with the use of CLC Combined Workbench 3.5.1 and pDRAW32. The nine phage integrase genes (Figure 5.6) were identified within seven of the 60 contigs. Analysis of the adjacent regions to these integrase genes indicated the absence of any prophage related genes but rather genes encoding membrane proteins; metabolic proteins; ribosomal proteins; cell membrane function proteins or membrane export proteins. The remaining 132 ORF's form part of four types of prophages.

A putative Mu-like prophage was identified and the genes forming part of this prophage were mapped and the prophage was named, ΦAvpmuC-2M. A putative genome map was drawn based on the genome map of Mu-like prophages reported by Morgan *et al.* (2002) (Figure 5.7).

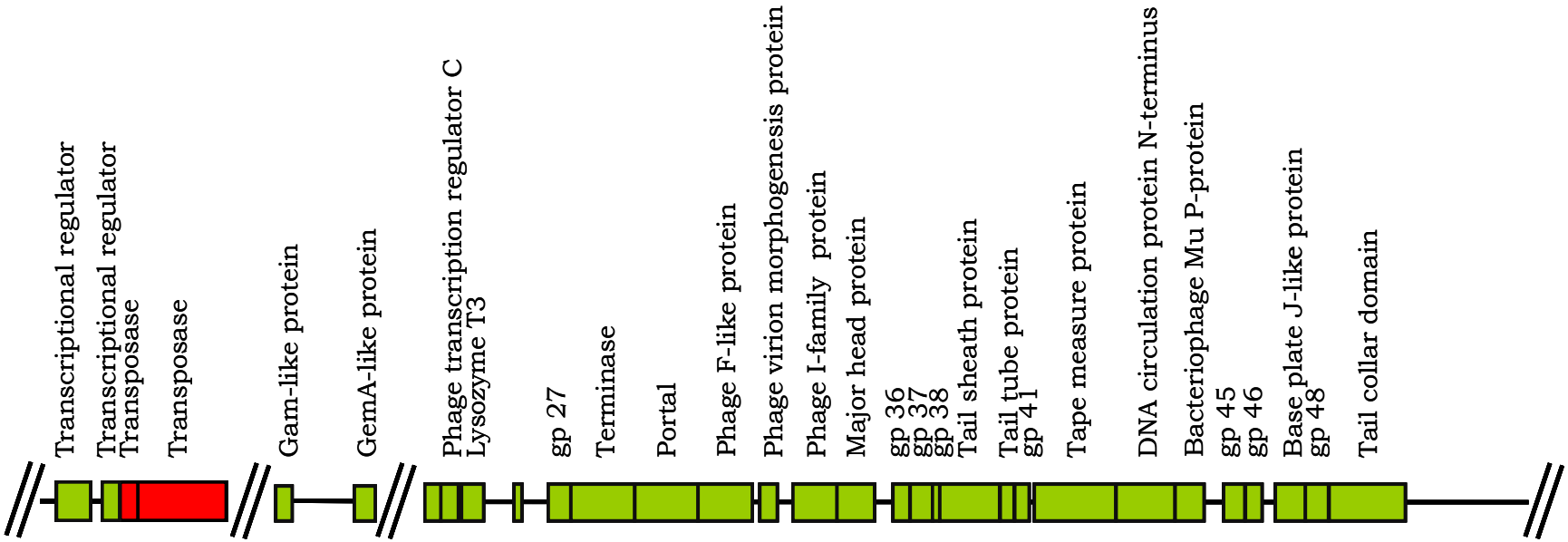


Figure 5.7. ΦAvpmuC-2M putative genome map. Beginnings and ends of contigs are indicated by //.

The presence of one or more Mu-like genes was also identified on eight other contigs. These genes were present in either a different arrangement, not all the phage genes were present or specific phage ORF's could not be assigned a specific function as reported by Morgan *et al.* (2002). Although these genes could be identified as Mu-like genes and assigned to similar function as that of ΦAvpmuC-2M, low similarity was observed when multiple alignments were performed. The Mu-like transposase within the ΦAvpmuC-2M genome map was chosen because it fitted the profile according to Morgan *et al.* 2002 but the possibility still exists that it may be any one of the seven phage transposases identified.

Forty-one lamboid genes were mapped on 11 different contigs. At this time it can not be concluded if this lamboid phage is complete. The various contigs implicated along with their associated phage functions can be observed in Figure 5.8. Lamboid genes identified play a role in the regulation, replication, structural capabilities and lysis with regard to the prophage. The exact arrangement of the genes can unfortunately not be determined because the genome of *Av. paragallinarum* Modesto could not be completely assembled due to repeated regions and current techniques not allowing the clarification thereof.

A complete lysogenic temperate prophage similar to the *Haemophilus influenzae* HP2 prophage (Williams *et al.*, 2002) was identified within the genome of *Av. paragallinarum* Modesto and was named ΦAvpC-2M-HP2 (Figure 5.9:C). This complete HP2 prophage was identified on a single contig. A nucleotide-to-nucleotide comparative study was performed where the *H. influenzae* HP2 phage were compared to the HP2 related prophage of *Av. paragallinarum* Modesto. Results indicated that ΦAvpC-2M-HP2 harbours little homology to the *H. influenzae* HP2 prophage.

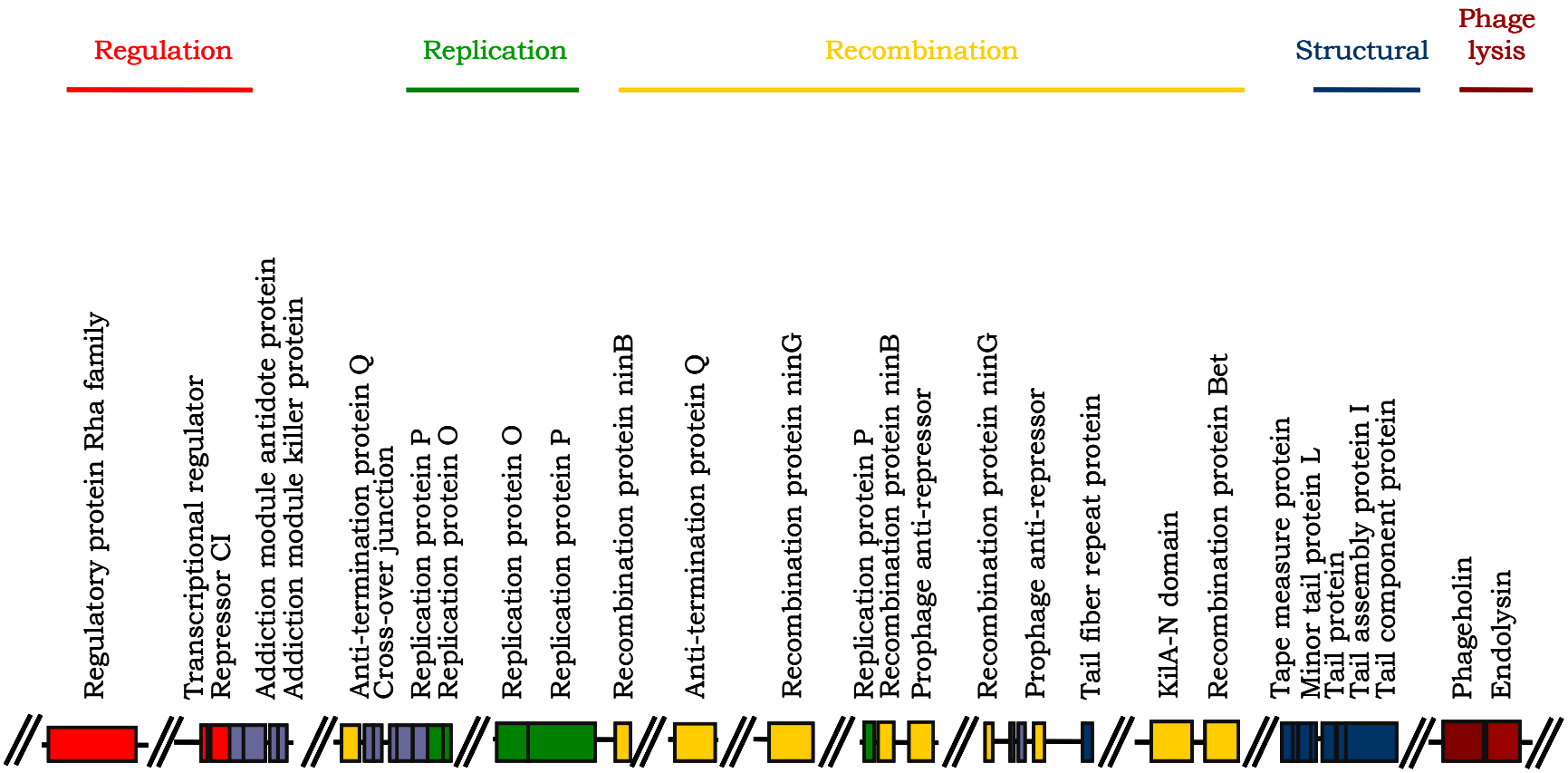


Figure 5.8. Lamboid gene mapping. Various lambda genes have been mapped along with their associated functions. Beginnings and ends of contigs are indicated by //.

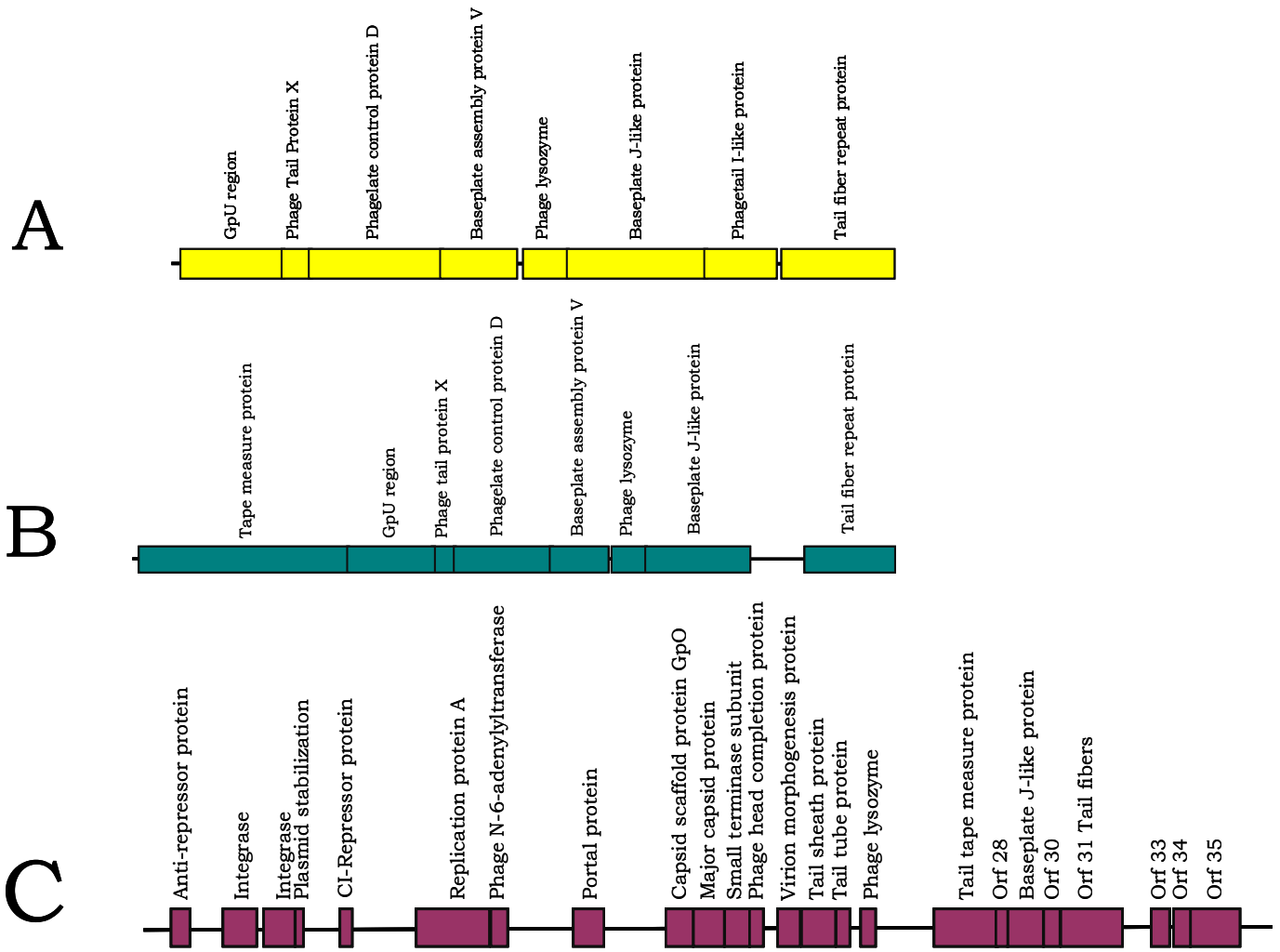


Figure 5.9. HP2-like prophage mapping. The various contigs implicating HP2-like ORF's are illustrated in the diagram a complete HP2-like prophage, ΦAvpC-2M-HP2, of *Av. paragallinarum* Modesto is illustrated in diagram C.

During the mapping of ΦAvpC-2M-HP2 two additional contigs were identified that harboured HP2-like genes from the GpU region to the tail fibre repeat regions (Figure 5.9: A & B). When these regions were compared with one another (Figure 5.10) a 80% similarity was observed between the HP2-like genes from contig A and B. However, a very low similarity was observed between the genes on contig A and B *versus* contig C.

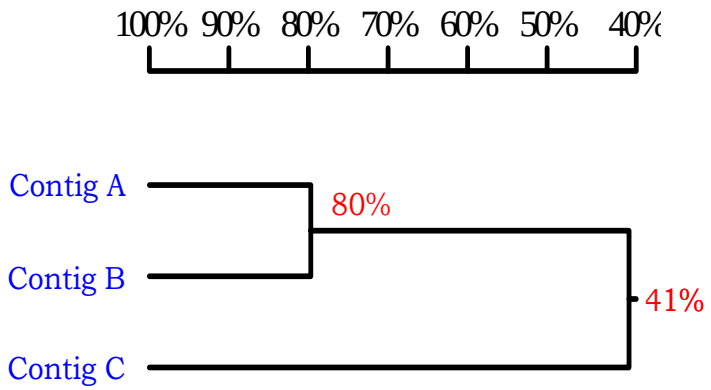


Figure 5.10. Homology enquiry of HP2 implicated contigs. Contigs harbouring the same HP2-like genes show little homology towards one another.

Figure 5.10 clearly illustrates that although the prophage genes identified on the three contigs shares the same genomic map of HP2-like genes in the same order they are not identical and share little homology.

One of the nine site-specific integrases belongs to the prophage P4 family. Annotation of the *Av. paragallinarum* Modesto indicates this to be the only gene associated with the P4 bacteriophage family. Various other phage genes have also been identified but could not be assigned to any specific phage family. These phage genes might have previously formed part of a complete prophage residing within *Av. paragallinarum* Modesto but have been lost over time. The *Av. paragallinarum* 0083 contig set were compared to the various phage moieties found within *Av. paragallinarum* Modesto a number of these

moieties are shared between these two strains while other seem to be unique to strain Modesto.

5.5. Concluding remarks

Av. paragallinarum undoubtedly forms a unique diverse species when looking at the numerous mobile and extrachromosal elements present within its genome. These elements consist of site-specific integrases, transposons and prophage elements. These elements, as well as numerous repeat regions, present in the genome of *Av. paragallinarum* may serve as hot spots for genetic recombination.

It was clearly established that the plasmid p250 of *Av. paragallinarum* has been integrated within the genome. The exact mode of integration is however unclear. The integration can be ascribed to a site-specific integrase, however attachment sites for site-specific integration for p250 could not be identified. Sequence data indicate that there are two integrase genes present on opposite sides of the haemocin operon. Following integration one of the integrase genes is divided into three regions, these regions are however not distinct cut-offs from one another, there is a degree of overlapping between these three segments. The second integrase gene is present between the *repB* gene and the haemocin operon from p250. The reason for the presence of a duplication of the integrase gene or the division of one of these genes in three sections is not clear.

In total nine site-specific integrase genes were implicated within the genome of *Av. paragallinarum* Modesto. A putative attachment site was identified for only one of the integrase genes. These genes share very little homology and are all unique.

Alongside the nine integrases, ten transposases were implicated within the genome of *Av. paragallinarum* Modesto. Three of these transposases belong to IS families namely IS150 and IS605; the rest belong to the bacteriophage Mu transposase family. Comparing these transposases with one another indicate that they are all unique and share little homology, even the two transposases belonging to the IS150 family share very little homology.

The JCVI annotation data could assign 58% of ORF identified to protein function, 7% of the assigned functions implicated prophage function. With the use of this data two complete prophages could be assembled for *Av. paragallinarum* Modesto. One prophage resembles the Mu phages and the other resembles the HP2 phage from *H. influenzae*. Numerous lamboid genes were mapped and they seem to form a cryptic lamboid prophage.

With regards to the nature of these elements and their genetic presentation, they contributed to the inability to assemble the whole genome of *Av. paragallinarum* Modesto (Chapter 4).

The contribution of integrative plasmids, transposons, IS and bacteriophages with regards to virulence of micro-organisms have been widely published (Prescott *et al.*, 2002). All these elements have the capability to enhance the pathogenicity of their associated micro-organism. These elements may likely provide *Av. paragallinarum* with the appropriate mechanisms for emergence of new and virulent serovars (Prescott *et al.*, 2002).

5.6. Literature cited

Blackall, P. J., and R. Yamamoto. 1990. Infectious coryza. In: A laboratory manual for the isolation and identification of avian pathogens, 3rd ed. H. G. Parchase, L. H. Arp, C. H. Domermuth and J. E. Pearson, eds. *American association of avian pathologists, Ames, Iowa.* Pp. 27-31.

Balding, C., S. A. Bromley, R. W. Pickup and J. R. Saunders. 2005. Diversity of phage integrases in *Enterobacteriaceae*: development of markers for environmental analysis of temperate phages. *Environ. Microbiol.* **7**:1558-1567.

Bennet, P. M. 2008. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br. J. Pharmacol.* **153**:S347-S357.

Chen, X., J. K. Miflin, P. Zhang and P. J. Blackall. 1996. Development and amplification of DNA probes and PCR tests for *Haemophilus paragallinarum*. *Avian Dis.* **40**:398-407.

Groth, A. C. and M. P. Calos. 2003. Phage Integrases: Biology and Applications. *J. Mol. Biol.* **335**:667-678.

Groth, A. C., E. C. Olivares, B. Thyagarajan and M. P. Calos. 2000. A phage integrase directs efficient site-specific integration in human cells. *Proc. Natl. Acad. Sci. USA.* **97**:5995-6000.

Highlander, S. K., S. Weissenberger, L. E. Alvarez, G. M. Weinstock and P. B. Berget. 2006. Complete nucleotide sequence of a P2 family lysogenic

bacteriophage, ΦMhaA1-PHL101, from *Mannheimia haemolytica* serotype A1. *Virology*. **350**:79-89.

Hsu, Y. M., H. K. Shieh, W. H. Chen, T. Y. Sun and J. H. Shiang. 2007. Antimicrobial susceptibility, plasmid profiles and haemocin activities of Avibacterium paragallinarum strains. *Vet. Microbiol*. **124**:209-218.

Mahillon, J. and M. Chandler. 1998. Insertion Sequences. *Microbiol. Mol. Biol. Rev.* **62**:725-774.

Morgan G. J., G. F. Hatfull, S. Casjens and R. W. Hendrix. 2002. Bacteriophage Mu Genome Sequence: Analysis and Comparison with Mu-like prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. *J. Mol. Biol.* **317**:337-359.

Prescott, L. M., J. P. Harley and D. A. Klein. 2002. Microbiology 5th Edition. Pp294-820.

Resch, G., E. M. Kulik, F. S. Dietrich and J. Meyer. 2004. Complete Genomic Nucleotide Sequence of the Temperate Bacteriophage AaΦ23 of *Actinobacillus actinomycetemcomitans*. *J. Bacteriol.* **186**:5523-5528.

Rousseau, P., E. Gueguen, G. Duval-Valentin and M. Chandler. 2004. The helix-turn-helix motif of bacterial insertion sequence IS*911* transposase is required for DNA binding. *Nucleic Acids Res.* **32**(4):1335-1344.

Terry, T. D., Y. M. Zalucki, S. L. Walsh, P. J. Blackall and M. P. Jennings. 2003. Genetic analysis of a plasmid encoding haemocin production in *Haemophilus paragallinarum*. *Microbiology.* **149**:3177-3184.

Williams, B. J., M. Golomb, T. Phillips, J. Brownlee, M. V. Olson and A. L. Smith. 2002. Bacteriophage HP2 of *Haemophilus influenzae*. *J. Bacteriol.* **184**:6893-6905.

Chapter 6
Concluding Remarks

Nicotinamide adenine dinucleotide (NAD^+) plays an important role in various catabolic and anabolic reactions and it is of considerable importance in cellular metabolism. NAD^+ shuttle between its oxidized form, NAD^+, and its' reduced form, NADH, but the total concentration remains constant within the cells. NAD^+ is either produced biosynthetically or through recycling pathways and different combinations of these metabolic routes result in a widespread mixture of NAD^+ synthetic machinery in various species.

Avibacterium paragallinarum forms part of the *Pasteurellaceae* family of which certain members require exogenous NAD^+ for growth since they are unable to synthesize and recycle their own. As a result their basic NAD^+ pathway consists of an uptake system with negligible resynthesis activity and for that reason, these organisms rely on the supply of extracellular NAD^+. Genes involved in the NAD^+-dependent pathway for *Av. paragallinarum* were implicated with the use of genomic information obtained from the whole genome sequencing project for the NAD^+-dependent *Av. paragallinarum* Modesto. Genes implicated include *pnuC* (Ribosyl nicotinamide transporter); *nadR* (Nicotinamide mononucleotide adenylyltransferase / Nicotinamide riboside kinase) and *ppnK* (NAD^+ kinase).

Various strains of *Av. paragallinarum* has the ability to grow without NAD^+ or NAD^+ precursors and within some members of the *Pasteurellaceae* this NAD^+-independent trait can be ascribed to the presence of a *nadV* gene either chromosomally or plasmid located. This however is not the case for *Av. paragallinarum* NAD^+-independent strain 1750 since no *nadV* gene was obtained either plasmid or chromosomally located. An *nadC* and a partial *nadE* gene located on the genomic DNA were identified for this strain. The

presence of these genes does not complete the NAD^+-independent pathway for *Av. paragallinarum* but it does seem to indicate that *Av. paragallinarum* follows the classical route for NAD^+ production as set out for bacteria in general.

In order to unravel the genome of *Av. paragallinarum*, a whole genome sequencing project was launched. 454 Pyrosequencing technology was applied to obtain the whole genome sequence of *Av. paragallinarum* Modesto. This specific serovar was selected because currently it is the most virulent serovar prevalent in South Africa. Initially a pyrosequencing run was performed on a GS-20 instrument with an average read length of 100 bp resulting in a 11 X sequence coverage. Analysis of this data set produced 787 contigs of which the smallest was 85 bp long and the largest 43 928 bp. In addition to the excessive amount of contigs, which was discouraging for any form of genomic sequence analysis, a reasonable amount of repeats either within a contig or shared between contigs, was observed. A second sequencing run was performed on a GS-FLX Titanium instrument with an average read length of 400 bp which drastically reduced the number of contigs to 300 with a 27 X sequence coverage.

Mapping the GS-20 and the Titanium contig sets to one another was not possible as the coverage of the GS-20 contig set was not close to the required coverage and this suggested that the GS-20 contig set should not be used. A joint assembly was performed with the GS-20 and Titanium read set which was used during contra-mapping in an attempt to decrease the total number of Titanium contigs. Through this method, the total number of contigs was further decreased from 300 to 268 and indicated a minimal genome size of 2 561 700 bases. This is approximately 16% larger than that of any of the known genome sizes from members of the *Pasteurellaceae.* The inability to assemble the genome is ascribed to the various repeat regions present in the genome of *Av. paragallinarum*.

The initial 300 Titanium contigs were submitted for annotation to the J. Craig Venter Institute. More proteins were implicated in cell envelope functions than cellular metabolism which could explain the variety of serovars and serotypes observed within this bacterium. Various proteins were also implicated to play a role in mobile and extrachromosomal element functions.

These mobile and extrachromosomal elements included an integrated plasmid, numerous site-specific integrases, and transposases and prophage proteins. This study confirmed and illustrated an integrated p250 from *Av. paragallinarum*. Nine site-specific integrases were identified which share little homology with one another. Three of the 10 transposases identified are IS- and the rest are phage-transposases. These transposases share little homology with each other including the two transposases associated with the same IS-family. Numerous prophage functions were assigned which led to the identification of a putative Mu-like prophage for *Av. paragallinarum*, ΦAvpmuC-2M, as well as a HP2-like prophage ΦAvpC-2M-HP2.

During the latter part of this study we were contacted by Professor Henrik Christensen from the Department of Veterinary Disease Biology, Center for Applied Bioinformatics, Copenhagen University, Denmark. This research group also performed a titanium pyrosequencing run on the NAD^+-dependent *Av. paragallinarum* A-1 (0083) type strain. Prof. Christensen was kind enough to supply their raw contig set which included a total number of 485 contigs with the largest contig 51 124 bp and the smallest contig 106 bp. This contig set was analysed in parallel to that of the C-2 serovar. Similar problems were experienced in terms of the reduction of the number of contigs and this confirms that the difficulties experienced are species related rather than strain specific. It should also be noted that the combination of these two contig sets did not lead to the reduction in the number of contigs.

Analysis of the above mentioned A-1 raw contig set provided similar information regarding the NAD^+ pathways for the *Av. paragallinarum* A-1 as reported in this study for *Av. paragallinarum* Modesto. An identical NAD^+-dependent pathway was detected in both serovars and no NAD^+ *de novo* synthesis genes could be identified. Moreover, both contig sets confirmed the integration of plasmid p250 within the genome of *Av. paragallinarum* and indicated that it is unlikely that plasmid pA14 as described by Hsu *et al.* (2007) exists within *Av. paragallinarum* Modesto and 0083. Although prophage genes are also present within the genome of *Av. paragallinarum* A-1 it seems that A-1 and C-2 may not share all the prophage moieties.

Chapter 7
Summary

Avibacterium paragallinarum is an avian pathogen and has the ability to cause vast economical losses. This bacterium forms part of the *Pasteurellaceae* family and factors contributing to pathogenicity, immunogenicity and serotyping are not clearly understood.

One of the main questions that were addressed in this study was the identification of genetic tools are that is responsible for the NAD^+-independence ability of this organism. NAD^+ recycling genes were implicated for this bacterium and *Av. paragallinarum* seem to follow the recycling pathway as set out for the *Pasteurellaceae* family. Still the question regarding NAD^+-independence remains unanswered as no complete pathway could be implicated for this trait. Furthermore, no plasmid(s) could be isolated that conferred this trait and no complete NAD^+ synthesis pathway could be implicated. Only two genes were identified to form part of a NAD^+-independent pathway which indicated that *Av. paragallinarum* may use genetic tools for NAD^+-independence as set out for bacteria in general and not as the rest of the *Pasteurellaceae* family members.

Plasmid isolation studies revealed only the plasmid p250 identical to the established plasmid p250 for *Av. paragallinarum*. Literature reports on two other native plasmids namely pYMH5 and pA14, however, these two plasmids could not be detected in any of the strains used in this study during plasmid screens. This study confirms and illustrates the integration of plasmid p250 within the genome of *Av. paragallinarum* by different serovars.

In order to study this bacterium on a genomic level a whole genome sequencing project was launched. This project experienced vast amount of

difficulties regarding the assembly of a complete genome for *Av. paragallinarum*. This bacterium contains numerous repeated regions within its genome which, with current sequencing technology, prevent the genome assembly into a single chromosome. A complete assembled genome can only be achieved when longer read lengths are available in high-throughput sequencing technologies and will thus enable clarification regarding the repeated regions.

A pseudo-assembled molecule was sent to the JCVI for genome annotation. Annotated data revealed that *Av. paragallinarum* contains a high degree of complexity regarding its cell envelope which may shed light on pathogenicity and immunogenicity problems endured. This bacterium contains numerous mobile and extrachromosomal elements which contribute to the inability to close the genome of *Av. paragallinarum*. Nine site-specific integrases and 10 transposases were identified. These tools could allow for a high degree of genetic diversity within the *Av. paragallinarum* specie. Alongside these tools two putative prophages was identified for *Av. paragallinarum*. One prophage resembles the Mu-like prophages and was termed ΦAvpmuC-2M, the other resembles the HP2-like prophages and was termed ΦAvpC-2M-HP2.

Keywords: Avibacterium paragallinarum; NAD+-dependence; NAD+-independence; NAD+ pathways; whole genome sequencing; repeated regions; integrases; transposases; prophages.

Chapter 8
Opsomming

Avibacterium paragallinarum is 'n voëlspesiepatogeen en kan tot groot finansiële skade lei. Hierdie bakterium vorm deel van die *Pasteurellaceae* familie en faktore wat bydrae tot die patogenisiteit; die immuunresponsreaksie en serotipering van hierdie bakterium is onduidelik.

Een van die vrae wat geadresseer is in hierdie studie het betrekking op die teenwoordigheid van genetiese inligting wat dit moontlik maak vir die organisme om onafhanklik te wees van NAD^+. NAD^+ herwinningsgene vir die bakterium is getoon in hierdie studie en dit blyk dat *Av. paragallinarum* die weg vir herwinning volg soos die res van die *Pasteurellaceae* familie. Die vraag na die NAD^+-onafhanklike weg bly huidiglik onbeantwoord aangesien geen plasmied(e), wat hierdie eienskap toelaat, geïsoleer is uit NAD^+-onafhanklike stamme nie. Daarmee saam blyk dit dat daar geen volledige NAD^+-onafhanklike weg teenwoordig is nie. Slegs twee gene wat deel vorm van die NAD^+-onafhanklike weg soos uiteengesit vir bakterieë oor die algemeen is in hierdie studie bevestig. Hierdie is in teenstelling met die weë soos aangedui vir die *Pasteurellaceae* familie.

Plasmied isolasies het slegs 'n identiese plasmied aan p250 opgelewer soos reeds beskryf in die literatuur. Literatuur noem ook die bestaan van twee ander plasmiede eie aan *Av. paragallinarum* naamlik pYMH5 en pA14. Nie een van hierdie plasmiede is egter geïsoleer uit stamme gebruik tydens plasmiedstudies nie. Hierdie studie bevestig en illustreer ook die feit dat plasmied p250 integreer in die genoom van verskeie serovars.

Om die bakterium beter te kan verstaan op genomiese vlak is 'n heelgenoom basispaaropeenvolgingstudie van stapel gestuur. Hierdie projek het eindelose

probleme opgelewer rakende die samevoeging van 'n volledige genoom vir *Av. paragallinarum*. Een van die hoof struikelblokke wat ervaar was, is die feit dat die genoom van hierdie bakterium oor veelvuldige herhalende gebiede beskik. Hierdie gebiede maak dit onmoontlik om die genoom volledig saam te voeg in 'n enkele chromosoom. 'n Voltooide genoom kan slegs verkry word indien hoë kapasiteit basispaarbepalingstegnologie langer leeslengtes toelaat wat dit moontlik mag maak om die probleem met herhalende gebiede op te los.

'n Sinteties saamgevoegde DNA molekuul is na die JCVI gestuur vir genoomannotering. Hierdie resultate het getoon dat *Av. paragallinarum* 'n hoë vlak van kompleksiteit beskik rakende die sel buitewand wat probleme rakende immuun response kan verklaar. Hierdie bakterium beskik ook oor heelwat mobiele en ekstra-chromosomale elemente wat daartoe bydra dat die genoom nie voltooi kan word nie. Nege plek-spesifieke integrasegene en 10 transposasegene is geïdentifiseer. Hierdie gene maak dit moontlik vir *Av. paragallinarum* om genetiese materiaal uit te ruil wat dus bydra tot genetiese diversiteit binne hierdie spesie. Daarmee saam is twee moontlike profage geïdentifiseer waarvan een verwant is aan die Mu-tipe fage en word dus genoem ΦAvpmuC-2M en die ander fage verwant is aan die HP2-tipe en is genoem ΦAvpC-2M-HP2.

Sleutelwoorde: Avibacterium paragallinarum; NAD$^+$-afhanklik; NAD$^+$-onafhanklik; NAD$^+$ weë; geheel-genoom-basispaaropeenvolging; herhalende gebiede; integrase gene; transposase gene; profage.