CHAPTER II

A Comparative Perspective on the Genomes of *Histophilus somni* Strains 2336 and 129Pt

ABSTRACT

Whole genome sequencing is an eventful strategy for providing the molecular foundation for an organism. In view of the economic and biological importance of *H. somni*, the genomes of pneumonia strain 2336 and preputial strain 129Pt have been sequenced. The size of *H. somni* strain 2336 chromosome is 2,263,870 bp and contains 2,160 open reading frames (ORFs). The size of *H. somni* strain 129Pt chromosome is 2,008,359 bp and contains 1,890 ORFs. Using the genome sequence data and comparative analyses with other members of the *Pasteurellaceae*, several *H. somni* genes that may be involved in virulence have been identified. Most of these genes/loci are located in different regions of the chromosomes and occur in opposite orientations, indicating genome rearrangement during evolution of the two *H. somni* strains. Since the genome of strain 129Pt is ~250,000 bp smaller than that of strain 2336, these two genomes provide yet another paradigm for studying evolutionary gene loss and/or gain in regard to virulence repertoire and pathogenic ability. Using the complete genome sequences, it has been possible to examine bacteriophage- and transposon-mediated horizontal gene transfer and gene loss at several loci in the chromosomes of *H. somni* strains 2336 and 129Pt. It is shown that these mobile genetic elements have played a significant role in creating genomic diversity and phenotypic variability among the two *H. somni* strains. In conclusion, comparative analyses of the genomes have provided new insights into the evolution of *H. somni* strains.

Introduction

Large scale sequencing of DNA is a powerful method to obtain genetic information about an organism in a short period of time. A "genome" is defined as "the sum total of all the genetic information (contained in the chromosomes, plasmids, and phages) belonging to an organism". The terms "genomics" and "genomic methods" describe "the molecular and bioinformatics techniques that employ all or part of the genome to answer a question about an organism or a group of organisms" (Buckley, 2004). The advent of genomics has found immense applications in the quest to understand nature, and comparative microbial genomics is an indispensable tool for modern molecular pathogenic bacteriology. The first complete genome sequence from a free-living organism was that of *Haemophilus influenzae* strain Rd KW20 (Fleischmann et al., 1995), a close relative of *H. somni*. This pioneering work at The Institute for Genomic Research (TIGR) introduced and popularized the concept of whole genome random sequencing by the 'shotgun' approach. Since the completion of the first bacterial genome sequence, a total of 471 microbial (435 bacterial and 36 archeal) genomes have been sequenced and annotated (NCBI microbial genome projects database, as of March 06 2007) using novel tools and techniques.

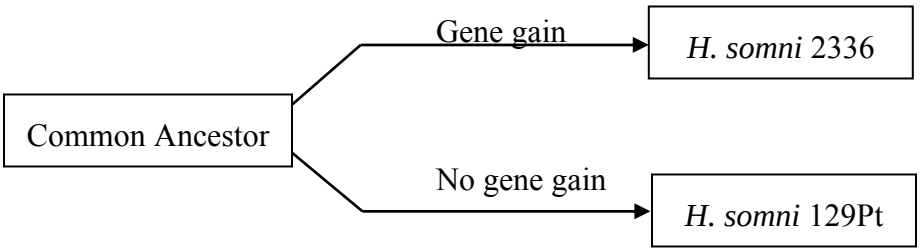
Among the *Pasteurellaceae*, the genomes of one or more species that are pathogenic to humans or animals from the genera *Actinobacillus*, *Pasteurella*, and *Mannheimia* have also been sequenced. The availability of these genome sequences has facilitated whole genome comparisons that have shed new light on the physiology and pathogenic evolution of the corresponding bacteria. Furthermore, whole genome analyses have been useful for "comprehensive genome phylogeny" and "prokaryotic genomic taxonomy" (Snel et al., 1999; Konstantinidis and Tiedje, 2005). These concepts have also raised the prospect of a meaningful classification of bacteria within the *Pasteurellaceae* (Coenye et al., 2005; Gioia et al., 2006).

Horizontal gene transfer (HGT: defined as the "acquisition of new genes either directly by transformation with naked DNA, transduction with phages, or the uptake of plasmids or chromosomal fragments by conjugation") plays a significant role in driving the evolution of pathogenic bacteria (Hacker et al., 1997). Reduction in genome size (referred to as reductive evolution) can occur as a result of continuous loss of genetic material due to gene deletion and/or mutation followed by DNA erosion (Olson, 1999). Comparative analyses of genome sequences of bacterial pathogens *in silico* using bioinformatics tools is a powerful method to infer their biochemical dissimilarity, virulence attributes, pathogenic ability, and adaptive evolution (Fraser et al., 2000). Comparative genome analyses have provided insights into the role of HGT in the evolution of different *Escherichia coli* strains (Lawrence and Ochman, 1997). By similar analyses, it has been proposed that *Mycobacterium leprae* has undergone a reductive evolution in relation to *Mycobacterium tuberculosis* (Cole et al., 2001). Comparisons of the genomes of three *Brucella* species have revealed similar genetic content and gene organization, despite the presence of genetic polymorphisms due to insertion-deletion events (Halling et al., 2005).

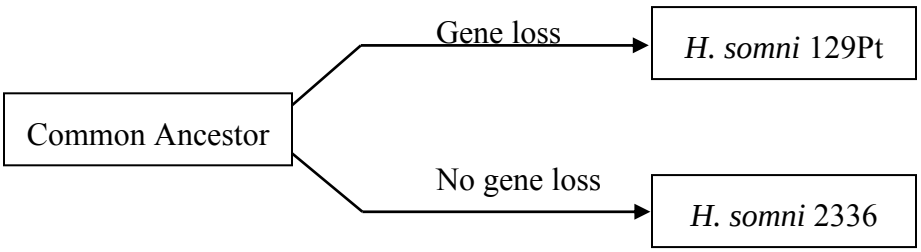
As outlined in chapter I, numerous *in vitro* and *in vivo* studies during the pre-genomic era have shed light on the differences in virulence properties and genetic traits between *H. somni* pathogenic isolates from sick animals and commensal isolates from the urogenital tract (Corbeil et al., 1995). *H. somni* pneumonia strain 2336 and preputial strain 129Pt have been phenotypically well characterized in the laboratory and utilized in several comparative studies (Inzana et al., 1992; Corbeil et al., 1997; Inzana et al., 2002). However, thus far only a few genes involved in metabolic and virulence functions have been identified and characterized in *H. somni* strains using DNA/DNA and DNA/Protein comparisons (Cole et al., 1993; Wu et al., 2000; McQuiston et al., 2000).

In the light of previous knowledge, three possible mechanisms that may have engendered the genetic differences between *H. somni* strains 2336 and 129Pt can be proposed, as below;

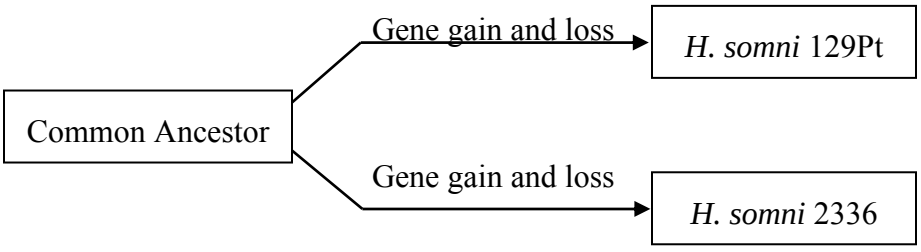
Mechanism I: The two strains originated from a common ancestor; only one of them has acquired genes by HGT while the other one has not acquired genes.



Mechanism II: The two strains originated from a common ancestor; only one of them has lost genes by deletion/mutation and undergone 'reductive evolution'.



Mechanism III: The two strains originated from a common ancestor; both have independently and continuously acquired and lost genes. The net loss or net gain of genes is a determinant of their divergent evolution.



By whole genome sequencing and comparisons, it may be possible to comprehend which of the above three mechanisms were responsible for genetic variability within the two strains.

A comparative genomics and bioinformatics approach is also expected to facilitate identification of *H. somni* genes putatively involved in virulence and pathogenesis. The *Microbial Genomics Workshop* conducted by the United States Department of Agriculture (USDA) in the year 2000 ranked the complete sequencing of the genome of *H. somni* among the top 15 priorities (USDA, 2000). Following the recommendations of this workshop, the genomes of *H. somni* strains 2336 and 129Pt have been sequenced. The principal objectives in sequencing the genomes of the *H. somni* strains were to;

(i) Create a *H. somni* genetic database and a catalog of putative genes

(ii) Compare and contrast the genetic profiles of *H. somni* strains 2336 and 129Pt

(iii) Provide a genomic perspective on the evolution of *H. somni* strains

In this chapter, data on the general features of the genomes of *H. somni* strains 2336 and 129Pt are presented. Several loci unique to each strain have been examined and these contain putative genes that encode proteases, restriction-modification (RM) enzymes, hemagglutinins, kinases, helicases, and transcription regulators. Most of the genes unique to each strain are shown to be associated with genetic elements such as prophages and/or transposons, which make up a significant portion of the genome in both *H. somni* strains. Chromosomal loci containing putative genes encoding hemagglutinins have been examined in detail along with regions containing prophage-like sequences.

Whole genome comparison has also been used to acquire information about the synteny of genes present on the chromosomes of *H. somni* and *H. influenzae* strains. Insights into the evolution of *H. somni* strains 2336 and 129Pt obtained through comparative genomic analyses have been discussed. Strategies for further analyses of these and other related genomes have been recommended.

Materials and Methods

Genome sequencing of *H. somni* strain 129Pt

The materials and methods used for sequencing the genome of *H. somni* strain 129Pt have been described by Challacombe et al. (2007).

Genome sequencing of *H. somni* strain 2336

This project was a collaboration between different institutions and draft sequencing of strain 2336 was performed at the Oklahoma University Health Sciences Center (OUHSC) and initial gap closure was performed at Virginia Tech. The finishing was performed at the Joint Genome Institute's Los Alamos National Laboratory (LANL).

Bulk genomic DNA (2 mg) from *H. somni* strain 2336 was purified using the Puregene protocol (Gentra Systems, Minneapolis, MN). The DNA was physically sheared in a nebulizer in 40% glycerol under pressure (6 psi N_2) in a dry ice-alcohol bath.. The nebulized DNA was ethanol precipitated, end-repaired (End-It DNA End-Repair Kit, Epicentre, Madison, WI) and size-selected by agarose gel electrophoresis to obtain 2-4 kb fragments. The fragments were gel purified, ethanol precipitated, then ligated into the *SmaI* site of plasmid vector pUC18 using the Fast-Link DNA ligase kit (Epicentre, Madison, WI) and transformed into *E. coli* strain XL-1 Blue MRF' (Stratagene) to obtain small insert libraries. Colonies that contained inserts were identified by blue-white screening on LB plates containing ampicillin (100µg/ml) and X-gal/IPTG. White colonies were transferred into 96-well blocks (Beckman Coulter) containing 1.5 ml of 2xYT broth with ampicillin (100 µg/ml) in each well (2 ml capacity). Colony picking was done using a Genomic Solutions G3 robot. The deep-well plate cultures were incubated for 18-22 hours at 37^0C.

The size of the *H. somni* genome was estimated to be ~2.2 Mb (Dr. Vivek Kapur, University of Minnesota, personal communication). According to Lander and Waterman (1988) calculations, the shotgun sequencing phase for this genome thus required ~35,200 sequence reads to reach 8-fold coverage. Since the colony-to-sequence read efficiency at the Laboratory for Genomics and Bioinformatics (LGB), Oklahoma University Health Sciences Center, was approximately 85%, it was necessary to pick at least 41,412 colonies/templates to obtain 35,200 reads (8 fold coverage) of sufficient quality for assembly, with the caveat that each colony would yield one template. A Beckman Biomek 2000 automated robotics workstation was used for template preparation following the method previously described (Chissoe et al., 1991). The purified DNA templates were dissolved in 50 µl distilled H_2O, and 2 µl of this sample was used for each DNA sequencing reaction.

After rearraying the templates into 384-well blocks, they were cycle-sequenced using Big-Dye terminators V2.0 or V3.0 (Applied Biosystems, Foster City, CA) with M13 forward and reverse universal primers. The DNA sequence reaction mixture was purified by passage through Sephadex G50 columns and was analyzed on an ABI 3700 capillary electrophoresis DNA sequencer.

Phred-Phrap was used for data analysis and initial assembly. Sequence data were transferred without intervening analysis from the sequencing computer to a Sun Enterprise 450 server for assembly and analysis. Each sequence file was automatically trimmed to remove vector sequence at the 5' end and low quality sequence at the 3' end, and was analyzed by Phred for data quality. Phrap was then used to assemble the accumulated files. Our default assembly parameters were a minmatch of 30 and a minscore of 55. In other words, the assembly software Phrap made pairwise comparisons between sequence reads to find a region between two

sequences where 30 bp in a given sequence read matched a second sequence read exactly. The assembler then examined either side of this match for additional similarity, using a point system where a match = 1 point, and a mismatch = -2. For the assembler to assemble two sequence reads into a contiguous sequence, the contig must have a minimum score of 55. These highly stringent assembly parameters ensured that contigs arriving in the final database were assembled with high fidelity.

The sequence data assembled with Phred-Phrap were viewed using Consed for assessing data quality and the design of closure experiments. Consed was also used to identify putative repeat regions so that the problems associated with assembling these regions could be resolved by way of combinatorial PCR experiments to isolate the repeat sequences on PCR amplicons. The location and exact sequence of each repeat was confirmed by isolating PCR fragments that contain each repeat in its entirety, followed by primer walking across the PCR product.

Gap closure of the *H. somni* genome relied on primer walking, with primer-directed finishing employed to raise sequence quality in low-coverage regions of the genome. This process was significantly automated by the use of AUTOFINISH (Gordon et al., 2001). For initial gap closure, Single Primer Amplification of Contig Ends (SPACE), which is similar to the single-primer PCR procedure for rapid identification of transposon insertion sites (Karlyshev et al., 2000), was the primary method used. In brief the PCR products from the SPACE amplification were sequenced using specific sequencing primers designed to bind downstream of the SPACE PCR primers. Sequences generated from SPACE PCR products as templates were used to assemble the contigs using Sequencher (Gene Codes Corporation, Ann Arbor, MI). Additional primers were designed to verify the correct assembly of contigs by confirmatory PCR.

Simultaneously, a fosmid library was constructed for scaffolding purposes using the vector pCC1fos (Epicentre) with 40 kb inserts. Sequencing of the fosmids was necessary to close gaps across sequences that occur more than once in the genome, such as insertion sequences and ribosomal genes. Additional gaps that had not been closed by SPACE-walking were closed using the already available partial sequence of *H. somni* strain 129Pt as a scaffold. Thus, a full length genomic sequence with no gaps and no low quality regions was generated. As mentioned above, the location and sequence of each repeat was confirmed by primer walking across a PCR product containing the repeat of interest.

During genome closure at Los Alamos National Laboratory, reads were assembled with parallel phrap (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher (unpublished, C. Han) or transposon bomb of bridging clones using an EZ::TN™ kit (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing, custom primer walking, or PCR amplification.

The National Human Genome Research Institute (NHGRI) standards for the human genome project (1 error per 10, 000 assembled bases) were followed to obtain sufficient quality genomic data for *H. somni*. The completed genome sequence of *H. somni* strain 2336 contained 26,472 reads, achieving an average of 9.7-fold sequence coverage per base with an error rate less than 1 in 100,000.

Preliminary annotation was done using the program Manatee created by the bioinformatics department at TIGR. After the final assembly, the Bacterial Annotation System (BASys) was used for gene prediction (Van Domselaar et al., 2005). Genes predicted by BASys were compared with those predicted by Manatee to enhance accuracy and to check for discrepancies.

Protein domains were identified by comparing each predicted protein against a Hidden Markov Model protein family database (Finn et al., 2006). To estimate the number of proteins unique to each strain, the Smith-Waterman algorithm (Smith and Waterman, 1981) was used to compare all predicted proteins from strain 129Pt against those from strain 2336. Proteins deemed to be unique to each strain were compared against the NCBI Non-Redundant protein database to determine whether they were hypothetical or conserved hypothetical.

NCBI GenePlot, which is a tool that compares the sequentially numbered genes along the entire length of two genomes based on the symmetrical best BLAST scores of the proteins they encode, was used to compare the order of genes in the chromosomes of *H. somni* strain 129Pt and *H. influenzae* strain Rd KW20 or *H. influenzae* strain 86-028NP.

NCBI TaxPlot, which is a tool for simultaneous comparison of related proteins encoded by three different genomes based on matching their BLAST scores, was used to analyze the similarity of genes present in the chromosomes of *H. somni* and *H. influenzae* strains. The genomes were compared for proteins of all functional categories with the BLAST score cutoff value set to 10 on a linear scale.

Gene acquisition and loss among the two strains were determined by comparing gene order, orientation of genes (forward/reverse), GC content of genes, features of intergenic regions, and the homology of proteins encoded by the genes at a locus of interest.

Protein and DNA sequences were aligned using the ClustalW multiple sequence alignment program. For protein sequence analyses, the ClustalW sequence submission form was used with default settings except the 'output format' set to 'pir'. The ClustalW 'pir' output was further analyzed using the program BOXSHADE. Linear maps of chromosomal loci were drawn using BioEdit. BioEdit, BOXSHADE, and ClustalW procedures are detailed in chapter III.

Results

General features of the chromosomes of *H. somni* strains 2336 and 129Pt

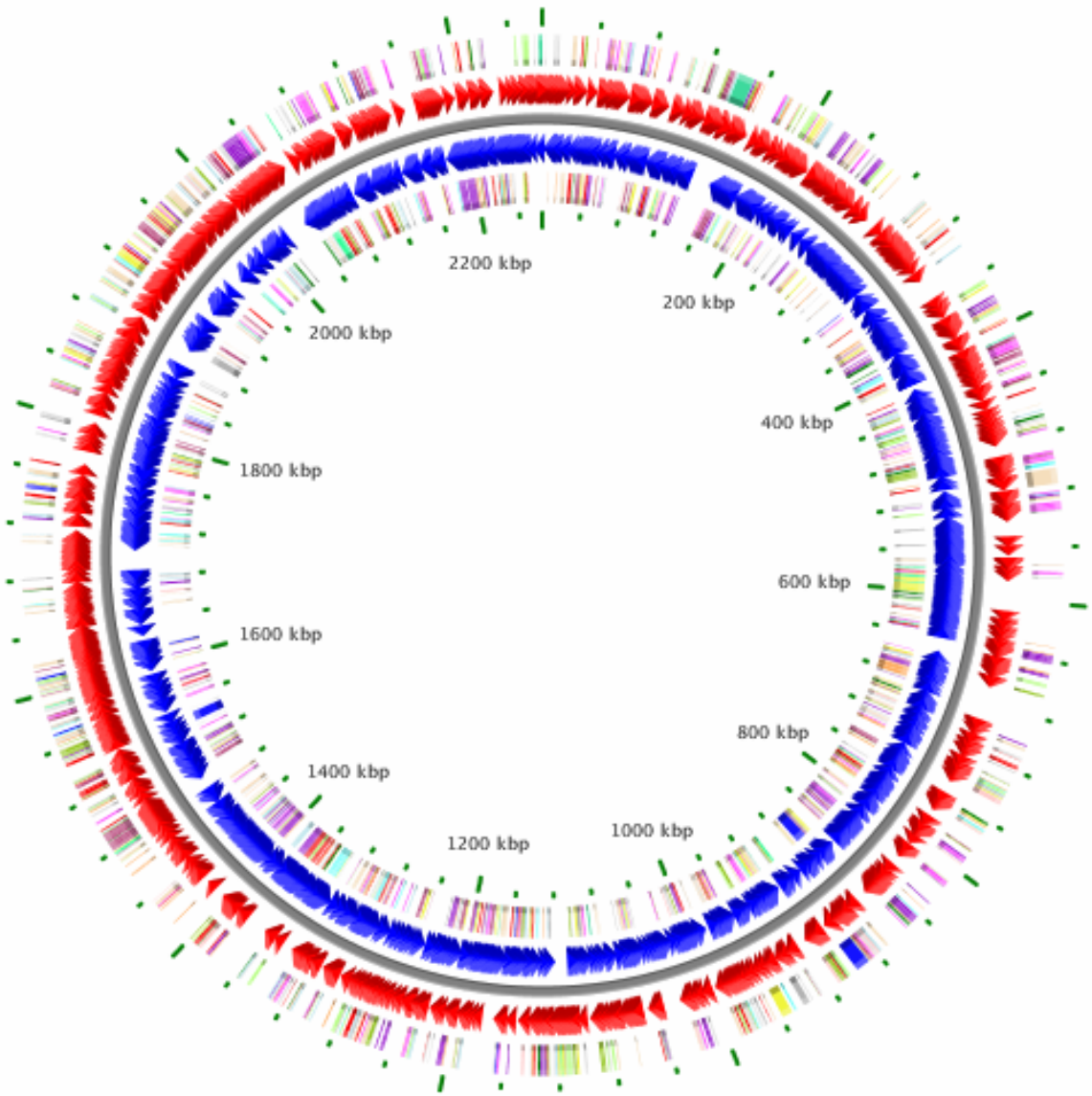
The length of *H. somni* strain 2336 chromosome is 2,263,870 bp and is bigger than that of *Haemophilus ducreyi* strain 35000HP (1,698,955 bp), *H. influenzae* strain Rd KW20 (1,830,137 bp), *H. influenzae* strain 86-028NP (1,913,428 bp), *H. somni* strain 129Pt (2,008,359 bp), *Actinobacillus actinomycetemcomitans* strain HK1651 (2,105,503 bp), *Neisseria gonorrhoeae* strain FA 1090 (2,153,922 bp), *Neisseria meningitidis* serogroup A strain Z2491 (2,184,406 bp), and *Pasteurella multocida* strain Pm70 (2,257,487 bp). However, the length of *H. somni* strain 2336 chromosome is smaller than that of *N. meningitidis* serogroup B strain MC58 (2,272,351 bp), *Mannheimia succiniciproducens* strain MBEL55E (2,314,078 bp), and *Mannheimia haemolytica* strain BAA-410 (~2,500,000 bp, draft sequence).

Circular maps of the chromosomes of *H. somni* strains 2336 and 129Pt showing the distribution of different categories of genes are presented in figs. 2.01 and 2.02, respectively. Using BASys, the chromosome of *H. somni* strain 2336 was predicted to encode 1,182 ORFs on the plus strand and 978 ORFs on the minus strand (total 2,160 ORFs). Whereas Challacombe et al. (2007) reported 1,844 ORFs in *H. somni* strain 129Pt, BASys predicted 1,010 ORFs on the plus strand and 880 ORFs on the minus strand in this strain (total 1,890 ORFs). For comparative purposes, predictions based on BASys were used. Although the chromosomes of the two strains differ in size by ~250,000 bp (equivalent of ~11%), their average GC content, gene density, and percentage of sequence that encodes proteins do not differ significantly. *H. somni* strain 2336 does not contain plasmids, but *H. somni* strain 129Pt contains a single, circular, multicopy plasmid (described in chapter III). Some of the other relevant features of the chromosomes of *H. somni* strains 2336 and 129Pt are compared in table 2.01.

A comparison of the orientation and order of genes on both DNA strands can be used to infer the degree of rearrangement of two related circular or linear chromosomes. An analysis of protein homologs using GenePlot revealed extensive colinearity of genes in the chromosomes of *H. influenzae* strains Rd KW20 and 86-028NP (Fig. 2.03), and lack of synteny of genes in the chromosomes of *H. somni* strain 129Pt and *H. influenzae* strains (Figs. 2.04 and 2.05). A similar comparison of the chromosomes of *H. somni* strains 129Pt and 2336 using Mummer showed several major inversion and deletion events (Fig. 2.06).

A three-way analysis of the proteins encoded by the chromosomes of *H. somni* strain 129Pt as well as *H. influenzae* strains Rd KW20 and 86-028NP using TaxPlot showed that 261 proteins were equally similar among the three organisms (Fig. 2.07) A similar analysis of the proteins encoded by the chromosomes of *H. somni* strain 2336 as well as *H. influenzae* strains Rd KW20 and 86-028NP showed that 297 proteins were equally similar among the three organisms (Fig. 2.08). However, both *H. somni* strains had more proteins with greater similarity to *H. influenzae* strain 86-028NP proteins than *H. influenzae* strain Rd KW20 proteins.

H. somni strain 129Pt contains seven regions that are associated with prophage-like sequences with a total length of ~82,000 bp (equivalent of ~4%). *H. somni* strain 2336 also contains seven regions that are associated with prophage-like sequences with a total length of ~198,000 bp (equivalent of ~8.7%). The chromosomal location and GC content of the 14 prophage regions from the two strains are compared in tables 2.02 and 2.03. Genes present in the seven prophage regions of *H. somni* strain 129Pt are shown in tables 2.04 to 2.10. Genes present in the seven prophage regions of *H. somni* strain 2336 are shown in tables 2.11 to 2.17. Prophage region IV of strain 129Pt is partially similar to prophage region VI of strain 2336. Prophage region VII of strain 129Pt is highly similar to prophage region I of strain 2336.



Chromosome Features

Genes encoding proteins
- Forward strand
- Reverse strand

Genes encoding functional RNA
- Forward strand
- Reverse strand

COG functional categories

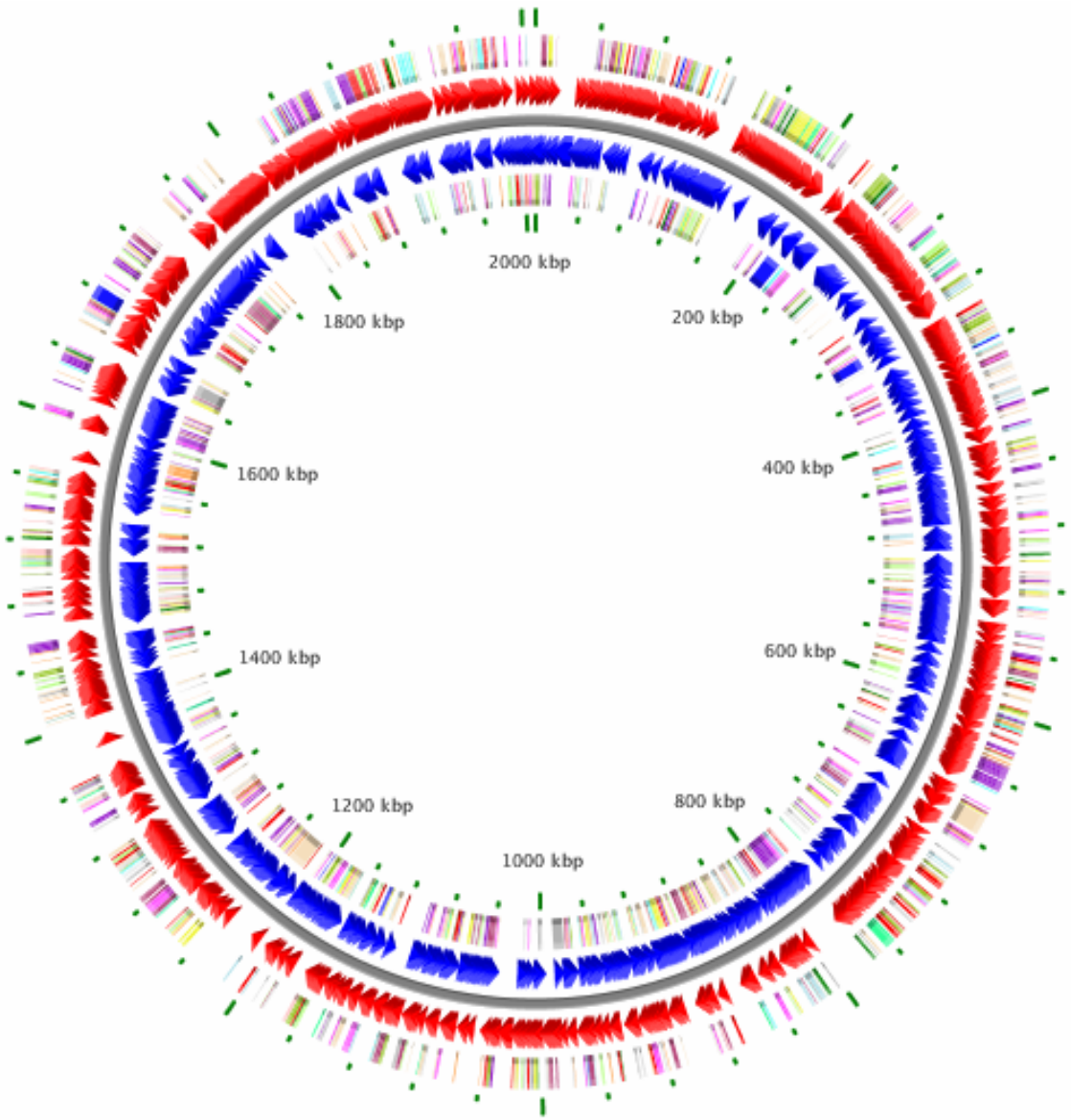
Information storage and processing
- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair

Cellular processes
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover, chaperones
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport and metabolism
- Signal transduction mechanisms

Metabolism
- Energy production and conversion
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme metabolism
- Lipid metabolism
- Secondary metabolites biosynthesis, transport and catabolism

Poorly characterized
- General function prediction only
- Function unknown

Fig. 2.01: Map of *H. somni* strain 2336 chromosome (Length: 2,263,870 bp; Genes: 2,160)



Chromosome Features

Genes encoding proteins
- Forward strand
- Reverse strand

Genes encoding functional RNA
- Forward strand
- Reverse strand

COG functional categories

Information storage and processing
- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair

Cellular processes
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover, chaperones
- Cell envelope biogenesis, outer membrane
- Cell motility and secretion
- Inorganic ion transport and metabolism
- Signal transduction mechanisms

Metabolism
- Energy production and conversion
- Carbohydrate transport and metabolism
- Amino acid transport and metabolism
- Nucleotide transport and metabolism
- Coenzyme metabolism
- Lipid metabolism
- Secondary metabolites biosynthesis, transport and catabolism

Poorly characterized
- General function prediction only
- Function unknown

Fig. 2.02: Map of *H. somni* strain 129Pt chromosome (Length: 2,008,359 bp; Genes: 1,890)

Table 2.01: Characteristics of the genomes of *H. somni* strains

Genome feature	***H. somni* strain 2336**	***H. somni* strain 129Pt**
Chromosome size	2,263,870 bp	2,008,359 bp
Number of ORFs	2,160	1,890
Gene density (bp per gene)	1048	1062
G+C content	37%	37%
5S ribosomal RNA genes	5	5
16S ribosomal RNA genes	5	5
23S ribosomal RNA genes	5	5
Number of tRNA genes	46	50
Plasmids	None	pHS129 (5,178 bp)
Regions containing prophage-like sequences	7	7
GenBank accession number	AACJ00000000	CP000436

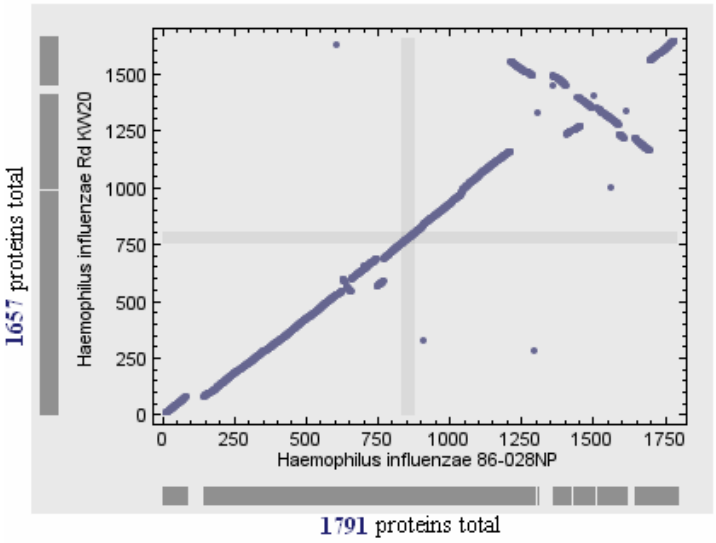


Fig. 2.03: GenePlot comparison of proteins from two strains of *H. influenzae*. Each point represents a pair of proteins from the two organisms showing a symmetrical best BLAST score; the coordinates of each point correspond to the position of the protein genes in the two genomes.

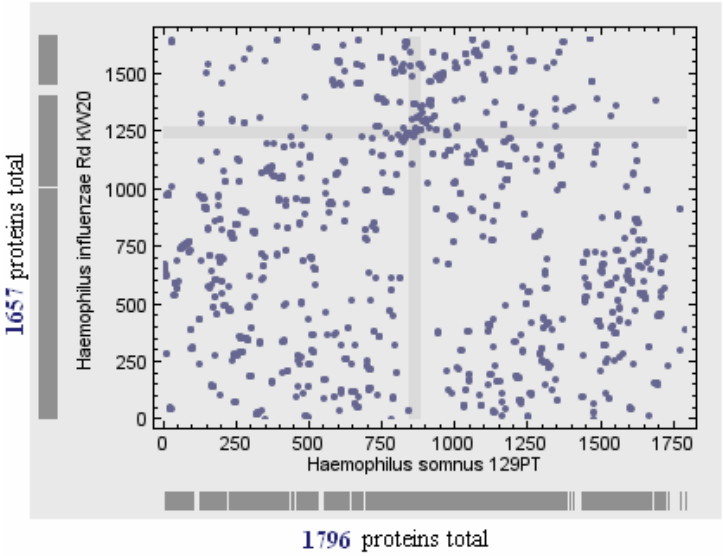


Fig. 2.04: GenePlot comparison of proteins from *H. somni* strain 129Pt and *H. influenzae* strain Rd KW20. Each point represents a pair of proteins from the two organisms showing a symmetrical best BLAST score; the coordinates of each point correspond to the position of the protein genes in the two genomes.

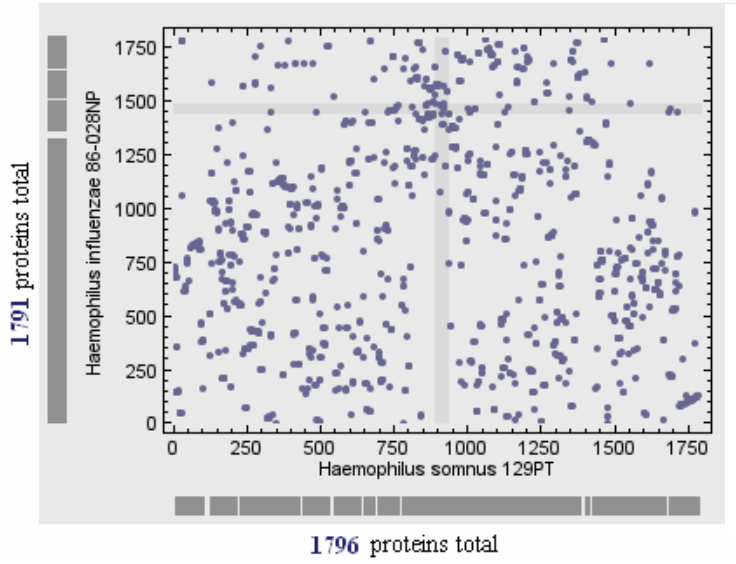


Fig. 2.05: GenePlot comparison of proteins from *H. somni* strain 129Pt and *H. influenzae* strain 86-028NP. Each point represents a pair of proteins from the two organisms showing a symmetrical best BLAST score; the coordinates of each point correspond to the position of the protein genes in the two genomes.

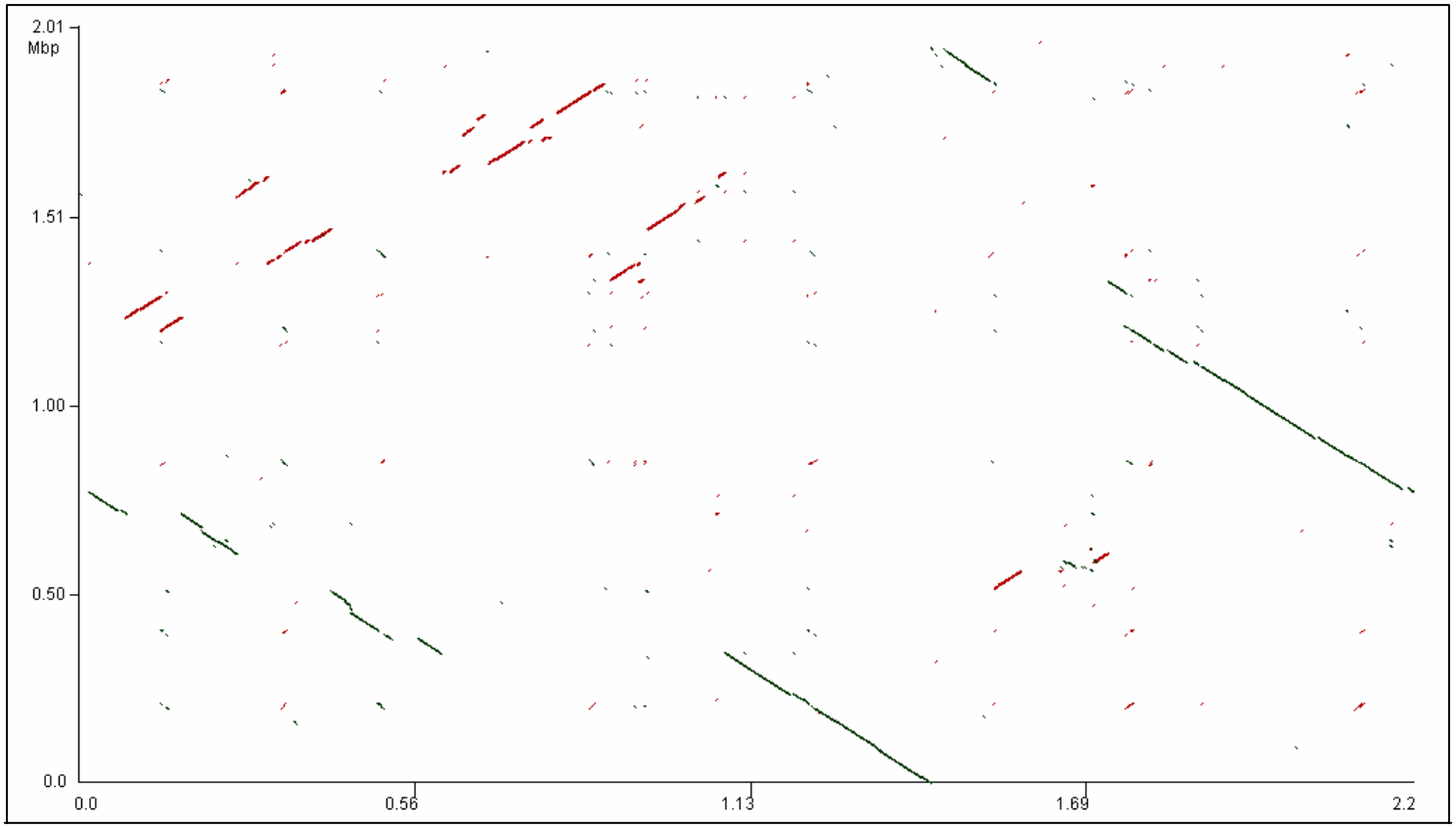


Fig. 2.06: Comparison of proteins from *H. somni* strains 129Pt and 2336 using Mummer

X-axis; *H. somni* strain 2336, Y-axis; *H. somni* strain 129Pt.

Forward matches are shown in red and reverse matches are shown in green.

Distribution of *H. somni* 129Pt homologs

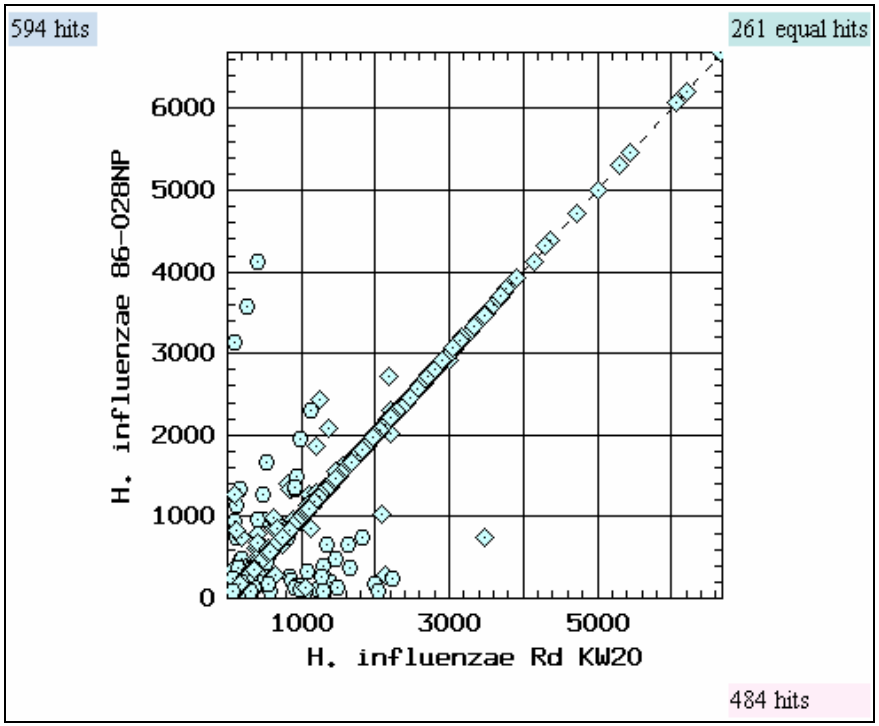


Fig. 2.07: TaxPlot for two strains of *H. influenzae* against *H. somni* strain 129Pt as the reference genome; 1,796 reference genome query proteins produced 1,339 hits.

Distribution of *H. somni* 2336 homologs

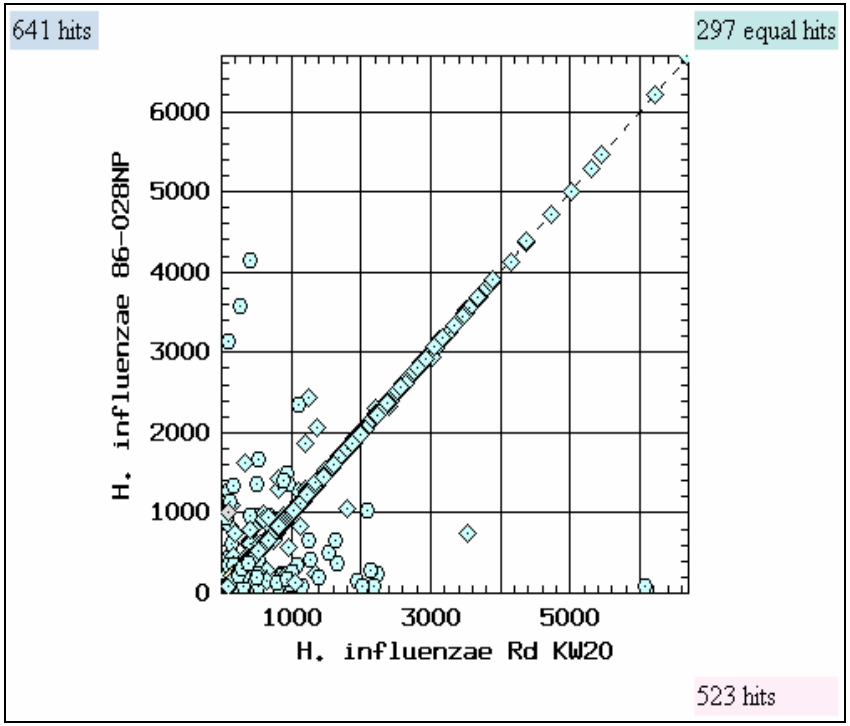


Fig. 2.08: TaxPlot for two strains of *H. influenzae* against *H. somni* strain 2336 as the reference genome; 2,063 reference genome query proteins produced 1,461 hits.

Each circle represents a single reference genome protein, plotted by its BLAST scores to the highest scoring protein from *H. influenzae* strain Rd KW20 (X-axis) and *H. influenzae* strain 86-028NP (Y-axis). Symmetrical hits are shown as diamonds.

Table 2.02: Characteristics of *H. somni* strain 129Pt prophage-like sequences*

Region name	Similar Region found in strain 2336	Chromosomal location	Total length	GC content
Prophage region I	Yes**	384 bp to 1,182 bp	798 bp	28.7%
Prophage region II	No	452,273 bp to 462,090 bp	9,817 bp	31%
Prophage region III	No	562,270 bp to 573,782 bp	11,512 bp	35%
Prophage region IV	Yes	589,932 bp to 593,221 bp	3,289 bp	33%
Prophage region V	No	1,379,985 bp to 1,381,127 bp	1,142 bp	41.3%
Prophage region VI	No	1,525,271 bp to 1,529,350 bp	4,079 bp	36%
Prophage region VII	Yes	1,558,147 bp to 1,610,766 bp	52,619 bp	40%

* Challacombe et al. (2007) predicted only five candidate prophage regions

** Similar gene found in strain 2336

Table 2.03: Characteristics of *H. somni* strain 2336 prophage-like sequences

Region name	Similar Region found in strain 129Pt	Chromosomal location	Total length	GC content
Prophage region I	Yes	266,093 bp to 319,633 bp	53,540 bp	39.5%
Prophage region II	No	528,571 bp to 574,528 bp	45,957 bp	42%
Prophage region III	No	1,056,356 bp to 1,079,503 bp	23,147 bp	39%
Prophage region IV	No	1,440,833 bp to 1,459,716 bp	18,883 bp	35%
Prophage region V	No	1,650,000 bp to 1,661,986 bp	11,986 bp	34.8%
Prophage region VI	Yes	1,703,580 bp to 1,719,485 bp	15,905 bp	35%
Prophage region VII	No	2,253,960 bp to 2,263,870 bp and 1 bp to 19,372 bp	29,282 bp	40%

Prophage regions containing homologous genes with partially conserved gene order among strains 2336 and 129Pt are shaded with similar colors in the tables 2.02 and 2.03.

Table 2.04: List of genes in prophage region I of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog
ABI24282	Phage integrase	ZP_00132752

Table 2.05: List of genes in prophage region II of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog
ABI24701	Possible phage-related protein	No homolog
ABI24702	Phage integrase	ZP_00132557
ABI24703	Hypothetical protein	No homolog
ABI24704	Possible phage DNA-polymerase or DNA-primase	No homolog
ABI24705	Possible virulence-associated protein E	No homolog
ABI24706	Conserved hypothetical protein	No homolog
ABI24707	Conserved hypothetical protein	No homolog
ABI24708	Type I restriction modification system, M protein	ZP_00133477
ABI24709	Type I restriction modification system, S protein	ZP_00204654
ABI24710	Type I restriction modification system, R protein	No homolog
ABI24711	Conserved hypothetical protein	ZP_00204701

Table 2.06: List of genes in prophage region III of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog
ABI24796	Phage integrase	ZP_00133240
ABI24797	Hypothetical protein	No homolog
ABI24798	Hypothetical protein	ZP_00132754
ABI24799	Possible phage transcriptional regulator	ZP_00132755
ABI24800	Conserved hypothetical protein	No homolog
ABI24801	Hypothetical protein	No homolog
ABI24802	Conserved hypothetical protein	ZP_00132552
ABI24803	Hypothetical protein	No homolog
ABI24804	Hypothetical protein	ZP_00204722
ABI24805	Conserved hypothetical protein	ZP_00348128
ABI24806	Hypothetical protein	No homolog
ABI24807	Hypothetical protein	No homolog
ABI24808	Hypothetical protein	No homolog
ABI24809	Conserved hypothetical protein	No homolog
ABI24810	Phage DNA primase-like protein	ZP_00133253
ABI24811	Cytotoxic translational repressor of toxin-antitoxin stability system	ZP_00348202
ABI24812	Possible transcriptional regulator	ZP_00348201
ABI24813	Conserved hypothetical protein	No homolog
ABI24814	Conserved hypothetical protein	No homolog
ABI24815	Conserved hypothetical protein	No homolog
ABI24816	Hypothetical protein	No homolog

Table 2.07: List of genes in prophage region IV of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog
ABI24831	Type I restriction modification system, S protein	ZP_00132561
ABI24832	Phage integrase	ZP_00132557
ABI24833	Type I restriction modification system, S protein	ZP_00132561
ABI24834	Phage P1-related protein	No homolog
ABI24835	Phage P1-related protein	No homolog

Table 2.08: List of genes in prophage region V of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog
ABI25494	*Haemophilus*-specific protein, uncharacterized	ZP_00204638
ABI25495	Phage integrase	ZP_00204639

Table 2.09: List of genes in prophage region VI of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog
ABI25606	Hypothetical protein	No homolog
ABI25607	Hypothetical protein	No homolog
ABI25608	*Haemophilus*-specific protein, uncharacterized	No homolog
ABI25609	Possible phage terminase, small subunit	No homolog
ABI25610	Possible prophage CP4-57 regulatory protein	No homolog
ABI25611	Hypothetical protein	No homolog
ABI25612	Bacteriophage P4 integrase	ZP_00133240

Table 2.10: List of genes in prophage region VII of *H. somni* strain 129Pt

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog*
ABI25640	*Haemophilus*-specific protein, uncharacterized	ZP_00131966
ABI25641	*Haemophilus*-specific protein, uncharacterized	ZP_00131965
ABI25642	*Haemophilus*-specific protein, uncharacterized	ZP_00131964
ABI25643	*Haemophilus*-specific protein, uncharacterized	ZP_00131963
ABI25644	Conserved hypothetical protein	ZP_00131962
ABI25645	Conserved hypothetical protein	ZP_00131961
ABI25646	Conserved hypothetical protein	ZP_00132834
ABI25647	Possible phage-related tail fiber protein	ZP_00132160
ABI25648	*Haemophilus*-specific protein, uncharacterized	ZP_00131957
ABI25649	*Haemophilus*-specific protein, uncharacterized	ZP_00133217
ABI25650	*Haemophilus*-specific protein, uncharacterized	ZP_00131956
ABI25651	*Haemophilus*-specific protein, uncharacterized	ZP_00131955
ABI25652	Conserved hypothetical protein	ZP_00131954
ABI25653	Conserved hypothetical protein	ZP_00131953
ABI25654	*Haemophilus*-specific protein, uncharacterized	ZP_00131952
ABI25655	Possible transcription regulator	ZP_00131951
ABI25656	Conserved hypothetical protein	ZP_00131950
ABI25657	Conserved hypothetical protein	No homolog
ABI25658	Hypothetical protein	No homolog

Table 2.10 continued from previous page

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog*
ABI25659	*Haemophilus*-specific protein, uncharacterized	ZP_00133493
ABI25660	Conserved hypothetical protein	ZP_00348111
ABI25661	Conserved hypothetical protein	ZP_00348112
ABI25662	Conserved hypothetical protein	ZP_00348114
ABI25663	Possible phage-related lysozyme	ZP_00348115
ABI25664	*Haemophilus*-specific protein, uncharacterized	ZP_00348116
ABI25665	Transposase	ZP_00133241
ABI25666	Transposase	ZP_00133242
ABI25667	Conserved hypothetical protein	No homolog
ABI25668	Conserved hypothetical protein	No homolog
ABI25669	Bacteriophage Lambda NinG recombination protein	ZP_00133501
ABI25670	Conserved hypothetical protein	ZP_00133502
ABI25671	Conserved hypothetical protein	ZP_00133503
ABI25672	Conserved hypothetical protein	No homolog
ABI25673	Conserved hypothetical protein	ZP_00133504
ABI25674	Conserved hypothetical protein	ZP_00133505
ABI25675	possible DNA methylase	ZP_00133506
ABI25676	Conserved hypothetical protein	ZP_00133507
ABI25677	Conserved hypothetical protein	ZP_00348118
ABI25678	Conserved hypothetical protein	ZP_00348119
ABI25679	Conserved hypothetical protein	No homolog
ABI25680	Conserved hypothetical protein	ZP_00133340
ABI25681	Conserved hypothetical protein	ZP_00204636
ABI25682	*Haemophilus*-specific protein, uncharacterized	No homolog
ABI25683	*Haemophilus*-specific protein, uncharacterized	No homolog

Table 2.10 continued from previous page

H. somni 129Pt NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 2336 homolog*
ABI25684	*Haemophilus*-specific protein, uncharacterized	No homolog
ABI25685	*Haemophilus*-specific protein, uncharacterized	No homolog
ABI25686	Hypothetical protein	No homolog
ABI25687	*Haemophilus*-specific protein, uncharacterized	No homolog
ABI25688	Conserved hypothetical protein	No homolog
ABI25689	Hypothetical protein	No homolog
ABI25690	Conserved hypothetical protein	ZP_00133504
ABI25691	*Haemophilus*-specific protein, uncharacterized	ZP_00204658
ABI25692	*Haemophilus*-specific protein, uncharacterized	ZP_00204659
ABI25693	*Haemophilus*-specific protein, uncharacterized	ZP_00133436
ABI25694	Possible phage recombinase	No homolog
ABI25695	Conserved hypothetical protein	No homolog
ABI25696	Conserved hypothetical protein	ZP_00133438
ABI25697	Cyt_C5_DNA_methylase	ZP_00133440
ABI25698	*Haemophilus*-specific protein, uncharacterized	No homolog
ABI25699	Conserved hypothetical protein	ZP_00132755
ABI25700	Phage integrase	ZP_00133240

*Genes listed in the third column that are also present in prophage region I of strain 2336 (table 2.11) in a similar order have been shaded with analogous colors.

Table 2.11: List of genes in prophage region I of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00131968	COG4383: Mu-like prophage protein gp29	No homolog
ZP_00348204 ZP_00348198	Hypothetical protein	No homolog
ZP_00348203	Hypothetical protein	No homolog
ZP_00132153	Hypothetical protein, bacteriophage-related	No homolog
ZP_00131966	Hypothetical protein	ABI25640
ZP_00131965	Hypothetical protein	ABI25641
ZP_00131964	Hypothetical protein	ABI25642
ZP_00131963	Hypothetical protein	ABI25643
ZP_00131962	Conserved hypothetical protein	ABI25644
ZP_00131961	Conserved hypothetical protein, bacteriophage-related	ABI25645
ZP_00131960	Conserved hypothetical protein, bacteriophage-related	No homolog
ZP_00348202	COG2026: Cytotoxic translational repressor of toxin-antitoxin stability system	ABI24811
ZP_00348200	COG0625: Glutathione S-transferase	ABI25647
ZP_00131957	Conserved hypothetical protein, bacteriophage-related	ABI25648
ZP_00131956	Conserved hypothetical protein	ABI25650
ZP_00131955	Conserved hypothetical protein	ABI25651
ZP_00131954	Conserved hypothetical protein	ABI25652
ZP_00131953	Conserved hypothetical protein	ABI25653
ZP_00131952	Hypothetical protein	ABI25654
ZP_00131951	Possible transcription regulator	ABI25655

Table 2.11 continued from previous page

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00131950	Hypothetical protein	ABI25656
ZP_00131948	Conserved hypothetical protein, bacteriophage-related	No homolog
ZP_00133493	Hypothetical protein	ABI25659
ZP_00348111	Conserved hypothetical protein, bacteriophage-related	ABI25660
ZP_00348112	COG0840: Methyl-accepting chemotaxis protein, bacteriophage-related	ABI25661
ZP_00348113	Hypothetical protein	No homolog
ZP_00348114	COG0840: Methyl-accepting chemotaxis protein	ABI25662
ZP_00348115	COG3772: Phage-related lysozyme	ABI25663
ZP_00348116	Hypothetical protein	ABI25664
ZP_00133500	COG3692: Uncharacterized protein conserved in bacteria	No homolog
None	Hypothetical protein	ABI25668
ZP_00133501	NinG recombination protein	ABI25669
ZP_00133503	COG1598: Uncharacterized conserved protein	ABI25671
ZP_00348117	Hypothetical protein	No homolog
ZP_00133504	COG3617: Prophage antirepressor	ABI25673
ZP_00133505	Hypothetical protein	ABI25674
ZP_00133506	Possible DNA methylase	ABI25675
ZP_00348118	Conserved hypothetical protein, bacteriophage-related	ABI25677
ZP_00348119	Conserved hypothetical protein	ABI25678
ZP_00133510	Hypothetical protein	No homolog
ZP_00133511	COG1974: SOS-response transcriptional repressors	No homolog

Table 2.11 continued from previous page

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00133512	Hypothetical protein	No homolog
ZP_00133513	COG3636: Predicted transcriptional regulator	No homolog
ZP_00133514	COG3657: Uncharacterized protein conserved in bacteria	No homolog
None	Conserved hypothetical protein, bacteriophage-related	No homolog
ZP_00133425	COG0517: FOG: CBS domain	No homolog
ZP_00133426	Conserved hypothetical protein, bacteriophage-related	No homolog
ZP_00133427	COG2805: Tfp pilus assembly protein	No homolog
ZP_00133428	Conserved hypothetical protein, bacteriophage-related	No homolog
ZP_00133429	Hypothetical protein	No homolog
ZP_00133430	Hypothetical protein	No homolog
ZP_00133431	Conserved hypothetical protein	No homolog
ZP_00133432	Hypothetical protein	No homolog
ZP_00133433	Conserved hypothetical protein, bacteriophage-related	No homolog
ZP_00204658	Hypothetical protein	ABI25691
ZP_00204659	Hypothetical protein	ABI25692
ZP_00133436	Hypothetical protein	ABI25693
ZP_00133437	COG3298: Predicted 3'-5' exonuclease	No homolog
ZP_00133439	Hypothetical protein	No homolog
ZP_00204660	Hypothetical protein	No homolog
ZP_00133440	COG0270: Site-specific DNA methylase	ABI25697
ZP_00133441	COG0582: Phage integrase	No homolog

Table 2.12: List of genes in prophage region II of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00132844	COG4166: ABC-type oligopeptide transport system, periplasmic component	No homolog
ZP_00132843	COG4166: ABC-type oligopeptide transport system, periplasmic component	No homolog
ZP_00132842	COG0601: ABC-type dipeptide/oligopeptide/nickel transport system, permease component	No homolog
ZP_00132841	COG1173: ABC-type dipeptide/oligopeptide/nickel transport systems, permease component	No homolog
ZP_00132840	COG4172: ABC-type uncharacterized transport system, duplicated ATPase component	No homolog
ZP_00132839	COG4172: ABC-type uncharacterized transport system, duplicated ATPase component	No homolog
ZP_00132838	Hypothetical protein	No homolog
ZP_00348150	Hypothetical protein	No homolog
ZP_00132837	Hypothetical protein	No homolog
ZP_00132836	COG0270: Site-specific DNA methylase, bacteriophage-related	No homolog
ZP_00132835	Hypothetical protein	No homolog
ZP_00132834	Hypothetical protein	No homolog
ZP_00132833	COG5301: Phage-related tail fiber protein	No homolog
ZP_00132832	COG3778: Uncharacterized protein conserved in bacteria	No homolog
ZP_00132831	COG3299: Uncharacterized homolog of phage Mu protein gp47	No homolog
ZP_00132830	COG4381: Mu-like prophage protein gp46	No homolog

Table 2.12 continued from previous page

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00132829	COG4384: Mu-like prophage protein gp45	No homolog
ZP_00132828	COG4379: Mu-like prophage tail protein gpP	No homolog
ZP_00132827	COG4228: Mu-like prophage DNA circulation protein	No homolog
ZP_00132826	COG5283: Phage-related tail protein	No homolog
ZP_00132825	Hypothetical protein	No homolog
ZP_00132824	Hypothetical protein	No homolog
ZP_00132823	COG4386: Mu-like prophage tail sheath protein gpL	No homolog
ZP_00132822	Hypothetical protein	No homolog
ZP_00132821	COG0443: Molecular chaperone	No homolog
ZP_00132820	COG4387: Mu-like prophage protein gp36	No homolog
ZP_00132819	Hypothetical protein	No homolog
ZP_00132818	COG4397: Mu-like prophage major head subunit gpT	No homolog
ZP_00132817	COG4388: Mu-like prophage I protein	No homolog
ZP_00132816	COG5005: Mu-like prophage protein gpG	No homolog
ZP_00348148	Hypothetical protein	No homolog
ZP_00348149	Hypothetical protein	No homolog
ZP_00132814	COG2369: Uncharacterized protein, homolog of phage Mu protein gp30	No homolog
ZP_00132813	COG4383: Mu-like prophage protein gp29	No homolog
ZP_00132812	Conserved hypothetical protein, phage-related	No homolog
ZP_00132811	Conserved hypothetical protein, phage-related	No homolog

Table 2.12 continued from previous page

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00132810	Hypothetical protein	No homolog
ZP_00348147	COG0840: Methyl-accepting chemotaxis protein	No homolog
ZP_00348146	Hypothetical protein	No homolog
ZP_00132806	COG3023: Negative regulator of beta-lactamase expression	No homolog
ZP_00132805	Hypothetical protein	No homolog
ZP_00132804	COG0556: Helicase subunit of the DNA excision repair complex	No homolog
ZP_00132803	COG4382: Mu-like prophage protein gp16	No homolog
ZP_00348145	Hypothetical protein	No homolog
ZP_00132802	Hypothetical protein	No homolog
ZP_00348144	Hypothetical protein	No homolog
ZP_00348143	Hypothetical protein	No homolog
ZP_00132800	Hypothetical protein	No homolog
ZP_00132799	Hypothetical protein	No homolog
ZP_00132797	Hypothetical protein	No homolog
ZP_00348141	Hypothetical protein	No homolog
ZP_00132265	Hypothetical protein	No homolog
ZP_00204743	COG2842: Uncharacterized ATPase, putative transposase	No homolog
ZP_00132267	COG2801: Transposase, phage-related	No homolog
ZP_00204744	Hypothetical protein	No homolog
ZP_00132268	Hypothetical protein	No homolog
ZP_00132269	COG2932: Predicted transcriptional regulator	No homolog
ZP_00132270	COG4172: ABC-type uncharacterized transport system, duplicated ATPase component	No homolog
ZP_00132271	Conserved hypothetical protein	No homolog

Table 2.13: List of genes in prophage region III of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00133259	COG3636: Predicted transcriptional regulator	No homolog
ZP_00133258	COG3657: Uncharacterized protein conserved in bacteria	No homolog
ZP_00133257	COG1396: Predicted transcriptional regulator	No homolog
ZP_00204671	COG0178: Exonuclease ATPase subunit	No homolog
ZP_00204670	Hypothetical protein	No homolog
None	Conserved hypothetical protein	ABI24802
ZP_00133256	COG0863: DNA modification methylase	No homolog
ZP_00133255	Hypothetical protein	ABI24804
ZP_00348128	Conserved hypothetical protein	ABI24805
ZP_00348127	Hypothetical protein	No homolog
ZP_00133253	COG5519: Superfamily II helicase	No homolog
ZP_00133252	Hypothetical protein	No homolog
ZP_00133251	Type I restriction modification system, M protein	No homolog
ZP_00348126	Type I restriction modification system, S protein	No homolog
ZP_00133250	Hypothetical protein, bacteriophage-related	No homolog
ZP_00133249	COG5518: Bacteriophage capsid portal protein	No homolog

Table 2.13 continued from previous page

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00133248	COG1196: Chromosome segregation ATPase	No homolog
ZP_00133247	COG0468: RecA/RadA recombinase	No homolog
ZP_00133246	Conserved hypothetical protein	No homolog
ZP_00133245	Bacteriophage capsid protein	No homolog
ZP_00133244	Bacteriophage-related protein	No homolog
ZP_00133243	Bacteriophage glycoprotein	No homolog
ZP_00348125	COG0443: Molecular chaperone	No homolog
ZP_00133242	COG1943: Transposase	ABI24943
ZP_00133241	COG0675: Transposase	ABI24763
ZP_00133240	COG0582: Phage integrase	ABI25612

Table 2.14: List of genes in prophage region IV of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00133360	COG0457: FOG: TPR repeat, bacteriophage-related protein	No homolog
ZP_00204666	COG1404: Subtilisin-like serine proteases	No homolog
ZP_00133357	COG0464: ATPases of the AAA+ class	No homolog
ZP_00133356	Transposase	No homolog
ZP_00133355	Hypothetical protein	No homolog
ZP_00133354	Transposase	No homolog
ZP_00133353	Transposase	No homolog
ZP_00133352	Transposase	ABI25391
ZP_00133351	Hypothetical protein	No homolog
ZP_00133350	Hypothetical protein	No homolog
ZP_00204665	Type I restriction modification system, M protein	No homolog
ZP_00133349	Hypothetical protein	No homolog
ZP_00204664	Hypothetical protein	No homolog
ZP_00133348	Ferric reductase like transmembrane component	No homolog
ZP_00133347	Predicted transcriptional regulator	ABI24897
ZP_00204663	Hypothetical protein	No homolog
ZP_00133346	Hypothetical protein	No homolog
ZP_00133345	Transposase	No homolog
ZP_00133344	Hypothetical protein	No homolog
ZP_00133343	COG2842: Uncharacterized ATPase, putative transposase	No homolog
ZP_00133342	Transposase	No homolog
ZP_00133341	COG3423: Predicted transcriptional regulator	No homolog
ZP_00133340	COG2932: Predicted transcriptional regulator	No homolog

Table 2.15: List of genes in prophage region V of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00132749	Hypothetical protein	No homolog
ZP_00132750	Type I restriction modification system, M protein	No homolog
ZP_00132751	Transposase	No homolog
ZP_00132752	Phage integrase	ABI24282
ZP_00132753	Phage-related ATPase involved in DNA repair	No homolog
None	Phage integrase	ABI24796
ZP_00132754	Phage shock protein A	No homolog
ZP_00132755	Predicted transcriptional regulator	ABI24799
None	Conserved hypothetical protein	ABI24800
None	Hypothetical protein	ABI24801
ZP_00132757	Transposase	ABI25666
ZP_00132553 ZP_00133241 ZP_00132550	Transposase	ABI24860 ABI24763 ABI25665

Table 2.16: List of genes in prophage region VI of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00132894	Hypothetical protein	No homolog
None	Hypothetical protein	No homolog
ZP_00132893	Hypothetical protein	No homolog
ZP_00204686	Hypothetical protein	No homolog
ZP_00132892	COG1943: Transposase	ABI25666
ZP_00133241	COG0675: Transposase	ABI25665
ZP_00132551	Hypothetical protein	No homolog
ZP_00204721	Hypothetical protein	ABI24805
ZP_00204722	COG1196: Chromosome segregation ATPase	ABI24804
ZP_00132552	Hypothetical protein	ABI24802
ZP_00132553	COG0675: Transposase	ABI24860
ZP_00132554	COG1943: Transposase	ABI25666
ZP_00204723	COG1835: Predicted acyltransferases	ABI24830
ZP_00132557	COG0582: Phage integrase	ABI24832
ZP_00132558	Type I restriction modification system, S protein	ABI24833
ZP_00132559	Type I restriction modification system, M protein	No homolog
ZP_00132560	COG2184: Protein involved in cell division	No homolog
ZP_00132561	Type I restriction modification system, S protein	ABI24833
ZP_00204724	COG4823: Abortive infection bacteriophage resistance protein	No homolog

Table 2.17: List of genes in prophage region VII of *H. somni* strain 2336

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00132143	COG5484: Uncharacterized conserved protein, bacteriophage-related	No homolog
ZP_00132144	Phage capsid scaffolding protein	No homolog
ZP_00132145	Phage major capsid protein	No homolog
ZP_00132146	Phage small terminase subunit	No homolog
ZP_00132147	COG0457: FOG: TPR repeat, Phage glycoprotein	No homolog
ZP_00132148	COG5004: P2-like prophage tail protein	No homolog
ZP_00132149	Bacteriophage holin protein	No homolog
ZP_00348196	COG3772: Phage-related lysozyme	ABI25663
ZP_00348197	COG3096: Uncharacterized protein involved in chromosome partitioning	No homolog
None	Bacteriophage-related protein	No homolog
ZP_00132152	P2 phage tail completion protein	No homolog
ZP_00132153	Hypothetical protein, bacteriophage-related	No homolog
ZP_00348198	Hypothetical protein	No homolog
ZP_00132154	COG5283: Phage-related tail protein	No homolog
ZP_00132155	Hypothetical protein	No homolog
ZP_00348199	Plasmid stabilisation system protein	No homolog
ZP_00132156	COG4540: Phage P2 baseplate assembly protein gpV	No homolog
ZP_00132157	COG3628: Phage baseplate assembly protein W	No homolog
ZP_00132158	COG3948: Phage-related baseplate assembly protein	No homolog
ZP_00132159	COG4385: Bacteriophage P2-related tail formation protein	No homolog
ZP_00132160	COG5301: Phage-related tail fiber protein	ABI25647

Table 2.17 continued from previous page

H. somni 2336 NCBI locus tag	Annotation	NCBI locus tag of closest *H. somni* 129Pt homolog
ZP_00348201	COG1396: Predicted transcriptional regulator	ABI24812
ZP_00348202	COG2026: Cytotoxic translational repressor of toxin-antitoxin stability system	ABI24811
ZP_00131960	Putative phage tail fiber assembly protein	No homolog
ZP_00132835	Conserved hypothetical protein	No homolog
ZP_00133561	Hypothetical protein	No homolog
ZP_00204632	Uncharacterized conserved protein	No homolog
ZP_00204633	COG2161: Antitoxin of toxin-antitoxin stability system	No homolog
ZP_00133564	COG3497: Phage tail sheath protein	No homolog
ZP_00133565	COG3498: Phage tail tube protein	No homolog
ZP_00133566	Phage-related protein	No homolog
ZP_00204634	COG3499: Phage protein U	No homolog
ZP_00204635	COG3500: Phage protein D	No homolog
ZP_00204636	COG2932: Predicted transcriptional regulator	No homolog
ZP_00204637	COG4197: Uncharacterized protein conserved in bacteria, prophage-related	No homolog
ZP_00133569	Hypothetical protein	No homolog
ZP_00133570	COG1196: Chromosome segregation ATPase	No homolog
ZP_00133571	Hypothetical protein	No homolog
ZP_00133572	Hypothetical protein	No homolog
ZP_00133573	Bacteriophage replication protein	No homolog
ZP_00204638	ABC-type dipeptide transport system, periplasmic component	ABI25494
ZP_00204639	COG0582: Bacteriophage Integrase	ABI25495

Acquisition and loss of genes encoding hemagglutinins by *H. somni* strains 2336 and 129Pt

Several types of two-partner virulence systems have been identified and characterized in Gram-negative bacteria (Jacob-Dubuisson et al., 2001). Filamentous hemagglutinins (Fha), consisting of a membrane-anchored protein (FhaC), which is involved in the activation/secretion of the cognate hemagglutinin/adhesin (FhaB), are prototypes of the two-partner virulence systems. Homologs of FhaB and FhaC that play a role in pathogenesis have been found in members of the *Pasteurellaceae* (Fuller et al., 2000; May et al., 2001; Ward et al., 2004).

The chromosome of *H. somni* strain 2336 contains four loci that have 11 putative genes encoding proteins that are homologous to FhaB, and 3 putative genes encoding proteins that are homologous to FhaC. Locus I of strain 2336 is 7,285 bp (GC content of 26.6%) and contains genes encoding FhaB and FhaC that appear to be associated with a transposon (Fig. 2.09). Strain 129Pt has a locus that contains genes that flank locus I of strain 2336, but does not contain the FhaB and FhaC homologs (Fig. 2.09).

Locus II of strain 2336 is 18,106 bp (GC content of 40%) and contains genes encoding four FhaB homologs and one FhaC homolog. No transposon or phage regions are apparent in this locus (Fig. 2.10). Strain 129Pt has a locus that contains genes that flank locus II of strain 2336, but does not contain the FhaB and FhaC homologs (Fig. 2.10).

Locus III of strain 2336 is 22,350 bp (GC content of 37.5%) and contains genes encoding five FhaB homologs. No FhaC homologs are found in this locus (Fig. 2.11). This locus is present in between prophage regions V and VI (Tables 2.03, 2.15, and 2.16). Partial homologs of *fhaB* genes from locus III of strain 2336 are found in strain 129Pt downstream of prophage region III (Tables 2.02 and 2.06; Fig. 2.11). This appears to be the only region in the chromosome of strain 129Pt that has a sequence related to the *fhaB* genes.

Locus IV of strain 2336 is 14,670 bp (GC content of 37.2%) and contains genes encoding FhaB and FhaC (Fig. 2.12). The *fhaB* of this locus is 12,288 bp and is the largest gene in the genome of strain 2336. This gene encodes a protein that is homologous to the high molecular weight immunoglobulin-binding protein of *H. somni* described by Tagawa et al. (2005), and the large supernatant protein 2 of *H. ducreyi* described by Ward et al. (1998). No transposon or phage regions are apparent in this locus. Strain 129Pt has a locus that contains genes that flank locus IV of strain 2336, but does not contain the FhaB and FhaC homologs (Fig. 2.12).

The eleven putative FhaB homologs found in the four loci of strain 2336 vary in size (the smallest has 444 amino acids and the largest has 4,095 amino acids). However, the eleven homologs appear to contain a core region that is conserved (Fig. 2.13), indicating a possible evolutionary and functional relationship. The hypothetical branching order and the inferred evolutionary change (or rate of evolution) of the eleven FhaB homologs is depicted in the phylogram in fig. 2.14. The phylogram indicates that the FhaB homologs within a locus are more similar to each other than to FhaB homologs from a different locus (Fig. 2.14).

The FhaC homolog found in locus IV is shorter than the FhaC homologs found in loci I and III. However, these three proteins appear to be related and contain several identical amino acids and conserved substitutions (Fig. 2.15).

H. somni strain 129Pt contains 12 genes randomly distributed throughout the chromosome (NCBI locus tags HS_0209, HS_0383, HS_0478, HS_0589, HS_0602, HS_0790, HS_1058, HS_1085, HS_1154, HS_1234, HS_1616, and HS_1632) that encode putative large adhesins (Challacombe et al., 2007). Homologs of these adhesins also occur in strain 2336 (data not shown). However, the adhesins have very low or no homology to the FhaB homologs of strain 2336 (data not shown).

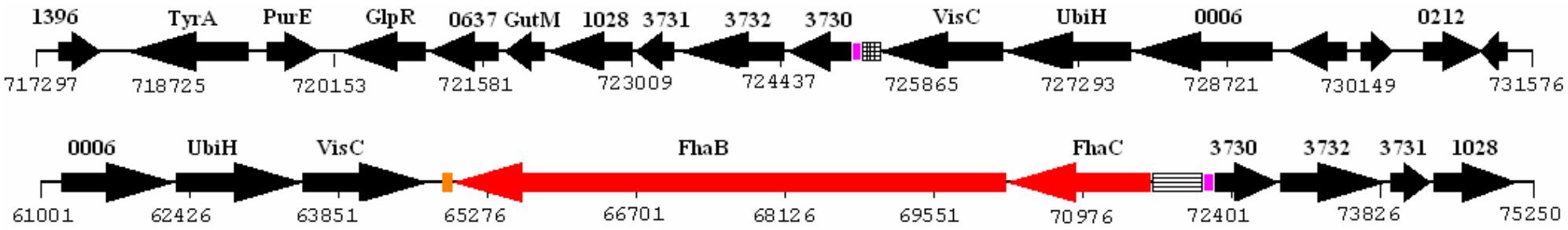


Fig. 2.09: Map of *H. somni* strain 2336 chromosomal locus I containing FhaB and FhaC compared with strain 129Pt

Top; *H. somni* strain 129Pt, 1028; Sorbitol-6-phosphate 2-dehydrogenase, 3731; Phosphotransferase system sorbitol-specific component IIA, 3732; PTS system, glucitol/sorbitol-specific IIBC component, 3730; Phosphotransferase system sorbitol-specific component IIC, Pink box; 96 bp sequence found only in *H. somni* strains 129Pt and 2336, Checkered box; 173 bp sequence found only in *H. somni* strain 129Pt, VisC; monooxygenase family protein, UbiH; 2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductase, 0006; Xaa-Pro aminopeptidase.

Bottom; *H. somni* strain 2336, 0006; Xaa-Pro aminopeptidase, UbiH; 2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductase, VisC; monooxygenase family protein, Orange box; 78 bp transposon-related sequence, FhaB; Large exoprotein involved in heme utilization or adhesion (contains 1755 amino acids), FhaC; Hemolysin activation/secretion protein (contains 450 amino acids), White box with horizontal lines; 516 bp sequence found only in *H. somni* strain 2336, Pink box; 96 bp sequence found only in *H. somni* strains 2336 and 129Pt, 3730; Phosphotransferase system (PTS) sorbitol-specific component IIC, 3732; PTS sorbitol-specific component IIB, 3731; PTS sorbitol-specific component IIA, 1028; Sorbitol-6-phosphate 2-dehydrogenase.

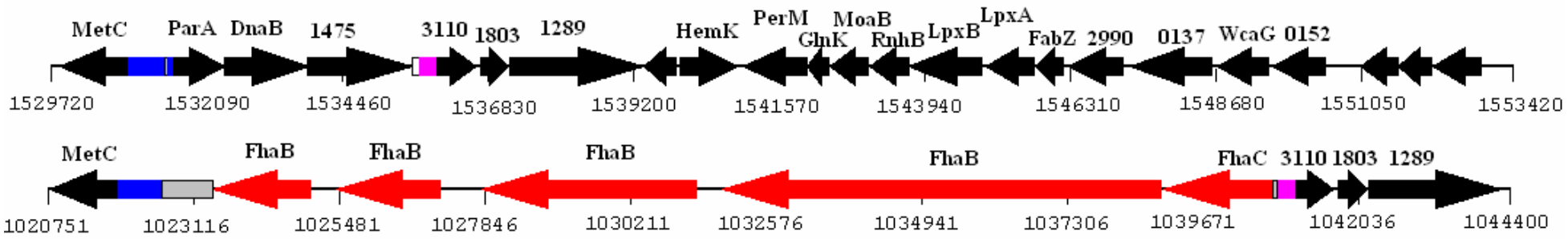


Fig. 2.10: Map of *H. somni* strain 2336 chromosomal locus II containing FhaB and FhaC compared with strain 129Pt

Top; *H. somni* strain 129Pt, MetC; cystathionine gamma-synthase, Blue box; 597 bp sequence found only in *H. somni* 2336 and 129Pt, ParA; chromosome partitioning-related protein, DnaB; replicative DNA helicase, 1475: Predicted transcriptional regulators, White box; 100 bp sequence found only in *H. somni* strain 129Pt, Pink box; 282 bp sequence found only in *H. somni* 2336 and 129Pt, 3110; Uncharacterized protein conserved in bacteria, 1803; Methylglyoxal synthase, 1289; Predicted membrane protein.

Bottom; *H. somni* strain 2336, MetC; cystathionine gamma-synthase, Blue box; 597 bp sequence found only in *H. somni* 2336 and 129Pt, Gray box; 827 bp sequence found only in *H. somni* 2336, FhaB; Large exoprotein involved in heme utilization or adhesion (contains 522 amino acids), FhaB; Large exoprotein involved in heme utilization or adhesion (contains 561 amino acids), FhaB; Large exoprotein involved in heme utilization or adhesion (contains 1159 amino acids), FhaB; Large exoprotein involved in heme utilization or adhesion (contains 2372 amino acids), FhaC; Hemolysin activation/secretion protein (contains 578 amino acids), Gray box; 57 bp sequence found only in *H. somni* 2336, Pink box; 282 bp sequence found only in *H. somni* 2336 and 129Pt, 3110; Uncharacterized protein conserved in bacteria, 1803; Methylglyoxal synthase, 1289; Predicted membrane protein.

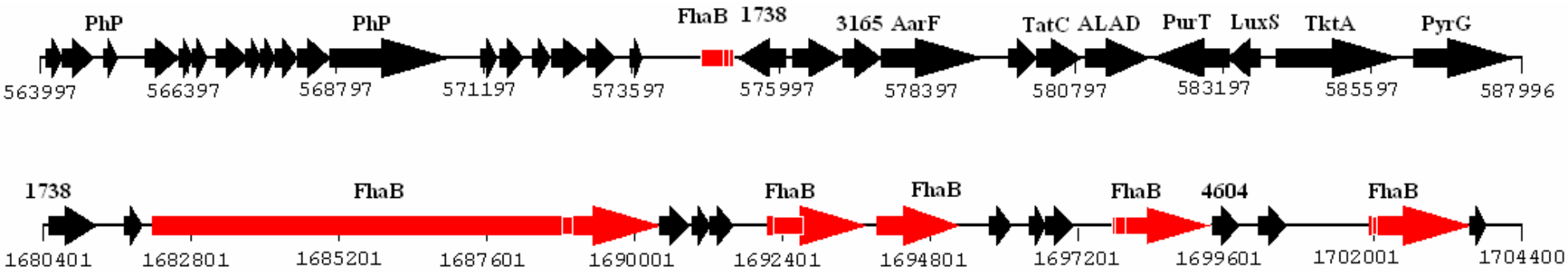


Fig. 2.11: Map of *H. somni* strain 2336 chromosomal locus III containing FhaB compared with strain 129Pt

Top; *H. somni* strain 129Pt, PhP; bacteriophage-related proteins, FhaB; sequence showing homology to sequences marked within FhaB homologs in *H. somni* strain 2336, 1738; Uncharacterized conserved protein.

Bottom; *H. somni* strain 2336, 1738; Uncharacterized conserved protein, FhaB; Large exoprotein involved in heme utilization or adhesion (contains 2744 amino acids), FhaB; Large exoprotein involved in heme utilization or adhesion (contains 531 amino acids), FhaB; Large exoprotein involved in heme utilization or adhesion (contains 444 amino acids), FhaB; Large exoprotein involved in heme utilization or adhesion (contains 542 amino acids), 4604; ABC-type enterochelin transport system, ATPase component, FhaB; Large exoprotein involved in heme utilization or adhesion (contains 544 amino acids). White bordered boxes within the four FhaB homologs show regions that have homology to the red box in strain 129Pt). In both maps, unmarked arrows indicate hypothetical or conserved hypothetical proteins.

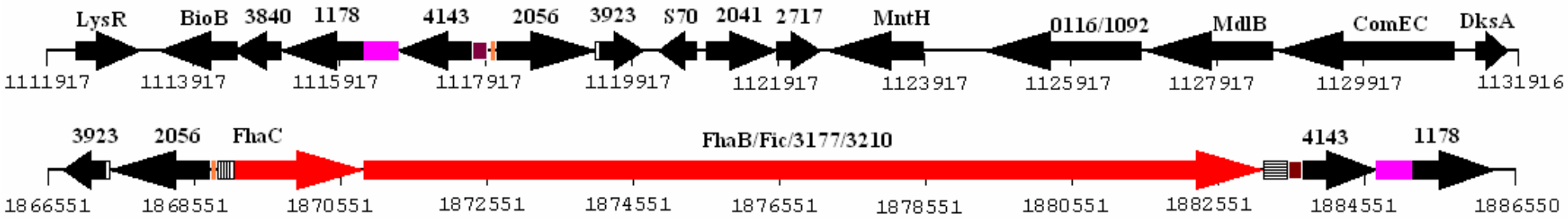


Fig. 2.12: Map of *H. somni* strain 2336 chromosomal locus IV containing FhaB and FhaC compared with strain 129Pt

Top; *H. somni* strain 129Pt, 1178; ABC-type Fe3+ transport system, permease component, Pink box; 475 bp sequence found only in *H. somni* strains 129Pt and 2336, 4143; ABC-type thiamine transport system, periplasmic component, Brown box; 198 bp sequence conserved in *H. somni* strains 129Pt and 2336, Orange box; 45 bp sequence found only in *H. somni* strains 129Pt and 2336, 2056; Predicted permease, 3923; Primosomal replication protein N.

Bottom; *H. somni* strain 2336, 3923; Primosomal replication protein N, 2056; Predicted permease, Orange box; 45 bp sequence found only in *H. somni* strains 129Pt and 2336, White box with vertical lines; 256 bp sequence found only in *H. somni* strain 2336, FhaC; Hemolysin activation/secretion protein (contains 586 amino acids), FhaB/Fic/3177/3210; Large exoprotein involved in heme utilization or adhesion (high molecular weight-immunoglobulin binding protein, contains 4095 amino acids), White box with horizontal lines; 341 bp sequence found only in *H. somni* strain 2336, Brown box; 198 bp sequence conserved in *H. somni* strains 129Pt and 2336, 4143; ABC-type thiamine transport system, periplasmic component, Pink box; 475 bp sequence found only in *H. somni* strains 129Pt and 2336, 1178; ABC-type Fe3+ transport system, permease component.

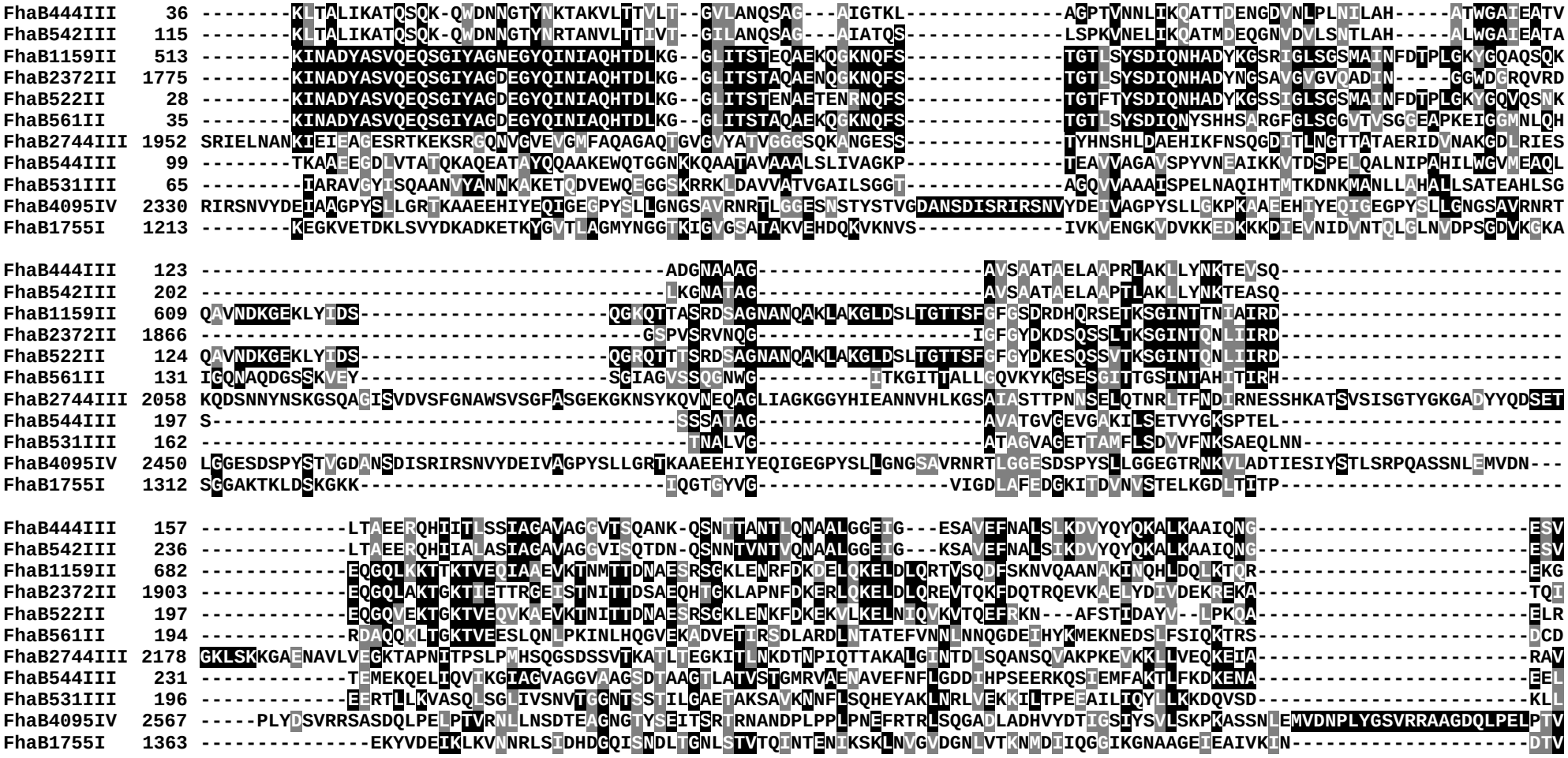


Fig. 2.13: ClustalW-BOXSHADE multiple sequence alignment of *H. somni* strain 2336 FhaB homologs

(Figure continued in next page)

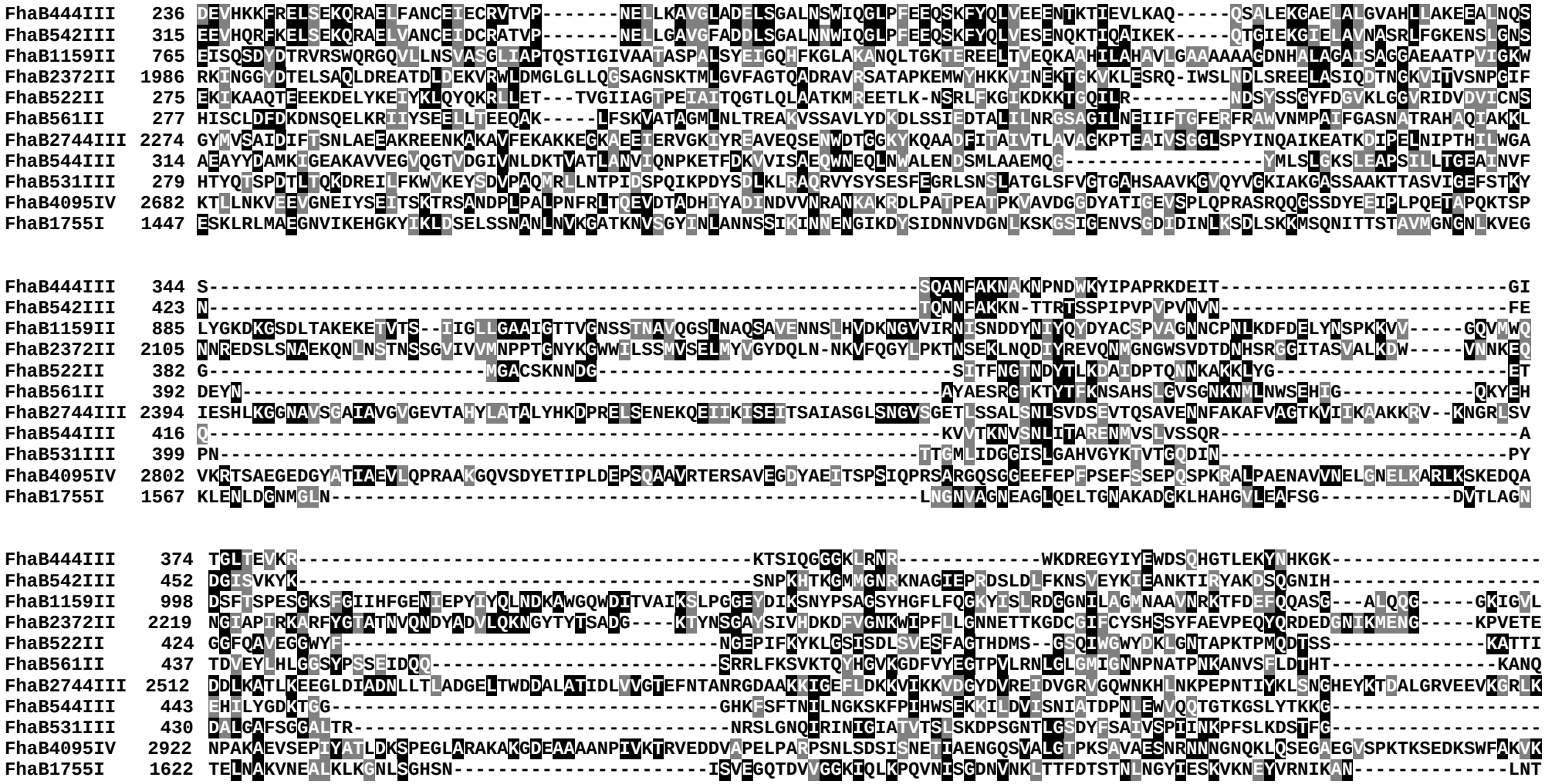


Fig. 2.13: ClustalW-BOXSHADE multiple sequence alignment of *H. somni* strain 2336 FhaB homologs

(detailed legend appears in next page)

Fig. 2.13: ClustalW-BOXSHADE multiple sequence alignment of *H. somni* strain 2336 FhaB homologs

FhaB444III; FhaB homolog from locus III containing 444 amino acids, FhaB542III; FhaB homolog from locus III containing 542 amino acids, FhaB1159II; FhaB homolog from locus II containing 1,159 amino acids, FhaB2372II; FhaB homolog from locus II containing 2,372 amino acids, FhaB522II; FhaB homolog from locus II containing 522 amino acids, FhaB561II; FhaB homolog from locus II containing 561 amino acids, FhaB2744III; FhaB homolog from locus III containing 2,744 amino acids, FhaB544III; FhaB homolog from locus III containing 544 amino acids, FhaB531III; FhaB homolog from locus III containing 531 amino acids, FhaB4095IV; FhaB homolog from locus IV containing 4,095 amino acids, FhaB1755I; FhaB homolog from locus I containing 1,755 amino acids. The number next to each protein sequence indicates amino acid position. The figure only shows the regions where all 11 proteins were found to be homologous. Homologous regions are box-shaded dark (identical amino acid residues) and gray (conserved amino acid substitutions).

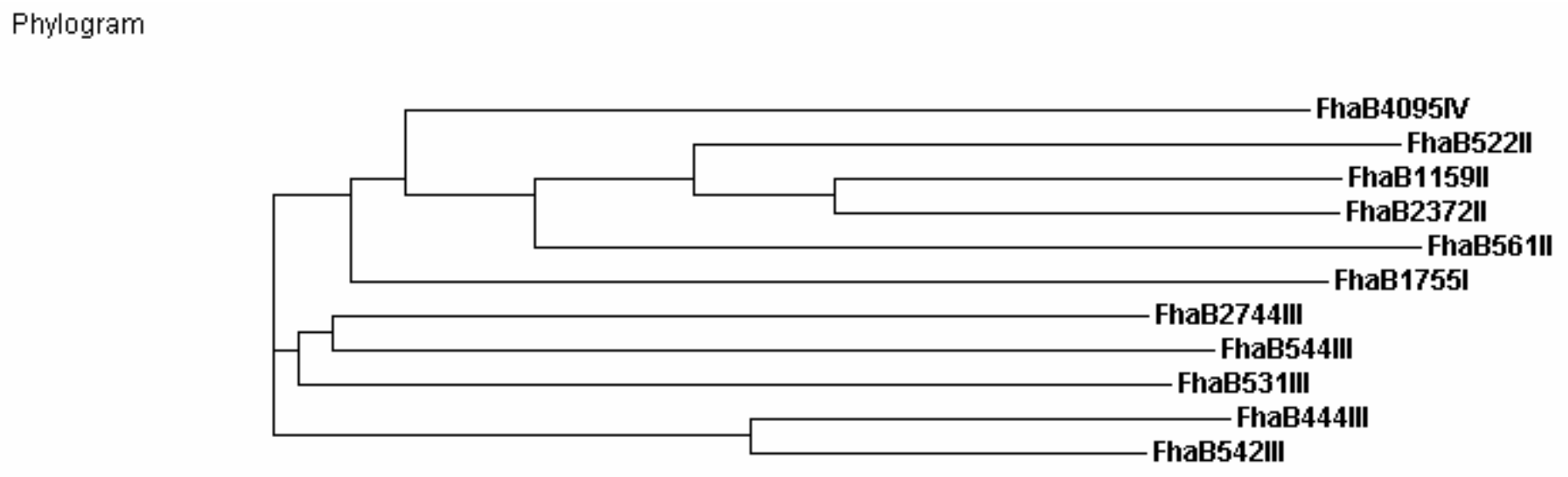


Fig. 2.14: Phylogram showing the relationship

between *H. somni* strain 2336 FhaB homologs aligned in fig. 2.13

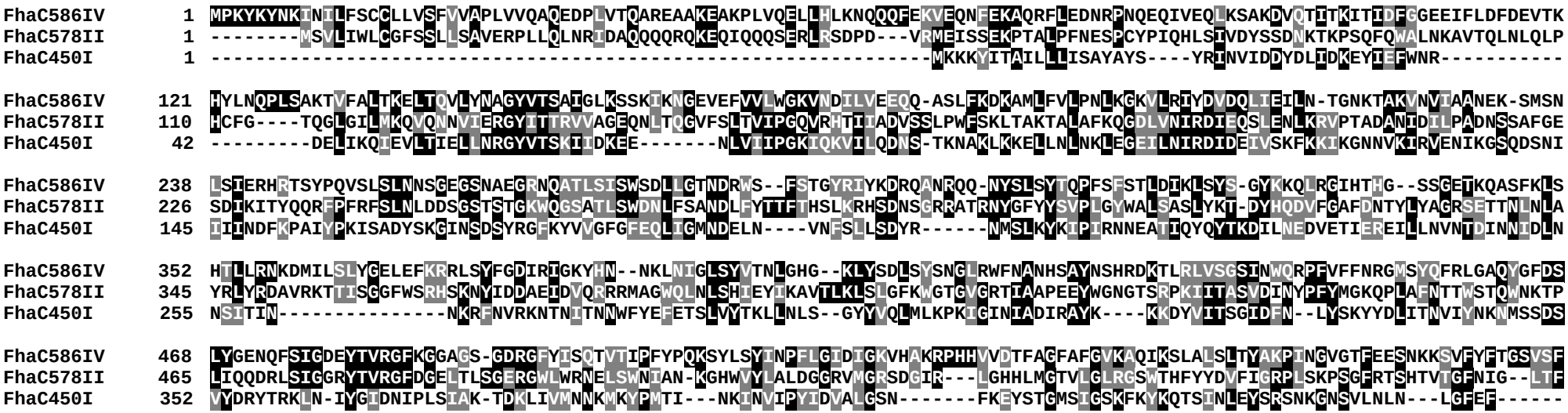


Fig. 2.15: ClustalW-BOXSHADE multiple sequence alignment of *H. somni* strain 2336 FhaC homologs

FhaC586IV; FhaC homolog from locus IV containing 586 amino acids, FhaC578II; FhaC homolog from locus II containing 578 amino acids, FhaC450I; FhaC homolog from locus I containing 450 amino acids, The number next to each protein sequence indicates amino acid position. Homologous regions are box-shaded dark (identical amino acid residues) and gray (conserved amino acid substitutions).

Comparison of *H. somni* strains 2336 and 129Pt loci encoding transferrin-binding proteins

The ability to acquire and metabolize iron is an important determinant of bacterial survival and adaptability. Several members of the *Pasteurellaceae* possess special receptors consisting of two unrelated transferrin-binding proteins, TbpA and TbpB, which facilitate acquisition of transferrin-bound iron from their hosts (Cornelissen, 2003). It has been shown that *H. somni* strain 649 has a TbpA-TbpB receptor system that acquires iron only from bovine transferrin, and a second, probably redundant, TbpA2 receptor system that can acquire iron from bovine, caprine, or ovine transferrins (Ekins et al., 2004). It is thought that transferrin-binding receptor systems may be responsible, in part, for the host specificity of *H. somni* strains (Yu et al., 1992; Ward et al., 2006).

A comparison of the chromosomes of *H. somni* strains 2336 and 129Pt revealed that both strains contain genes encoding TbpA and TbpB (Fig. 2.16). The two strains also contain a 200 bp transposon-related sequence in the region examined in fig. 2.16. Strain 129Pt contains a decaying prophage-related region immediately downstream of the genes encoding TbpA and TbpB (Fig. 2.16). Upstream of this decaying prophage-related region is a 9,817 bp sequence that has been designated as prophage region II of strain 129Pt (Table 2.02). This prophage region contains genes that encode a putative virulence-associated protein E (VirE) and a RM system that has homology to the proteins of the Type I RM system of *H. somni* strain 2336 (Table 2.05 and Fig. 2.16). With the exception of phage-related genes, the sequences up- and down-stream of genes encoding TbpA and TbpB in the two strains are similar and have a conserved gene order.

Furthermore, homologs of *H. somni* strain 649 TbpA2 are present in strain 129Pt (HS_0582, heme uptake protein, associated with a transposon) and strain 2336 (ZP_00348205 and ZP_00133056, COG1629: Outer membrane receptor proteins involved in Fe transport).

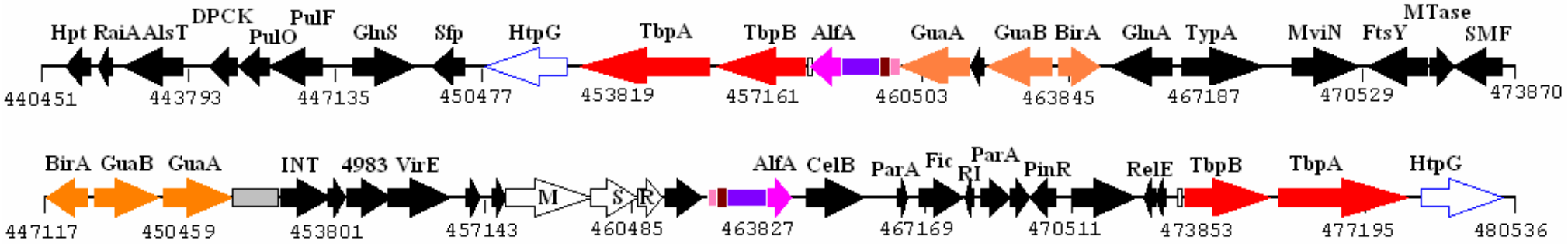


Fig. 2.16: Maps of *H. somni* strain 2336 and 129Pt chromosomal loci encoding transferrin binding proteins

Top; *H. somni* strain 2336, HtpG; Molecular chaperone, HSP90 family, TbpA; Transferrin-binding protein A, TbpB; Transferrin-binding protein B, White box; 91 bp sequence conserved in *H. somni* strains, AlfA; alpha-L-fucosidase, Purple box; 896 bp sequence conserved in *H. somni* strains 2336 and 129Pt, Brown box; 200 bp transposon-related sequence, Tan box; 246 bp sequence conserved in *H. somni* strains 2336 and 129Pt, GuaA; GMP synthase, PP-ATPase domain/subunit, GuaB; IMP dehydrogenase/GMP reductase, BirA; Biotin-(acetyl-CoA carboxylase) lyase (biotin operon repressor).

Bottom; *H. somni* strain 129Pt, BirA, GuaB, and GuaA; as in strain 2336, Gray box; 1,095 bp sequence found only in *H. somni* strain 129Pt, INT; phage integrase, 4983; possible phage DNA-polymerase or DNA-primase, VirE; possible virulence-associated protein E, M; restriction enzyme, site-specific DNA-methyltransferase, S; restriction enzyme, DNA specificity protein, R; restriction enzyme, REase subunit, Tan, Brown, and Purple boxes; as in strain 2336, AlfA; as in strain 2336, CelB; PTS, cellobiose-specific IIC component, ParA; conserved hypothetical protein, Fic; cell filamentation-like protein, RI; phage-related resolvase/integrase-like protein, ParA; ParA-like protein, PinR; phage-related resolvase/integrase-like protein, RelE; Plasmid addiction system poison protein, TbpB, TbpA, and HtpG; as in strain 2336. In the above maps, unmarked arrows indicate hypothetical or conserved hypothetical proteins.

Acquisition of a gene encoding a subtilisin-like serine protease by *H. somni* strain 2336

Subtilisin-like proteases (subtilases) are a large superfamily of functionally diverse serine endo- and exo-peptidases that occur in prokaryotes and eukaryotes. Bacterial subtilisins may have a role in pathogenesis besides facilitating protein degradation and nutrient acquisition (Siezen et al., 2007). However, subtilisin-like serine proteases from members of the *Pasteurellaceae* have not been characterized previously.

H. somni strain 2336 prophage region IV contains a gene that encodes a putative subtilisin-like serine protease (COG1404; Table 2.14 and Fig. 2.17). This gene is not found in strain 129Pt or other members of the *Pasteurellaceae* whose genome sequences are available. A blastp search revealed that *H. somni* strain 2336 putative subtilase is 74% identical to an unnamed protein of *E. coli* (NCBI locus tag AAA64865), 44% identical to a putative subtilisin of *Ralstonia eutropha* JMP134, 39% identical to a conserved hypothetical protein of *Mesorhizobium* spp. BNC1, 37% identical to a putative subtilisin of *Delftia acidovorans* SPH-1, 35% identical to a putative subtilisin of *Chromohalobacter salexigens* DSM 3043, and 22% identical to a putative subtilisin of *Neisseria lactamica*.

The prokaryotic subtilase database (www.cmbi.ru.nl/subtilases) contains two proteins of *H. ducreyi* strain 35000HP annotated as possible serine proteases (NCBI locus tags HD1094 and HD1278). However, these two proteins are unrelated to each other and to the *H. somni* strain 2336 putative subtilase. Furthermore, none of the other members of the *Pasteurellaceae* whose genome sequences are available have homologs of HD1094 or HD1278 (data not shown). Although *H. somni* strain 2336 appears to have acquired the subtilisin gene by bacteriophage-mediated HGT, the mechanism of acquisition of *H. ducreyi* subtilisins HD1094 or HD1278 remains uncharacterized.

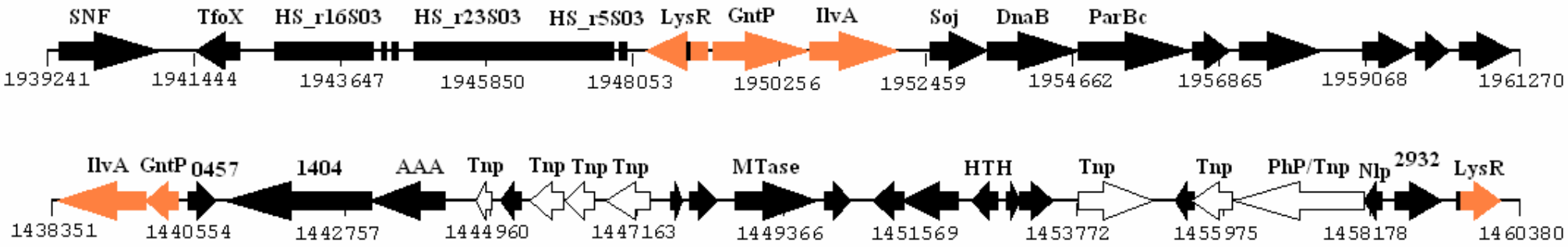


Fig. 2.17: Map of *H. somni* strain 2336 chromosomal region containing the subtilisin gene compared with *H. somni* strain 129Pt

Top; *H. somni* strain 129Pt, LysR; transcriptional regulator, LysR family, GntP; H+/gluconate symporter and related permease, IlvA; D-serine dehydratase.

Bottom; *H. somni* strain 2336, IlvA; D-serine dehydratase, GntP; H+/gluconate symporter and related permease, 0457; FOG: TPR repeat, 1404; Subtilisin-like serine protease, AAA; ATPase of the AAA+ class, Tnp; transposases for IS1668, MTase; Type I restriction-modification system methyltransferase subunit, HTH_XRE; Predicted transcriptional regulator, Tnp; transposases, PhP/Tnp; Uncharacterized ATPase (putative phage transposase), Nlp; Ner-like DNA-binding protein, 2932; Predicted transcriptional regulator, LysR; transcriptional regulator, LysR family.

In the above maps, unmarked arrows indicate hypothetical or conserved hypothetical proteins.

Comparison of other loci in *H. somni* strains 2336 and 129Pt containing prophage sequences

Prophage region I of strain 129Pt is the smallest (798 bp) and contains a single gene that encodes a phage integrase (Table 2.04). A homolog of this integrase is found in prophage region V of *H. somni* strain 2336 (Table 2. 15). Prophage region IV of strain 129Pt is 3,289 bp and appears to have caused the loss of the Type I RM locus in this strain (Table 2.07, detailed analysis appears in chapter V). Prophage region V of strain 129Pt is 1,142 bp and contains only two genes whose protein products have homologs in prophage region VII of strain 2336 (Tables 2.08 and 2.17). Prophage region VI of strain 129Pt is 4,079 bp and a similar region is not found in strain 2336 (Table 2.09 and Fig. 2.18). Prophage region VII of strain 129Pt is the largest (52,619 bp) and contains several genes that are related to genes in prophage region I of strain 2336, and the order of these related genes is conserved in the two strains (Tables 2.10 and 2.11). Although prophage region VII of strain 129Pt is unrelated to prophage region III of strain 2336, these two regions are flanked by similar genes (*wcaG* and *ffh*) (Figs. 2.19 and 2.20).

Prophage region II of strain 2336 is the second largest (45,957 bp) and does not have homologous regions in the genome of strain 129Pt (Table 2.12). However, prophage region III of strain 2336 contains three genes whose protein products have homologs in prophage region III of strain 129Pt (Tables 2.06 and 2.13). Prophage region V of strain 2336 contains four genes whose protein products have homologs in prophage region III of strain 129Pt (Tables 2.06 and 2.15). Prophage region VI of strain 2336 contains a Type I RM locus (Table 2.16, detailed analysis appears in chapter V). Prophage region VII of strain 2336 contains four genes whose protein products have homologs in prophage regions III and VII of strain 129Pt (Tables 2.06, 2.10, and 2.17). Furthermore, prophage regions I-VII of strain 129Pt and prophage regions I, III, V, VI, and VII of strain 2336 each contain at least one putative gene encoding a phage integrase.

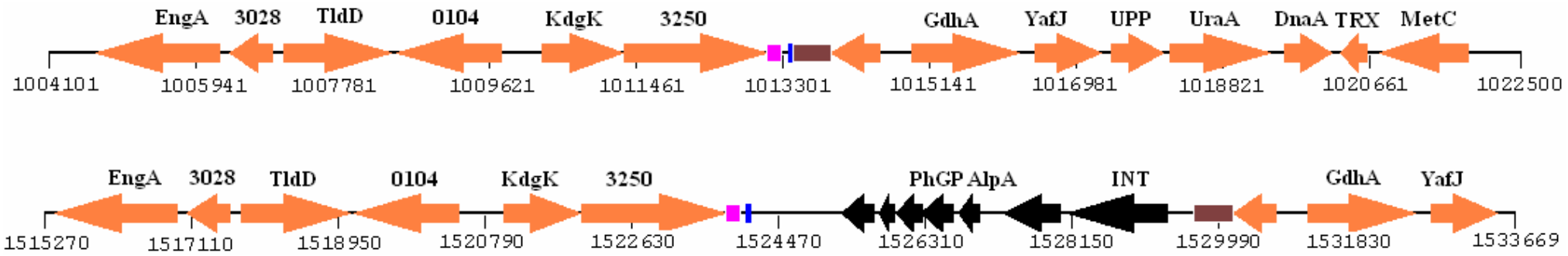


Fig. 2.18: Map of *H. somni* strain 129Pt chromosomal region containing phage sequences compared with *H. somni* strain 2336

Top; *H. somni* strain 2336, EngA; Predicted GTPase, 3028; Uncharacterized protein conserved in bacteria, TldD; Predicted Zn-dependent protease, 0104; Adenylosuccinate synthase, KdgK; Sugar kinase, ribokinase family, 3250; Beta-galactosidase/beta-glucuronidase, Pink box; 188 bp sequence found only in *H. somni* strains 2336 and 129Pt, Blue box; 57 bp sequence found only in *H. somni* strains 2336 and 129Pt, Brown box; 486 bp sequence conserved in *H. somni* strains 2336 and 129Pt, GdhA; Glutamate dehydrogenase/leucine dehydrogenase, YafJ; Predicted glutamine amidotransferase, Upp; Uracil phosphoribosyltransferase, Ura; Uracil permease, DnaA; possible DNA replication factor, TRX; Thiol-disulfide isomerase and thioredoxins, MetC; Cystathionine beta-lyases/cystathionine gamma-synthases.

Bottom; *H. somni* strain 129Pt, EngA, 3028, TldD, 0104, KdgK, 3250; as in strain 2336, Pink and Blue boxes; as in strain 2336, PhGP; bacteriophage-related protein (possible terminase small subunit), AlpA; possible prophage CP4-57 regulatory protein, INT; phage integrase, Brown box; as in strain 2336, GdhA and YafJ; as in strain 2336.

In the above maps, unmarked arrows indicate hypothetical or conserved hypothetical proteins.

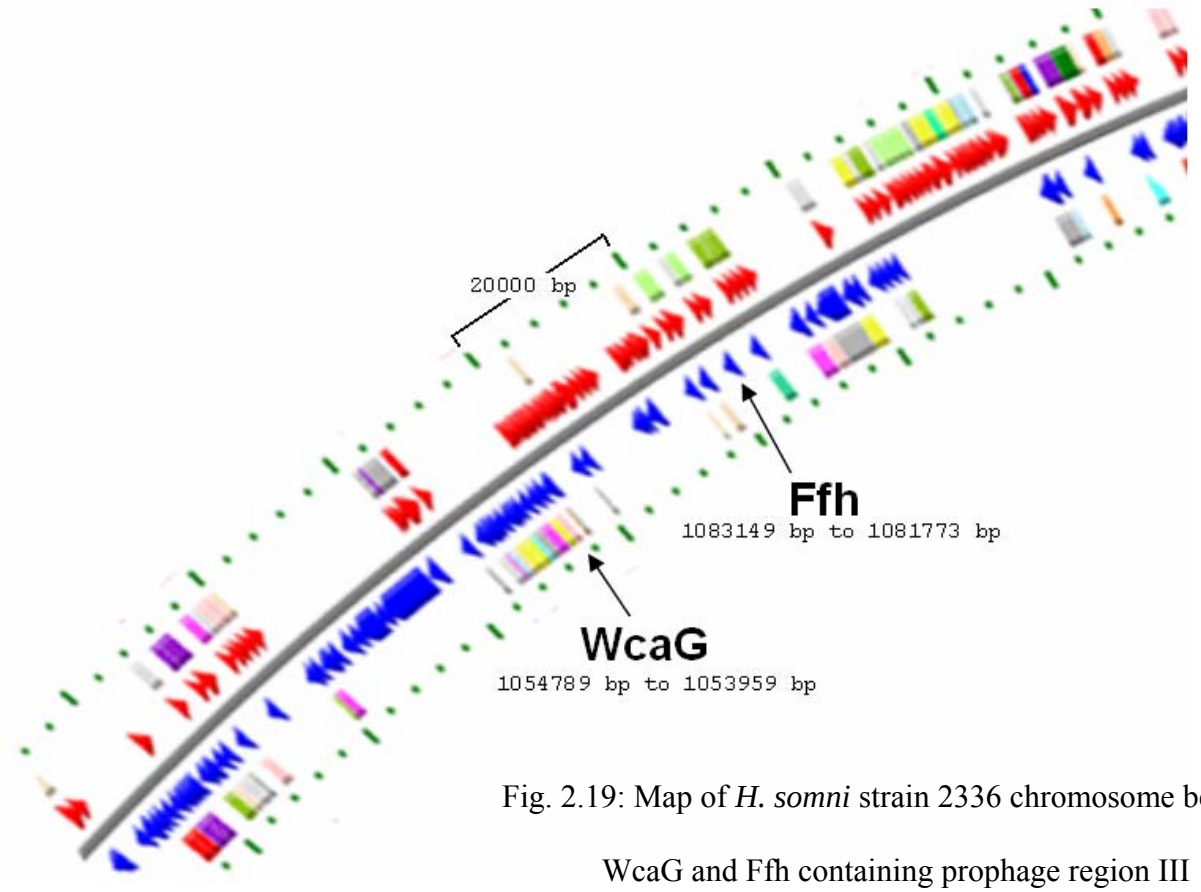


Fig. 2.19: Map of *H. somni* strain 2336 chromosome between WcaG and Ffh containing prophage region III

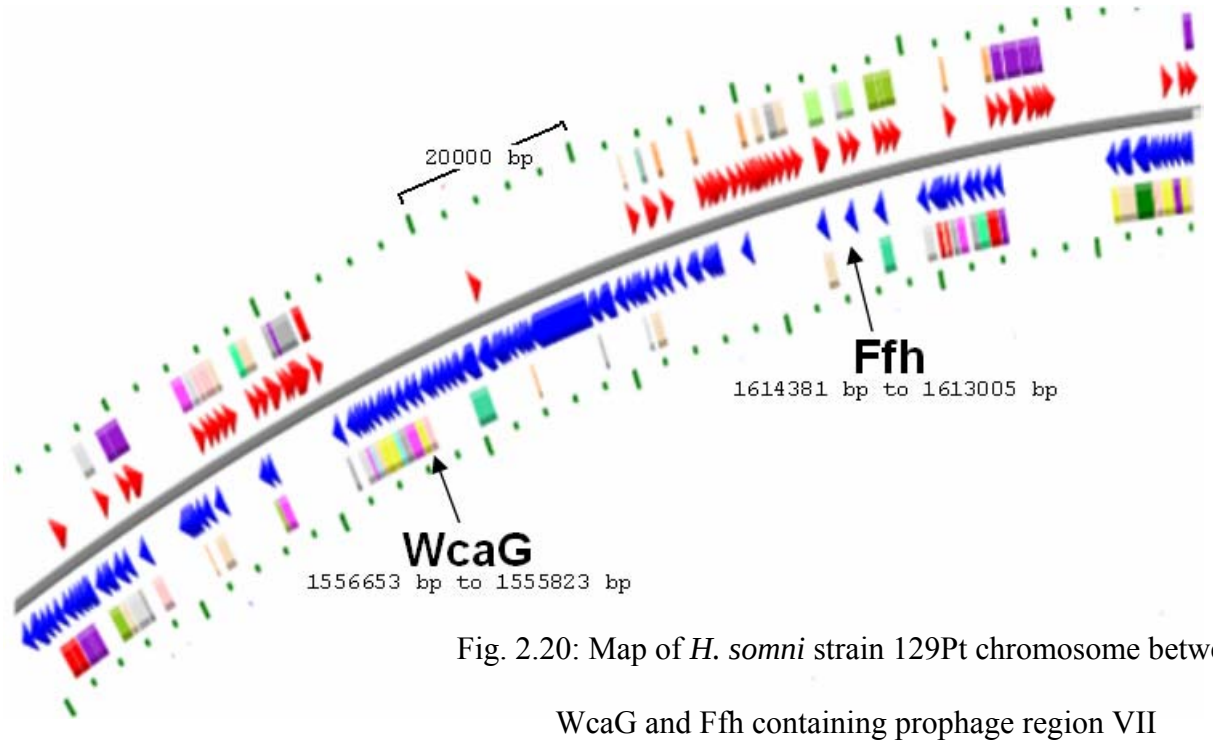


Fig. 2.20: Map of *H. somni* strain 129Pt chromosome between WcaG and Ffh containing prophage region VII

Discussion

As a significant step toward understanding the physiological properties of the bovine commensal and opportunistic pathogen *H. somni*, the complete genome sequences of a pneumonia isolate (strain 2336) and a preputial isolate (strain 129Pt) have been deciphered. The fact that the chromosomes of the two *H. somni* strains are slightly larger in size than those of *H. ducreyi* and *H. influenzae* is an indicator of their relative genetic richness. Despite the difference in chromosome size, the two *H. somni* strains have similar genome features, encode more ORFs on the plus strand than on the minus strand, and contain a vast common subset of highly homologous genes. The disparity in the number of ORFs predicted by BASys in relation to Challacombe et al. (2007) for strain 129Pt is probably due to different criteria used by the two annotation systems. Nevertheless, their common features point to the possibility that the two *H. somni* strains diverged from a common ancestor in the not too distant past.

Genetic events such as deletions, duplications, insertions, and inversions are relatively common in bacterial chromosomes as a result of bacteriophage infection, integration and excision of plasmids, transpositions, and/or replication-directed translocations (Tillier and Collins, 2000; Brussow et al., 2004). In addition, different prophages embedded within a single chromosome can contain similar genes encoding integration and structural functions, and it is not uncommon for these genes to undergo homologous recombination. One of the consequences of such homologous recombination is the rearrangement of the host chromosome (Canchaya et al., 2004). These events are known to be the precursors of evolution and can bring about a significant change in the number, linear order, as well as the orientation of genes on the circular chromosomes of different strains/species of closely related bacteria (Moret et al., 2002). In view of the biochemical similarities and differences that have been observed among strains of

H. somni and *H. influenzae*, a comprehensive analysis of their genetic characteristics was imminent. The availability of complete genome sequences of *H. somni* and *H. influenzae* strains facilitated the characterization of not only the colinearity and distribution of genes, but also the evolutionary changes within these bacterial species.

The lack of synteny of genes among *H. somni* strain 129Pt and *H. influenzae* strains indicates extensive rearrangements and genetic alterations of their chromosomes probably due to such generic events previously described by Tillier and Collins (2000) and Brussow et al. (2004). The non-colinearity of genes between the chromosomes of *H. somni* strains 2336 and 129Pt could also be attributed to similar events. The observation based on TaxPlot that strains 2336 and 129Pt have more proteins with greater similarity to proteins from the otitis media isolate of *H. influenzae* (strain 86-028NP) than to proteins from the nonpathogenic isolate of *H. influenzae* (strain Rd KW20) indicates the conservation of a subset of genes among strains 2336, 129Pt, and 86-028NP during their evolution.

The presence of multiple prophage regions in the chromosomes of strains 2336 and 129Pt is perhaps the most significant feature since the number and diversity of genes associated with these prophage sequences far exceed those described in *H. influenzae* strains Rd KW20 and 86-028NP (Fleischmann et al., 1995; Harrison et al., 2005). Of the ~250,000 bp sequence of strain 2336 that is not found in strain 129Pt, more than half (~150,000 bp) is located within these prophage regions. Similar observations have been made in other bacteria wherein prophage-associated sequences constitute a large portion of strain-specific DNA (Canchaya et al., 2004).

Of the seven prophage regions in the chromosome of strain 129Pt, regions I and V are the smallest and it is possible that these were part of one of the other large prophage regions and have been displaced. The only locus containing partial homologs of *fha* genes in strain 129Pt

appears to have lost the native full length genes because of a phage attack (phage region III). Similarly, the locus containing relics of Type I RM genes may have lost the resident genes due to a phage attack (phage region IV). Prophage regions II and VI of strain 129Pt contain several ORFs annotated as hypothetical and conserved hypothetical genes and some of these may encode metabolic and virulence functions.

It is interesting to note that several genes found in prophage region VII of strain 129Pt are also present in prophage region I of strain 2336 in a similar order. However, the two prophage sequences are found in different parts of the respective chromosomes. One possible explanation is that these regions represent a single phage insertion that occurred in the common ancestor of strains 2336 and 129Pt, and that this phage insertion has relocated within the chromosome of at least one of the two strains during their subsequent evolution. An alternate explanation is that these two phage insertions were similar, but independent, events, which occurred after strains 2336 and 129Pt diverged from their common ancestor. Yet another notable feature is the presence of similar genes (*wcaG* and *ffh*) flanking otherwise unrelated prophage regions (region VII of strain 129Pt and region III of strain 2336) present in different parts of the respective chromosomes. It is likely that these two phage insertions were independent events that targeted the locus between *wcaG* and *ffh* genes.

The seven prophage regions of strain 2336 are distributed randomly in the chromosome and they appear to have enriched the metabolic and pathogenic potential of this strain since they contain several strain-specific genes. Based on homology to similar genes in other bacteria, some of these genes have been predicted to encode a subtilase (region IV), a Type I RM system (region VI), and hemagglutinins (*fha* locus III, present immediately downstream of region VI). Many other ORFs have been annotated as hypothetical and conserved hypothetical genes.

Among the other 3 loci containing *fha* homologs in strain 2336, locus I appears to be an acquisition mediated by a transposon and locus II appears to be an acquisition due to homologous recombination. However, the mechanism of acquisition of locus IV containing the 12,288 bp *fhaB* homolog is unknown. Together, the 11 genes distributed over 4 loci in strain 2336 represent a large collection of *fhaB* homologs in a single genome. It is possible that *fhaB* homologs in locus II of strain 2336 are paralogs as are *fhaB* homologs in locus III. Furthermore, the presence of these multiple *fhaB* homologs in strain 2336 may partially be responsible for the serum resistance of this strain, in contrast to strain 129Pt that does not contain full length *fhaB* homologs and is serum sensitive.

The presence in strain 2336 of a gene encoding a putative subtilase whose homologs are not found in other members of the *Pasteurellaceae* is yet another classic case of bacteriophage-mediated HGT. Since subtilases have been shown to enhance pathogenesis in other bacteria, the virulence potential of *H. somni* subtilase homolog is worth investigating. Similarly, bacteriophage-mediated acquisition of genes encoding RM enzymes by strains 2336 and 129Pt is also an interesting feature and may provide yet another paradigm to test the "selfish gene hypothesis", "cellular defense hypothesis", and the "variation hypothesis" proposed for such genes by Kobayashi (2004). In addition, detailed phylogenetic analyses may reveal the evolutionary relationships between *H. somni* genes encoding RM enzymes, hemagglutinins, subtilases, and their respective orthologs in other bacteria.

Despite the genetic enrichment of strains 2336 and 129Pt by HGT, it is surprising that genes encoding antibioitic resistance have not been found in these strains and that antibiotic resistant strains of *H. somni* are rare, if any, in the field and in the literature. Furthermore, there appears to be no correlation between chromosome size and the number of tRNA genes (a

comparison of the genomes of *H. somni* 2336, *H. somni* 129Pt, *H. influenzae* 86-028NP, *H. ducreyi* 35000HP, and *P. multocida* Pm70 revealed that they contain 46, 50, 58, 48, and 57 tRNA genes, respectively). Whether the reduction in the number of tRNA genes in *H. somni* strains is due to disruptive integration of bacteriophages into tRNA genes [as in 'bacteriophage disruption of tRNA genes in *Lactobacillus johnsonii*' (Ventura et al., 2003)] or is a result of compensatory gene loss in lieu of the acquisition of new genes [as in 'genome reduction in pathogenic and symbiotic bacteria' (Hacker et al., 2003)] remains to be examined while taking into account the number of pseudogenes. Nevertheless, comparison of the chromosomes of strains 129Pt and 2336 bolsters the proposition that bacteriophages have played a major role in creating genomic diversity and phenotypic variability among the two strains. It is also apparent that strains 2336 and 129Pt have independently and intermittently acquired and lost genes, and that the net gain in strain 129Pt is less than the net gain in strain 2336 (mechanism III, page 44).

Previous studies have shown that the nucleotide composition and pattern of codon usage in regions that were acquired by HGT differ from nucleotide composition and pattern of codon usage in the rest of the chromosome (Lawrence and Ochman, 1997; Gladitz et al., 2005). Although comparison of gene order in the chromosomes of strains 2336 and 129Pt has facilitated the identification of chromosome rearrangements and regions acquired by HGT, compositional analyses of nucleotides (A, T, G, C, and G+C) and pattern of codon usage within these regions are pending.

A phylogenetic tree of various strains/species constructed on the basis of gene content within wholes genomes ("comprehensive genome phylogeny") rather than homology of single genes appears to be a meaningful approach to bacterial classification and understanding their evolution. This type of phylogenetic analyses takes into account events like acquisition and loss

of genes as well as predicted gene functions in calculating the evolutionary distances between strains and species (Snel et al., 1999). Such trees have also been shown to be consistent with those constructed on the basis of rDNA genes. The availability of genome sequences of *H. somni* strains 2336 and 129Pt as well as other members of the *Pasteurellaceae* has raised the possibility of genome-based, comprehensive phylogenetic analyses. It is not unreasonable to expect that these analyses will be able to address questions related to genetic variation and ancestry. Such analyses may also become a milestone in the eventual classification of these bacteria within the context of "prokaryotic genomic taxonomy".

In conclusion, *H. somni* strain 2336 contains the largest chromosome in comparison to other *Haemophilus* and *Histophilus* strains whose genome sequences are available. Several regions that resemble the pathogenicity islands (PAIs) of other virulent bacteria are present in strain 2336. There is evidence to suggest that most of these regions were acquired via HGT mechanisms, whereas similar regions are not found in the commensal strain 129Pt. Although previous studies shed light on the genetic differences that correspond to the phenotypic dissimilarities between strains 2336 and 129Pt, comparative genome sequence analyses have provided a comprehensive account of innate and acquired genetic traits.

The post-genomic era for *H. somni* poses new challenges and opportunities in terms of functional characterization of genes and understanding their roles in colonization, survival, and pathogenesis. Continued analyses of the genomes of *H. somni* strains and comparing them to the newly sequenced genomes of other bacteria should enhance the current knowledge on virulence mechanisms. However, high throughput analyses like DNA microarrays and signature-tagged mutagenesis are promising approaches to address some of the challenges in a relatively short period of time, and a concerted effort is required to realize the genomic potentials of *H. somni*.

References

Brussow H, Canchaya C, Hardt WD (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* **68**: 560-602.

Buckley M (2004). The genomics of disease-causing organisms: Mapping a strategy for discovery and defense. *Colloquium Report of the American Academy of Microbiology* (available at http://www.asm.org/ASM/files/ccLibraryFiles/FILENAME/000000001048/PATHOGENESIS_F.pdf).

Canchaya C, Fournous G, Brüssow H (2004). The impact of prophages on bacterial chromosomes. *Molecular Microbiology* **53**: 9-18.

Challacombe JF, Duncan AJ, Brettin TS, Bruce D, Chertkov O, Detter JC, Han CS, Misra M, Richardson P, Tapia R, Thayer N, Xie G, Inzana TJ (2007). Complete genome sequence of *Haemophilus somnus* (*Histophilus somni*) strain 129Pt and comparison to *H. ducreyi* 35000HP and *H. influenzae* Rd. *Journal of Bacteriology* **189**: 1890-1898.

Chissoe SL, Wang YF, Clifton SW, Ma N, Sun HJ, Lobsinger JS, Kenton SM, White JD, Roe BA (1991). Strategies for rapid and accurate DNA sequencing. *Methods. A Companion To Methods In Enzymology* **3:** 55-65.

Coenye T, Gevers D, Van de Peer Y, Vandamme P, Swings J (2005). Towards a prokaryotic genomic taxonomy. *FEMS Microbiology Reviews* **29**: 147-167.

Cole SP, Guiney DG, Corbeil LB (1993). Molecular analysis of a gene encoding a serum-resistance-associated 76 kDa surface antigen of *Haemophilus somnus*. *Journal of General Microbiology* **139**: 2135-2143.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007-1011.

Cornelissen CN (2003). Transferrin-iron uptake by Gram-negative bacteria. *Frontiers in Bioscience* **8**: 836-847.

Corbeil LB, Gogolewski RP, Stephens LR, Inzana TJ (1995). *Haemophilus somnus*: antigen analysis and immune responses. In: Donachie W, Lainson FA, Hodgson JC (Editors). *Haemophilus, Actinobacillus, and Pasteurella.* New York, NY: Plenum Press, pp. 63-73.

Corbeil LB, Bastida-Corcuera FD, Beveridge TJ (1997). *Haemophilus somnus* immunoglobulin binding proteins and surface fibrils. *Infection and Immunity* **65**: 4250-4257.

Ekins A, Bahrami F, Sijercic A, Maret D, Niven DF (2004). *Haemophilus somnus* possesses two systems for acquisition of transferrin-bound iron. *Journal of Bacteriology* **186**: 4407-4411.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, Fitzhugh W, Fields C, Gocyne JD, Scott J, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. **269**: 496-512.

Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S (2000). Comparative genomics and understanding of microbial biology. *Emerging Infectious Diseases* **6**: 505-512.

Fuller TE, Kennedy MJ, Lowery DE (2000). Identification of *Pasteurella multocida* virulence genes in a septicemic mouse model using signature-tagged mutagenesis. *Microbial Pathogenesis* **29**: 25-38.

Gioia J, Qin X, Jiang H, Clinkenbeard K, Lo R, Liu Y, Fox GE, Yerrapragada S, McLeod MP, McNeill TZ, Hemphill L, Sodergren E, Wang Q, Muzny DM, Homsi FJ, Weinstock GM, Highlander SK (2006). The genome sequence of *Mannheimia haemolytica* A1: insights into virulence, natural competence, and *Pasteurellaceae* phylogeny. *Journal of Bacteriology* **188**: 7257-7266.

Gladitz J, Shen K, Antalis P, Hu FZ, Post JC, Ehrlich GD (2005). Codon usage comparison of novel genes in clinical isolates of *Haemophilus influenzae*. *Nucleic Acids Research* **33**: 3644-3658.

Gordon D, Desmarais C, Green P (2001). Automated finishing with autofinish. *Genome Research* **11**: 614-625.

Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Molecular Microbiology* **23**: 1089-1097.

Hacker J, Hentschel U, Dobrindt U (2003). Prokaryotic chromosomes and disease. *Science* **301**: 790-793.

Halling SM, Peterson-Burch BD, Bricker BJ, Zuerner RL, Qing Z, Li LL, Kapur V, Alt DP, Olsen SC (2005). Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. *Journal of Bacteriology* **187**: 2715-2726.

Harrison A, Dyer DW, Gillaspy A, Ray WC, Mungur R, Carson MB, Zhong H, Gipson J, Gipson M, Johnson LS, Lewis L, Bakaletz LO, Munson RS Jr. (2005). Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. *Journal of Bacteriology* **187**: 4627-4636.

Inzana TJ, Gogolewski RP, Corbeil LB (1992). Phenotypic phase variation in *Haemophilus somnus* lipooligosaccharide during bovine pneumonia and after *in vitro* passage. *Infection and Immunity* **60**: 2943-2951.

Inzana TJ, Glindemann G, Cox AD, Wakarchuk W, Howard MD (2002). Incorporation of N-acetylneuraminic acid into *Haemophilus somnus* lipooligosaccharide (LOS): enhancement of resistance to serum and reduction of LOS antibody binding. *Infection and Immunity* **70**: 4870-4879.

Jacob-Dubuisson F, Locht C, Antoine R (2001). Two-partner secretion in Gram-negative bacteria: a thrifty, specific pathway for large virulence proteins. *Molecular Microbiology* **40**: 306-313.

Karlyshev AV, Pallen MJ, Wren BW (2000). Single-primer PCR procedure for rapid identification of transposon insertion sites. *Biotechniques* **28**:1078.

Kobayashi I (2004). Restriction-modification systems as minimal forms of life. In: Pingoud A (Editor) *Restriction Endonucleases*. Berlin: Springer**,** pp.19-62.

Konstantinidis KT and Tiedje JM (2005). Towards a genome-based taxonomy for prokaryotes. *Journal of Bacteriology* **187**: 6258-6264.

Lander ES and Waterman MS (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231-239.

Lawrence JG, Ochman H (1997). Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution* **44**: 383-397.

May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V (2001). Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proceedings of the National Academy of Sciences of the U.S.A.* **98**: 3460-3465.

McQuiston JH, McQuiston JR, Cox AD, Wu Y, Boyle SM, Inzana TJ (2000). Characterization of a DNA region containing 5'-$(CAAT)_{(n)}$-3' DNA sequences involved in lipooligosaccharide biosynthesis in *Haemophilus somnus*. *Microbial Pathogenesis* **28**: 301-312.

Moret BME, Bader DA, Warnow T (2002). High-performance algorithm engineering for computational phylogenetics *The Journal of Supercomputing* **22**: 99-111.

Olson MV (1999). When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*. **64**: 18-23.

Siezen RJ, Renckens B, Boekhorst J (2007). Evolution of prokaryotic subtilases: genome-wide analysis reveals novel subfamilies with different catalytic residues. *Proteins* **67**: 681-694.

Snel B, Bork P, Huynen MA (1999). Genome phylogeny based on gene content. *Nature Genetics* **21**: 108-110.

Tagawa Y, Sanders JD, Uchida I, Bastida-Corcuera FD, Kawashima K, Corbeil LB (2005). Genetic and functional analysis of *Haemophilus somnus* high molecular weight-immunoglobulin binding proteins. *Microbial Pathogenesis* **39**: 159-170.

Tillier ER, Collins RA (2000). Genome rearrangement by replication-directed translocation. *Nature Genetics* **26**:195-197.

USDA (2000). Outlining priorities in the area of animal pathogen genomics and providing recommendations to the USDA for leadership opportunities. *USDA Microbial Genomics Workshop 2000: Report and Recommendations of Animal Agriculture Stakeholders* (available at http://genome.cvm.umn.edu/microbialgenomics.pdf).

Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Research* **33** (Web Server issue): W455-W459.

Ventura M, Canchaya C, Pridmore D, Berger B, Brussow H (2003). Integration and distribution of *Lactobacillus johnsonii* prophages. *Journal of Bacteriology* **185**: 4603-4608.

Ward CK, Lumbley SR, Latimer JL, Cope LD, Hansen EJ (1998). *Haemophilus ducreyi* secretes a filamentous hemagglutinin-like protein. *Journal of Bacteriology* **180**: 6013-6022.

Ward CK, Mock JR, Hansen EJ (2004). The LspB protein is involved in the secretion of the LspA1 and LspA2 proteins by *Haemophilus ducreyi*. *Infection and Immunity* **72**: 1874-1884.

Ward AC, Weiser GC, Anderson BC, Cummings PJ, Arnold KF, Corbeil LB (2006). *Haemophilus somnus* (*Histophilus somni*) in bighorn sheep. *Canadian Journal of Veterinary Research* **70**: 34-42.

Wu Y, McQuiston JH, Cox A, Pack TD, Inzana TJ (2000). Molecular cloning and mutagenesis of a DNA locus involved in lipooligosaccharide biosynthesis in *Haemophilus somnus.* *Infection and Immunity* **68**: 310-319.

Yu RH, Gray-Owen SD, Ogunnariwo J, Schryvers AB (1992). Interaction of ruminant transferrins with transferrin receptors in bovine isolates of *Pasteurella haemolytica* and *Haemophilus somnus*. *Infection and Immunity* **60**: 2992-2994.